

Health Services Research

*Series Editor:* Boris Sobolev

Adrian Levy · Sarah Goring

Constantine Gatsonis · Boris Sobolev

Ewout van Ginneken · Reinhard Busse *Editors*

# Health Services Evaluation



Springer Reference

---

# Health Services Research

**Series Editor**

Boris Sobolev  
University of British Columbia  
Vancouver, BC, Canada

Health services research is the study of the organization, uses and outcomes of health care. The societal value of health services research lies in identifying the ways in which health care can best be organized, financed, and delivered. This ambitious agenda brings together researchers from a wide range of disciplinary backgrounds that are required for evaluating the effectiveness of diagnostic technologies, treatment procedures, and managerial solutions. The series is envisaged as a collection that overviews the established knowledge and provides access to accepted information in the field. The content is grouped into six major areas.

1. Clinical evaluation of health care outcomes
2. Medical practice variations
3. Research methods
4. Health care systems and policies
5. Sources of data
6. Health economics in health services research.

The series will be of significant interest for healthcare professionals, program directors, service administrators, policy and decision makers, as well as for graduate students, educators, and researchers in healthcare evaluation.

More information about this series at <http://www.springer.com/series/13490>

---

Adrian Levy • Sarah Goring  
Constantine Gatsonis • Boris Sobolev  
Ewout van Ginneken • Reinhard Busse  
Editors

# Health Services Evaluation

With 142 Figures and 137 Tables

 Springer Reference

*Editors*

Adrian Levy  
Community Health and Epidemiology  
Dalhousie University  
Halifax, NS, Canada

Sarah Goring  
ICON plc  
Vancouver, BC, Canada

Constantine Gatsonis  
Department of Biostatistics  
Brown University  
Providence, RI, USA

Boris Sobolev  
University of British Columbia  
Vancouver, BC, Canada

Ewout van Ginneken  
Berlin University of Technology  
Berlin, Germany

Reinhard Busse  
Technische Universität Berlin  
Berlin, Germany

European Observatory on Health  
Systems and Policies, Department of  
Health Care Management  
Berlin University of Technology  
Berlin, Germany

Department Health Care Management  
Faculty of Economics and Management  
Technische Universität  
Berlin, Germany

ISSN 2511-8293

ISSN 2511-8307 (electronic)

ISBN 978-1-4939-8714-6

ISBN 978-1-4939-8715-3 (eBook)

ISBN 978-1-4939-8716-0 (print and electronic bundle)

<https://doi.org/10.1007/978-1-4939-8715-3>

Library of Congress Control Number: 2018960887

© Springer Science+Business Media, LLC, part of Springer Nature 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Science+Business Media, LLC part of Springer Nature.

The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

---

## Series Preface

Health Services Research has experienced explosive growth in the past three decades. The new field was formed at the interface of a number of disciplines, including medicine, statistics, economics, management science, and the social and behavioral sciences, which came together around the study of health care practice, delivery and outcomes. The rich, multidisciplinary research enterprise that developed from this fusion has already produced a growing and sophisticated body of subject matter research and has also defined a body of methodology that is integral to the field. True to the multidisciplinary origins of the field, its methods continue to benefit from developments in diverse disciplines, while formulating and addressing scientific questions that are unique to health care and outcomes research.

The societal value of health services research lies in identifying the ways in which health care can best be organized, financed, and delivered. This ambitious agenda brings together researchers from a wide range of disciplinary backgrounds who are required for evaluating the effectiveness of diagnostic technologies, treatments, procedures, and health delivery systems as no single discipline provides a full perspective on how the health systems operate.

A fundamental discovery was the persistent variation in health care utilization across providers, regions and countries, variation that cannot be explained by population illness level, known benefit or patient preference. Another discovery was that treatments and procedures that are meant to benefit patients may produce adverse events and unintended consequences. We have learned that results of randomized clinical trials cannot always be generalized to clinical practice because patients enrolled in trials can be highly selective. Researchers have been able to identify patients who may benefit from a treatment but there are groups of patients for whom the optimal treatment is not well defined or may depend on their personal preferences. Learning what works in real life gave rise to comparative effectiveness research.

The Health Services Research series addresses the increasing need for a comprehensive reference in the field of inquiry that welcomes interdisciplinary collaborations. This major reference work aims to be a source of information for everyone who seeks to develop an understanding of health services and health systems, and learn about the historic, political, and economic factors

that influence health policies at global, national, regional and local levels. The intended readership includes graduate students, educators, researchers, healthcare professionals, policy makers and service administrators.

The main reason for public support of health services research is the common understanding that new knowledge will lead to more effective health care. Over the past decades, we have witnessed the increased prominence of health services and health policy research since the knowledge, skills and approaches required for ground-breaking work distinguish it from other specialties. An important step towards the formation of the profession is a comprehensive reference work of established knowledge. The Health Services Research series is intended to provide the health services researcher a home for the foundations of the profession.

The Health Services Research series is available in both printed and online formats. The online version will serve as a web-based conduit of information that evolves as knowledge content expands. This innovative depository of knowledge will offer various search tools, including cross-referencing across chapters and linking to supplement data, other Springer reference works and external articles.

July 2015

Boris Sobolev

---

# Contents

<b>Part I</b>	<b>Data and Measures in Health Services Research</b>	<b>1</b>
<b>1</b>	<b>Health Services Data: Big Data Analytics for Deriving Predictive Healthcare Insights</b>	<b>3</b>
	Ankit Agrawal and Alok Choudhary	
<b>2</b>	<b>Health Services Data: Managing the Data Warehouse: 25 Years of Experience at the Manitoba Centre for Health Policy</b>	<b>19</b>
	Mark Smith, Leslie L. Roos, Charles Burchill, Ken Turner, Dave G. Towns, Say P. Hong, Jessica S. Jarmasz, Patricia J. Martens, Noralou P. Roos, Tyler Ostapyk, Joshua Ginter, Greg Finlayson, Lisa M. Lix, Marni Brownell, Mahmoud Azimae, Ruth-Ann Soodeen, and J. Patrick Nicol	
<b>3</b>	<b>Health Services Data, Sources and Examples: The Institute for Clinical Evaluative Sciences Data Repository</b>	<b>47</b>
	Karey Iron and Kathy Sykora	
<b>4</b>	<b>Health Services Data: The Centers for Medicare and Medicaid Services (CMS) Claims Records</b>	<b>61</b>
	Ross M. Mullner	
<b>5</b>	<b>Health Services Data: Typology of Health Care Data</b>	<b>77</b>
	Ross M. Mullner	
<b>6</b>	<b>Health Services Information: Application of Donabedian’s Framework to Improve the Quality of Clinical Care</b>	<b>109</b>
	A. Laurie W. Shroyer, Brendan M. Carr, and Frederick L. Grover	
<b>7</b>	<b>Health Services Information: Data-Driven Improvements in Surgical Quality: Structure, Process, and Outcomes</b>	<b>141</b>
	Katia Noyes, Fergal J. Fleming, James C. Iannuzzi, and John R. T. Monson	



<b>8</b>	<b>Health Services Information: From Data to Policy Impact (25 Years of Health Services and Population Health Research at the Manitoba Centre for Health Policy)</b> .....	171
	Leslie L. Roos, Jessica S. Jarmasz, Patricia J. Martens, Alan Katz, Randy Fransoo, Ruth-Ann Soodeen, Mark Smith, Joshua Ginter, Charles Burchill, Noralou P. Roos, Malcolm B. Doupe, Marni Brownell, Lisa M. Lix, Greg Finlayson, and Maureen Heaman	
<b>9</b>	<b>Health Services Information: Key Concepts and Considerations in Building Episodes of Care from Administrative Data</b> .....	191
	Erik Hellsten and Katie Jane Sheehan	
<b>10</b>	<b>Health Services Information: Lessons Learned from the Society of Thoracic Surgeons National Database</b> .....	217
	David M. Shahian and Jeffrey P. Jacobs	
<b>11</b>	<b>Health Services Information: Patient Safety Research Using Administrative Data</b> .....	241
	Chunliu Zhan	
<b>12</b>	<b>Health Services Information: Personal Health Records as a Tool for Engaging Patients and Families</b> .....	265
	John Halamka	
<b>13</b>	<b>A Framework for Health System Comparisons: The Health Systems in Transition (HiT) Series of the European Observatory on Health Systems and Policies</b> .....	279
	Bernd Rechel, Suszy Lessof, Reinhard Busse, Martin McKee, Josep Figueras, Elias Mossialos, and Ewout van Ginneken	
<b>14</b>	<b>Health Services Knowledge: Use of Datasets Compiled Retrospectively to Correctly Represent Changes in Size of Wait List</b> .....	297
	Paul W. Armstrong	
<b>15</b>	<b>Waiting Times: Evidence of Social Inequalities in Access for Care</b> .....	345
	Luigi Siciliani	
<b>16</b>	<b>Health Services Data: The Ontario Cancer Registry (a Unique, Linked, and Automated Population-Based Registry)</b> .....	363
	Sujohn Prodhan, Mary Jane King, Prithwish De, and Julie Gilbert	
<b>17</b>	<b>Challenges of Measuring the Performance of Health Systems</b> .....	391
	Adrian R. Levy and Boris G. Sobolev	

<b>Part II</b>	<b>Methods in Health Services Research</b>	<b>403</b>
18	<b>Analysis of Repeated Measures and Longitudinal Data in Health Services Research</b>	405
	Juned Siddique, Donald Hedeker, and Robert D. Gibbons	
19	<b>Competing Risk Models</b>	433
	Melania Pintilie	
20	<b>Modeling and Analysis of Cost Data</b>	447
	Shizhe Chen and XH Andrew Zhou	
21	<b>Instrumental Variable Analysis</b>	479
	Michael Baiocchi, Jing Cheng, and Dylan S. Small	
22	<b>Introduction to Causal Inference Approaches</b>	523
	Elizabeth A. Stuart and Sarah Naeger	
23	<b>Measurement of Patient-Reported Outcomes of Health Services</b>	537
	Joseph C. Cappelleri and Andrew G. Bushmakin	
24	<b>Micro-simulation Modeling</b>	559
	Carolyn M. Rutter	
25	<b>Network Meta-analysis</b>	577
	Georgia Salanti, Deborah Caldwell, Anna Chaimani, and Julian Higgins	
26	<b>Introduction to Social Network Analysis</b>	617
	Alistair James O'Malley and Jukka-Pekka Onnela	
27	<b>Survey Methods in Health Services Research</b>	661
	Steven B. Cohen	
28	<b>Two-Part Models for Zero-Modified Count and Semicontinuous Data</b>	695
	Brian Neelon and Alistair James O'Malley	
29	<b>Data Confidentiality</b>	717
	Theresa Henle, Gregory J. Matthews, and Ofer Harel	
30	<b>Qualitative Research</b>	733
	Cynthia Robins	
<b>Part III</b>	<b>Health Care Systems and Policies</b>	<b>753</b>
31	<b>Assessing Health Systems</b>	755
	Irene Papanicolas and Peter C. Smith	
32	<b>Health System in Canada</b>	769
	Gregory Marchildon	

---

<b>33</b>	<b>Health System in China</b> .....	779
	David Hipgrave and Yan Mu	
<b>34</b>	<b>Health System in Egypt</b> .....	809
	Christian A. Gericke, Kaylee Britain, Mahmoud Elmahdawy, and Gihan Elsis	
<b>35</b>	<b>Health System in France</b> .....	827
	Karine Chevreul and Karen Berg Brigham	
<b>36</b>	<b>Health System in Japan</b> .....	837
	Ryozo Matsuda	
<b>37</b>	<b>Health System in Mexico</b> .....	849
	Julio Frenk and Octavio Gómez-Dantés	
<b>38</b>	<b>Health System in the Netherlands</b> .....	861
	Madelon Kroneman and Willemijn Schäfer	
<b>39</b>	<b>Health System in Singapore</b> .....	877
	William A. Haseltine and Chang Liu	
<b>40</b>	<b>Health System in the USA</b> .....	891
	Andrew J. Barnes, Lynn Y. Unruh, Pauline Rosenau, and Thomas Rice	
<b>41</b>	<b>Health System Typologies</b> .....	927
	Claus Wendt	
<b>42</b>	<b>Organization and Governance: Stewardship and Governance in Health Systems</b> .....	939
	Scott L. Greer	
<b>43</b>	<b>Provision of Health Services: Long-Term Care</b> .....	949
	Vincent Mor and Anna Maresso	
<b>44</b>	<b>Provision of Health Services: Mental Health Care</b> .....	979
	Jon Cylus, Marya Saidi, and Martin Knapp	

---

## About the Series Editor



**Boris Sobolev** is a health services researcher from the University of British Columbia. He is author of *Analysis of Waiting-Time Data in Health Services Research and Health Care Evaluation Using Computer Simulation: Concepts, Methods and Applications*.

Dr. Sobolev started an academic career at the Radiation Epidemiology Institute in Kiev, studying the risk of cancer in relation to exposure resulting from the Chernobyl accident. In 1996, he came to Canada to work at Queen's University in Kingston, where he studied how people get access to health care, what services they use, and what happens to patients as a result. There, he pioneered the epidemiological approach to studying the risk of adverse events in relation to time of receiving medical services.

Later, Dr. Sobolev joined the University of British Columbia, Canada, where he is a Professor at the School of Population and Public Health. There, he has taught a variety of courses and introduced into the curriculum a new course on causal inferences in health sciences. He was awarded a Canada Research Chair in Statistics and Modelling of the Health Care System, a distinction he held through 2013. Currently, he serves as principal investigator for the Canadian Collaborative Study on Hip Fractures.

Dr. Sobolev also leads the Health Services and Outcomes Research Program at the Centre for Clinical Epidemiology and Evaluation at the Vancouver

General Hospital. The program's mission is closing the gap between health care that is possible and health care that is delivered. This ambitious agenda brings together researchers from a wide range of disciplinary backgrounds that are required for evaluating the effectiveness of diagnostic technologies, treatment procedures, and managerial solutions. The program's investigators empirically assess the benefits and harms of therapeutic and health care interventions in the acute and primary care setting, using patient registries and data from routine medical care. By learning what works in everyday clinical practice the program generates knowledge that helps physicians and patients to make shared decisions about the best approach to treatment.

Dr. Sobolev promotes and advances the causality perspective in health services research for informing policy and decision-making. In particular, his recent work helped to estimate the reduction in postoperative mortality expected from providing timely cardiac surgical care; the health effects of receiving hip fracture surgery within the government benchmark; the proportion of hospital readmissions that could be avoided had patients undergone medication review in emergency departments rather than in hospital wards; and the expected reduction of mortality had all coronary obstructive pulmonary disease patients had their second exacerbation prevented.

---

## About the Editors



**Adrian Levy** is professor of epidemiology and health services research working at Dalhousie University in Halifax, Nova Scotia. Dr. Levy commenced his academic career working for the Quebec Council for Health Technology Assessment doing applied health research on real-world use of health technologies such as extracorporeal shock wave lithotripsy and complex operations. His doctoral dissertation in epidemiology was completed at McGill University (1998) followed by postgraduate training in economic evaluation at McMaster University (2000). In 2000, Dr. Levy joined the faculty in the School of Population and Public Health at the University of British Columbia and was awarded British Columbia Michael Smith Foundation for Health Research Scholar (2001) and Senior Scholar (2006) awards and a New Investigator Award from the Canadian Institutes of Health Research (2004). There, he linked administrative health databases with patient and treatment registries to study access, quality, and cost of care in cardiac surgery, HIV, and transplant.

In 2009, Dr. Levy joined Dalhousie University in Halifax, Nova Scotia, Canada, to serve as head of the Department of Community Health and Epidemiology. As an integral part of the Medical School of the Maritimes, the Department's collective purpose is to enhance the capacity to improve the health of individuals, patients, communities, populations, and systems, by

serving as leaders who generate evidence and apply critical thinking to the health challenges of today and tomorrow. The Department's faculty generate evidence and engage in knowledge exchange that advances effective and sustainable systems for health services access and delivery.

As nominated principal investigator, Dr. Levy led the development and implementation of the Maritime Strategy for Patient-Oriented Research SUPPORT Unit. This initiative, co-funded by the Canadian Institutes of Health Research, offers research infrastructure designed to promote patient-centered outcomes and health services research in Canada's three Maritime provinces. The Unit's mission is to lead the development and application of patient-centered outcomes research, and the vision is to enhance the health and well-being of individuals and populations in the Maritimes and across Canada. The central goals include advancing research on health systems, knowledge translation and implementation of healthcare transformation, and implementing research at the point of care.



**Sarah Goring** has an M.Sc. in healthcare and epidemiology from the University of British Columbia and more than 10 years of experience consulting in the private sector, where she focuses on pharmacoepidemiology, evidence synthesis methods, and health services research.



**Constantine Gatsonis** is Henry Ledyard Goddard University Professor and founding chair of the Department of Biostatistics and the Center for Statistical Sciences at the Brown University School of Public Health. Dr. Gatsonis is a

leading authority on the evaluation of diagnostic and screening tests and has made major contributions to statistical methods for medical technology assessment and health services and outcomes research. His current research activity spans the spectrum of evidence-based diagnostic medicine, addresses both methodology and subject matter, and has a major focus on the comparative effectiveness of screening and diagnostic modalities. As the founding network statistician of the American College of Radiology Imaging Network (ACRIN) and a group statistician for the ECOG-ACRIN collaborative group, he has decades-long experience in the clinical evaluation of modalities for diagnosis and prediction in cancer and other chronic diseases. Dr. Gatsonis has served on numerous review and advisory panels. He chaired the NAS Committee on Applied and Theoretical Statistics and is a member of the NAS Committee on National Statistics. He served on the IOM Committee on Comparative Effectiveness Research Prioritization and the NAS Committee on Reproducibility and Replicability in Science and was the founding editor-in-chief of *Health Services and Outcomes Research Methodology*. Dr. Gatsonis was educated at Princeton and Cornell, was elected fellow of the American Statistical Association, and received a Long-Term Excellence Award from the Health Policy Statistics Section of ASA.



**Ewout van Ginneken** is coordinator of the Berlin office of the European Observatory on Health Systems and Policies at the Berlin University of Technology. He holds a master's degree in health policy and administration from Maastricht University in the Netherlands and a Ph.D. in public health from the Berlin University of Technology. His expertise is in comparative international health systems research and health policy research. His main interests include health financing, insurance competition, care purchasing, integrated care, cross-border care, and migrants' access to care. He has edited several Health Systems in Transition (HiT) reviews including on the healthcare systems of Bulgaria, the Czech Republic, Estonia, Lithuania, the Netherlands, Slovakia, Slovenia, and the United States. He has published widely on these topics in international peer-reviewed literature and the wider literature. Before joining the Observatory, Ewout was a senior researcher at the Berlin University of Technology and a 2011–2012 Commonwealth Fund Harkness Fellow in Health Care Policy and Practice at the Harvard School of Public Health.





**Reinhard Busse** is department head for healthcare management in the Faculty of Economics and Management at Technische Universität Berlin, Germany. He is also a faculty member of Charité, Berlin's medical faculty, co-director and head of the Berlin hub of the European Observatory on Health Systems and Policies, member of several scientific advisory boards, as well as regular consultant for the WHO, the EU Commission, the World Bank, the OECD, and other international organizations within Europe and beyond as well as national health and research institutions. From 2006 to 2009, he served as dean of his faculty.

His research focuses on methods and contents of comparative health system analysis and assessment as well as health services research (with emphasis on hospitals, human resources, cross-border care, health reforms in Germany, role of the EU, financing and payment mechanisms, as well as disease management), health economics, and health technology assessment (HTA).

His regular master-level teaching courses at TU Berlin include "Managing and Researching Health Care Systems"; "Health Technology Assessment" (blended learning, i.e., mainly online); "Health Care Management I, Insurance Management"; "Health Care Management II, Provider Management"; "Health Care Management III, Industry Management" (pharmaceuticals and medical devices); and "Health Care Management IV, Health Economic Evaluation." He is the principal editor of the German textbook on healthcare management (published with Springer, fourth edition 2017), author of a book on the German health system (fourth edition 2017), as well as co-editor of German textbooks on public health (third edition 2012) and on HTA (second edition 2014). Since 2015, he is speaker of the board of the newly founded inter-university Berlin School of Public Health.

Professor Busse is the director of the annual Observatory Summer School in Venice, which is directed at policy-makers and has covered a wide range of topics: human resources for health; hospital reengineering; innovation and health technology assessment; EU integration and health systems; the aging crisis; performance assessment for health system improvement; innovative ways of improving population health; integrated care – moving beyond the rhetoric; primary care – innovating for integrated, more effective care; and quality of care – improving effectiveness, safety, and responsiveness.

He was the PI/coordinator of the EU-funded project "EuroDRG: Diagnosis-Related Groups in Europe: Towards Efficiency and Quality" (Seventh Framework; 2009–2011). He has been and is also involved in several other

EU-funded projects under the Seventh Framework, e.g., on the relationship between nursing and patient outcome (RN4Cast; 2009–2011), mobility of health professionals (PROMeTHEUS; 2009–2011), evaluating care across borders (ECAB; 2009–2013), on healthcare data for cross-country comparisons of efficiency and quality (*EuroREACH*; 2010–2013), on the impact of new roles for health professionals (Munros; 2012–2016), and on advancing and strengthening HTA (Advance HTA; 2013–2015). Previously, he was PI/scientific coordinator of the EU-funded project “Health Benefits and Service Costs in Europe” (*HealthBASKET*; 2004–2007).

Since 2011, he is editor-in-chief of the international peer-reviewed journal *Health Policy*. Since 2012, he is the director of the Berlin Health Economics Research Centre (BerlinHECOR, overarching topic “Towards a Performance Assessment of the German Health Care System”), one of four centers in Germany funded by the Federal Ministry of Research. In 2016–2017, he was president of the German Health Economics Association (DGGÖ).

Professor Busse studied medicine in Marburg, Germany; Boston, USA; and London, UK, as well as public health in Hannover, Germany. Prior to his appointment at TU Berlin in 2002, he was head of the Observatory’s hub in Madrid, Spain (1999–2002); a senior research fellow in the Department of Epidemiology, Social Medicine and Health Systems Research (1994–1999, finishing with his “*habilitation*”/second Ph.D.) and a resident physician in the Department of Rheumatology (1992–1994), both at the Hannover Medical School; and a researcher in the Planning Group for a Problem-Based Medical Curriculum at the Freie Universität Berlin (1991–1992). In 1993, he earned a “*Dr. med.*” (Ph.D. in medicine) from Philipps-Universität in Marburg.

---

## Contributors

**Ankit Agrawal** Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA

**Paul W. Armstrong** London, UK

**Mahmoud Azimaee** ICES Central, Toronto, ON, Canada

**Michael Baiocchi** Department of Statistics, Stanford University, Stanford, CA, USA

**Andrew J. Barnes** Department of Health Behavior and Policy, School of Medicine, Virginia Commonwealth University, Richmond, VA, USA

**Karen Berg Brigham** University of Washington, Seattle, WA, USA

**Kaylee Britain** University of Queensland School of Public Health, Brisbane, Australia

**Marni Brownell** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**Charles Burchill** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**Andrew G. Bushmakin** Global Product Development, Pfizer Inc, Groton, CT, USA

**Reinhard Busse** Technische Universität, Berlin, Germany

Department Health Care Management, Faculty of Economics and Management, Technische Universität, Berlin, Germany

**Deborah Caldwell** School of Social and Community Medicine, University of Bristol, Bristol, UK

**Joseph C. Cappelleri** Global Product Development, Pfizer Inc, Groton, CT, USA

**Brendan M. Carr** Department of Emergency Medicine, Mayo Clinic, Rochester, MN, USA

**Anna Chaimani** Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece

**Shizhe Chen** Department of Biostatistics, University of Washington, Seattle, WA, USA

**Jing Cheng** Department of Preventive and Restorative Dental Sciences, University of California, San Francisco School of Dentistry, San Francisco, CA, USA

**Karine Chevreul** Health Economics and Health Services Research Unit, URC ECO Ile de France, Paris, France

**Alok Choudhary** Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA

**Steven B. Cohen** Division of Statistical and Data Sciences, RTI International, Washington, DC, USA

**Jon Cylus** The London School of Economics and Political Science, London, UK

**Prithwish De** Surveillance and Ontario Cancer Registry, Cancer Care Ontario, Toronto, ON, Canada

**Malcolm B. Doupe** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**Mahmoud Elmahdawy** Ministry of Health, Cairo, Egypt

**Gihan Elsisy** Ministry of Health, Cairo, Egypt  
Faculty of Pharmacy, Heliopolis University, Cairo, Egypt

**Josep Figueras** European Observatory on Health Systems and Policies, Brussels, Belgium

**Greg Finlayson** Finlayson and Associates Consulting, Kingston, ON, Canada

**Fergal J. Fleming** University of Rochester Medical Center, Rochester, NY, USA

**Randy Fransoo** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**Julio Frenk** University of Miami, Coral Gables, FL, USA

**Octavio Gómez-Dantés** National Institute of Public Health, Cuernavaca, MOR, Mexico

**Christian A. Gericke** Anton Breinl Centre for Health Systems Strengthening, James Cook University, Cairns, Australia  
University of Queensland School of Public Health, Brisbane, Australia

**Robert D. Gibbons** Departments of Medicine and Public Health Sciences, University of Chicago, Chicago, IL, USA

**Julie Gilbert** Planning and Regional Programs, Cancer Care Ontario, Toronto, ON, Canada

**Joshua Ginter** Montreal, QC, Canada

**Scott L. Greer** Department of Health Management and Policy, University of Michigan, Ann Arbor, MI, USA

**Frederick L. Grover** Department of Surgery, School of Medicine at the Anschutz Medical Campus, University of Colorado, Aurora, CO, USA

**John Halamka** Department of Emergency Medicine, Harvard Medical School and Beth Israel Deaconess Medical Center, Boston, MA, USA

**Ofer Harel** Department of Statistics, University of Connecticut, Storrs, CT, USA

**William A. Haseltine** ACCESS Health International, New York, NY, USA

**Maureen Heaman** College of Nursing, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB, Canada

**Donald Hedeker** Department of Public Health Sciences, University of Chicago, Chicago, IL, USA

**Erik Hellsten** Health Quality Ontario, Toronto, ON, Canada

**Theresa Henle** Department of Mathematics and Statistics, Loyola University, Chicago, IL, USA

**Julian Higgins** MRC Biostatistics Unit, Cambridge, UK  
Centre for Reviews and Dissemination, University of York, York, UK

**David Hipgrave** UNICEF, New York, NY, USA  
Nossal Institute for Global Health, University of Melbourne, Melbourne, VIC, Australia

**Say P. Hong** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**James C. Iannuzzi** University of Rochester Medical Center, Rochester, NY, USA

**Karey Iron** College of Physicians and Surgeons of Ontario, Toronto, ON, Canada

**Jeffrey P. Jacobs** Division of Cardiac Surgery, Department of Surgery, Johns Hopkins University School of Medicine, Baltimore, MA, USA  
Johns Hopkins All Children's Heart Institute, Saint Petersburg/Tampa, FL, USA

**Jessica S. Jarmasz** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**Alan Katz** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**Mary Jane King** Surveillance and Ontario Cancer Registry, Cancer Care Ontario, Toronto, ON, Canada

**Martin Knapp** The London School of Economics and Political Science, London, UK

**Madelon Kroneman** Netherlands Institute of Health Services Research (NIVEL), Utrecht, The Netherlands

**Suszy Lessof** European Observatory on Health Systems and Policies, Brussels, Belgium

**Adrian R. Levy** Community Health and Epidemiology, Dalhousie University, Halifax, NS, Canada

**Chang Liu** ACCESS Health International, New York, NY, USA

**Lisa M. Lix** Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, Canada

**Gregory Marchildon** Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

**Anna Maresso** European Observatory on Health Systems and Policies, London School of Economics and Political Science, London, UK

**Patricia J. Martens** Winnipeg, MB, Canada

**Ryozo Matsuda** Ritsumeikan University, Kyoto, Japan

**Gregory J. Matthews** Department of Mathematics and Statistics, Loyola University, Chicago, IL, USA

**Martin McKee** London School of Hygiene and Tropical Medicine, London, UK

**John R. T. Monson** Florida Hospital System Center for Colon and Rectal Surgery, Florida Hospital Medical Group Professor of Surgery, University of Central Florida, College of Medicine, Florida Hospital, Orlando, FL, USA

**Vincent Mor** Department of Health Services, Policy and Practice, Brown University School of Public Health, Providence, RI, USA

Providence Veterans Administration Medical Center, Center on Innovation, Providence, RI, USA

**Elias Mossialos** London School of Economics and Political Science, London, UK

**Yan Mu** UNICEF China, Beijing, China

**Ross M. Mullner** Division of Health Policy and Administration, School of Public Health, University of Illinois, Chicago, IL, USA

**Sarah Naeger** Behavioral Health Research and Policy, IBM Watson Health, Bethesda, MD, USA

**Brian Neelon** Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA

**J. Patrick Nicol** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**Katia Noyes** Department of Surgery, University of Rochester Medical Center, Rochester, NY, USA

**Alistair James O'Malley** The Dartmouth Institute for Health Policy and Clinical Practice, Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

Department of Health Care Policy, Harvard Medical School, Boston, MA, USA

**Jukka-Pekka Onnela** Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

**Tyler Ostapyk** University Advancement, Carleton University, Ottawa, ON, Canada

**Irene Papanicolas** The London School of Economics and Political Science, London, UK

Harvard T.H. Chan School of Public Health, Cambridge, MA, USA

**Melania Pintilie** University Health Network, Toronto, ON, Canada

**Sujohn Proadhan** Surveillance and Ontario Cancer Registry, Cancer Care Ontario, Toronto, ON, Canada

**Bernd Rechel** European Observatory on Health Systems and Policies, London School of Hygiene and Tropical Medicine, London, UK

**Thomas Rice** Department of Health Policy and Management, Fielding School of Public Health, University of California, Los Angeles, CA, USA

**Cynthia Robins** Westat, Rockville, MD, USA

**Leslie L. Roos** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**Noralou P. Roos** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**Pauline Rosenau** Division of Management, Policy and Community Health, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA

**Carolyn M. Rutter** RAND Corporation, Santa Monica, CA, USA

**Marya Saidi** The London School of Economics and Political Science, London, UK

**Georgia Salanti** Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece

**Willemijn Schäfer** Netherlands Institute of Health Services Research (NIVEL), Utrecht, The Netherlands

**David M. Shahian** Department of Surgery and Center for Quality and Safety, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

**Katie Jane Sheehan** School of Population and Public Health, The University of British Columbia, Vancouver, BC, Canada

**A. Laurie W. Shroyer** Department of Surgery, School of Medicine, Stony Brook University, Stony Brook, NY, USA

**Luigi Siciliani** Department of Economics and Related Studies, University of York, York, UK

**Juned Siddique** Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

**Dylan S. Small** University of Pennsylvania, Philadelphia, PA, USA

**Mark Smith** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**Peter C. Smith** Imperial College, London, UK  
University of York, York, UK

**Boris G. Sobolev** School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada

**Ruth-Ann Soodeen** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**Elizabeth A. Stuart** Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

**Kathy Sykora** Toronto, ON, Canada

**Dave G. Towns** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**Ken Turner** Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada

**Lynn Y. Unruh** Department of Health Management and Informatics, College of Health and Public Affairs, University of Central Florida, Orlando, FL, USA

**Ewout van Ginneken** Berlin University of Technology, Berlin, Germany  
European Observatory on Health Systems and Policies, Department of Health Care Management, Berlin University of Technology, Berlin, Germany

**Claus Wendt** University of Siegen, Siegen, Germany



**Chunliu Zhan** Department of Health and Human Services, Agency for Healthcare Research and Quality, Rockville, MD, USA

**XH Andrew Zhou** Beijing International Center for Mathematical Research, Peking University, Beijing, China

VA Puget Sound Healthcare System, University of Washington, Seattle, WA, USA

---

**Part I**

**Data and Measures in Health Services  
Research**



# Health Services Data: Big Data Analytics for Deriving Predictive Healthcare Insights

1

Ankit Agrawal and Alok Choudhary

## Contents

<b>Introduction</b> .....	3
<b>Big Data Analytics on SEER Lung Cancer</b>	
<b>Data</b> .....	6
Lung Cancer Survival Prediction System .....	6
Conditional Survival Prediction .....	10
Association Rule Mining .....	10
Illustrative Data Mining Results on SEER Data .....	10
Lung Cancer Outcome Calculator .....	12
Other Applications of Big Data Analytics in Healthcare .....	15
<b>Summary</b> .....	17
<b>References</b> .....	17

## Abstract

This chapter describes the application of big data analytics in healthcare, particularly on electronic healthcare records so as to make predictive models for healthcare outcomes and discover interesting insights. A typical workflow for such predictive analytics involves data collection, data transformation, predictive modeling, evaluation, and deployment, with each step tailored to the end goals of the project. To illustrate each of these steps, we shall take the example of recent advances in such predictive analytics on lung cancer data

from the Surveillance, Epidemiology, and End Results (SEER) program. This includes the construction of accurate predictive models for lung cancer survival, development of a lung cancer outcome calculator deploying the predictive models, and association rule mining on that data for bottom-up discovery of interesting insights. The lung cancer outcome calculator illustrated here is available at <http://info.eecs.northwestern.edu/LungCancerOutcomeCalculator>.

## Introduction

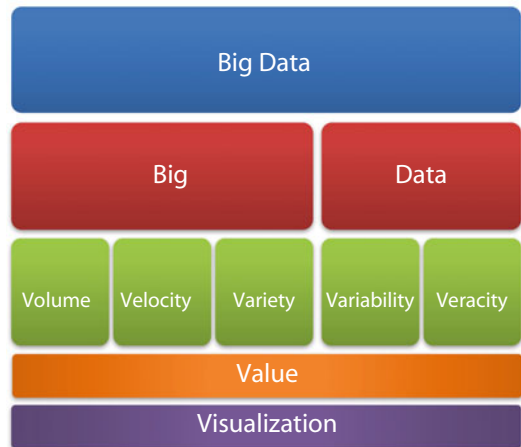
The term “big data” has become a ubiquitous buzzword today in practically all areas of science, technology, and commerce. It primarily denotes datasets that are too large, complex, or both, to be

A. Agrawal (✉) · A. Choudhary  
Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA  
e-mail: [ankita@eecs.northwestern.edu](mailto:ankita@eecs.northwestern.edu);  
[choudhar@eecs.northwestern.edu](mailto:choudhar@eecs.northwestern.edu)

adequately analyzed by traditional processing techniques. Scientific and technological advances in measurement and sensor devices, databases, and storage systems have made it possible to efficiently *collect*, *store*, and *retrieve* huge amounts of and different kinds of data. However, when it comes to the *analysis* of such data, we have to admit that our ability to generate big data has far outstripped our analytical ability to make sense of it. This is true in practically all fields, and the field of medicine and healthcare is no exception to it, where the fourth paradigm of science (data-driven analytics) is increasingly becoming popular and has led to the emergence of the new field of healthcare informatics. The fourth paradigm of science (Hey et al. 2009) unifies the first three paradigms of science – namely, theory, experiment, and simulation/computation. The need for such data-driven analytics in healthcare has also been emphasized by large-scale initiatives all around the world, such as Big Data to Knowledge (BD2K) and Precision Medicine Initiative of National Institutes of Health in the USA, Big Data for Better Outcomes Initiative in Europe, and so on.

The bigness (amount) of data is certainly the central feature and challenge of dealing with the so-called big data, but it is many times accompanied by one or more of other features that can make the collection and analysis of such data even more challenging. For example, the data could be from several heterogeneous sources, may be of different types, may have unknown dependencies and inconsistencies within it, parts of it could be missing or not reliable, the rate of data generation could be much more than what traditional systems could handle, and so on. All this can be summarized by the famous Vs associated with big data, as presented in Fig. 1 and briefly described below:

- *Volume*: It refers to the amount of data. Datasets of sizes exceeding terabytes and even petabytes are not uncommon today in many domains. This presents one of the biggest challenges in big data analytics.
- *Velocity*: The speed with which new data is generated. The challenge here is to be able to



**Fig. 1** The various Vs associated with big data. Volume, velocity, and variety are unique features of big data that represent its bigness. Variability and veracity are characteristics of any type of data, including big data. The goal of big data analytics is to unearth the value hidden in the data and appropriately visualize it to make informed decisions

effectively process the data in real time. A good example of high velocity data source is Twitter, where more than 5,000 tweets are posted every second.

- *Variety*: This refers to the heterogeneity in the data. For instance, many different types of healthcare data are generated and collected by different healthcare providers, such as electronic health records, X-rays, cardiograms, genomic sequence, etc. It is important to be able to derive insights by looking at all available heterogeneous data in a holistic manner.
- *Variability*: The inconsistency in the data. This is especially important since the correct interpretation of the data can vary significantly depending on its context.
- *Veracity*: It refers to how trustworthy the data is. The quality of the insights resulting from analysis of any data is critically dependent on the quality of the data itself. Noisy data with erroneous values or lot of missing values can greatly hamper accurate analysis.
- *Visualization*: It means the ability to interpret the data and resulting insights. Visualization can be especially challenging for big data due to its other features as described above.

- *Value*: The goal of big data analytics is to discover the hidden knowledge from huge amounts of data, which is akin to finding a needle in a haystack, and can be extremely valuable. For example, big data analytics in healthcare can help enable personalized medicine by identifying optimal patient-specific treatments, which can potentially improve millions of lives, reduce waste of healthcare resources, and save billions of dollars in healthcare expenditure.

The first three Vs above distinguish big data from small data, and other Vs are characteristics of any type of data, including big data. Further, each application domain can also introduce its own nuances to the process of big data management and analytics. For example, in healthcare, the privacy and security of patients' data are of paramount importance, and compliance to Health Insurance Portability and Accountability Act (HIPAA) and institutional review board (IRB) protocols is necessary to work with many types of healthcare data. It is also worth noting here that although the size and scale of healthcare data are not as large as in some other domains of science like high energy physics or in business and marketing, but the sheer complexity and variety in healthcare data becoming available nowadays require the development of new big data approaches in healthcare. For example, there are electronic healthcare records (EHRs), medical images (e.g., mammograms), time-series data (e.g., ECG signals), textual data (doctor's notes, research papers), genome sequence, and related data (e.g., SNPs).

So what can big data analytics do for a real-world healthcare application? A variety of personalized information such as patient's electronic health records is increasingly becoming available. What if we could intelligently integrate the hidden knowledge from such healthcare data during a real-time patient encounter to complement physician's expertise and potentially address the challenges of personalization, safe, and cost-effective healthcare? Note that the challenge here is to make the insights patient specific instead of giving generic population-wide statistics. Why is this

important? Let us try to understand with the help of an example. The benefits of medical treatments can vary depending on one's expected survival, and thus not considering an individual patient's prognosis can result in poor quality of care as well as nonoptimal use of healthcare resources. Developing accurate prognostic models using all available information and incorporating them into clinical decision support could thus significantly improve quality of healthcare (Collins et al. 2015), both in terms of improving clinical decision support and enhancing informed patient consent. Development of accurate data-driven models can also have a tremendous economic impact. The Centers for Disease Control and Prevention estimates that there are more than 150,000 surgical site infections annually (Magill et al. 2014), and it can cost \$11,000–\$35,000 per patient, i.e., about \$5 billion every year. Accurate predictions and risk estimation for healthcare outcomes can potentially avoid thousands of complications, resulting in improved resource management and significantly reduced costs. This requires development of advanced data-driven technologies that could effectively mine all available historical data, extract and suitably store the resulting insights and models, and make them available at the point of care in a patient-specific way.

In the rest of this chapter, we will see one such application of big data analytics on electronic healthcare records so as to make predictive models on it and discover interesting insights. In particular, we will take the example of lung cancer data from the Surveillance, Epidemiology, and End Results (SEER) program to build models of patient survival after 6 months, 9 months, 1 year, 2 years, and 5 years (Agrawal et al. 2011a) and for conditional survival as well (Agrawal et al. 2012). We will also see the application of association rule mining on this dataset for 5-year survival (Agrawal et al. 2011b) and 5-year conditional survival (Agrawal and Choudhary 2011). Finally, we will discuss the online lung cancer outcome calculator that resulted from the described predictive analytics on SEER data and conclude with some examples of big data analytics in other healthcare-related applications.

## Big Data Analytics on SEER Lung Cancer Data

Lung (respiratory) cancer is the second most common cancer and the leading cause of cancer-related deaths in the USA. In 2012 alone, over 157,000 people in the USA died from lung cancer. The 5-year survival rate for lung cancer is estimated to be just 15% (Ries et al. 2007). The Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute (NCI) is an authoritative repository of cancer statistics in the USA (SEER 2008). It is a population-based cancer registry covering about 26% of the US population and is the largest publicly available cancer dataset in the USA. It collects cancer data for all invasive and in situ cancers, except basal and squamous cell carcinomas of the skin and in situ carcinomas of the uterine cervix (Ries et al. 2007). The SEER data attributes can be broadly categorized into demographic attributes, diagnosis attributes, treatment attributes, and outcome attributes (see Table 1). The presence of outcome attributes makes the SEER data very useful for doing predictive analytics and making models for cancer survival.

## Lung Cancer Survival Prediction System

Till now we have seen what big data is and what big data analytics can do for healthcare applications. We have also had a brief introduction to SEER and what kind of data is present in the SEER database. So now let us dive deeper into what a typical workflow for predictive analytics looks like, with the specific example of lung cancer survival

prediction on SEER data. Figure 2 depicts the overall end-to-end workflow. It is worth mentioning here that this workflow for predictive lung cancer outcome analytics is essentially a healthcare adaptation of existing similar data science workflows in other domains, since most of the advanced techniques for big data management and analytics are invented in the field of computer science and more specifically high-performance data mining (Agrawal et al. 2013a; Xie et al. 1072), via applications in many different domains like business and marketing (Xie et al. 2012), climate science (Ganguly et al. 2014), materials informatics (Agrawal and Choudhary 2016), and social media analytics (Xie et al. 2013), among many others. Here we will only focus on the healthcare application of developing a lung cancer survival prediction system. As shown in Fig. 2, it has five stages described below.

### Data Collection

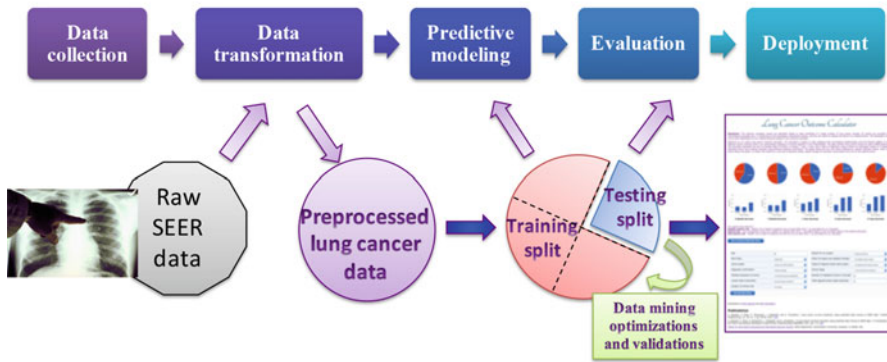
This is the obvious first step. Depending on the project, the kind of data required for it, and the license agreements associated with that data, this can be the easiest or the toughest step in the workflow. SEER has made it easy to get the “SEER limited-use data” from their website on submitting a SEER limited-use data agreement form. It creates a personalized SEER research data agreement for every user that allows the use of the data for only research purposes. In particular, there must be no attempt to identify the individual patients in the database. Of course, the obvious identification information like patient name, SSN, etc., are excluded from the data released by SEER, but it still has demographic information like age, sex, and race, which is very useful for research purposes but should not be misused to try to identify patients in any way. Such compliance to HIPAA regulations is important to preserve patient privacy.

**Table 1** SEER data attributes

Type	Examples
Demographic	Age, gender, location, race/ethnicity, date of diagnosis
Diagnosis	Tumor primary site, size, extension, lymph node involvement
Treatment	Primary treatment, surgical procedure, radiation therapy
Outcome	Survival time, cause of death

### Data Transformation

Once the data is available, the first step is to understand the data format and representation and do any necessary transformations to make it suitable for modeling. Let us assume the data is in a row-column (spreadsheet) format, such as in the case of SEER data. Each row corresponds to a



**Fig. 2** A typical workflow for predictive analytics, illustrated with the example of outcome prediction models for lung cancer using SEER data

patient’s medical record and can also be referred to as an instance, data point, or observation. The columns are the attributes, such as age, race, tumor size, surgery, outcome, etc. Data attributes can be of different types – numeric, nominal, ordinal, and interval – and it is important to have the correct representation of each attribute for analysis, for which some data transformation might be necessary. More broadly, data transformation is needed to ensure the quality of the data ahead of modeling and remove or appropriately deal with noise, outliers, missing values, duplicate data instances, etc.

Data transformation is usually unsupervised, which means that it does not depend on the outcome or target attributes. For example, SEER encodes all attributes as numbers, and many of them are actually nominal, like marital status, where “1” represents “Single,” “2” represents “Married,” “3” represents “Separated,” “4” represents “Divorced,” “5” represents “Widowed,” and “9” represents “Unknown.” Numbers have a natural order, and the operations of addition, subtraction, and division are defined, which may be fine for numeric attributes like “tumor size” but not for nominal attributes like marital status, sex, race, etc.,. Such attributes need to be explicitly converted to nominal for correct predictive modeling. Even numeric attributes need to be examined carefully. For example, the tumor size attribute in SEER data gives the exact size of tumor in mm, if it is known. But in some cases, the doctor notes may say “less than 2 cm,” in which case it is encoded as “992,”

which could easily be misinterpreted as 992 mm if not transformed appropriately. Another example of an unsupervised data transformation required in SEER data is to construct numeric survival time in months from the SEER format of YYMM, so that it can be modeled correctly.

The above data transformations are required due to the way SEER data is represented and may be necessary for almost any project dealing with this data. But there are also problem-specific data transformations that may be necessary for building a model as originally intended. For example, if we are interested in building a predictive model for lung cancer survival, then we should only include those patient records where the cause of patients’ death was lung cancer, which is given by the “cause of death” attribute. We also need to remove certain attributes from the modeling that directly or indirectly specify the outcome, e.g., cause of death, whether the patient is still alive. Further, for binary class prediction, we also need to derive appropriate binary attributes for survival time, e.g., 5-year survival.

There are also certain data transformation steps that could be supervised in some cases, meaning that they depend on the outcome attribute(s). Examples include feature selection/extraction, discretization, and sampling, and all of these can be supervised or unsupervised. If they are supervised, they should in general be considered together with other supervised analytics so as to avoid over-fitting (more about this later).

## Predictive Modeling

Once appropriate data transformation has been performed and the data is ready for modeling, we can employ supervised data mining techniques for feature selection and predictive modeling. Caution needs to be exercised here to appropriately split the data into training and testing sets (or use cross validation), or else the model may be subject to over-fitting and give overoptimistic accuracy. If the target attribute is numeric (e.g., survival time), regression techniques can be used for predictive modeling, and if it is categorical (e.g., whether a patient survived at least 5 years), classification techniques can be used. Some techniques are capable of doing both regression and classification. Further, there also exist several ensemble learning techniques that can combine the results from base learners in different ways and in some cases have shown to improve accuracy and robustness of the final model. Table 2 lists some of the popular predictive modeling techniques.

## Evaluation

Traditional statistical methods such as logistic regression are typically evaluated by building the model on the entire available data, and computing prediction errors on the same data, and it has been a common practice in statistical analysis of medical data as well for many years. Although this approach may work well in some cases, it is nonetheless prone to over-fitting and thus can give overoptimistic accuracy. It is easy to see that a data-driven model can, in principle, “memorize” every single instance of the dataset and thus result in 100% accuracy on the same data but will most likely not be able to work well on unseen data. For this reason, advanced data-driven techniques that usually result in black box models need to be evaluated on data that the model has not seen while training. A simple way to do this is to build the model only on random half of the data and use the remaining half for evaluation. This is called the train-test split setting for model evaluation. Further, the training and testing halves can then also be swapped for another round of

**Table 2** Popular predictive modeling algorithms

Modeling technique	Brief description
Naive Bayes	A probabilistic classifier based on Bayes theorem
Bayesian network	A graphical model that encodes probabilistic conditional relationships among variables
Logistic regression	Fits data to a sigmoidal S-shaped logistic curve
Linear regression	A linear least-squares fit of the data w.r.t. input features
Nearest neighbor	Uses the most similar instance in the training data for making predictions
Artificial neural networks	Uses hidden layer(s) of neurons to connect inputs and outputs, edge weights learnt using back propagation (called deep learning if more than two layers)
Support vector machines	Based on the structural risk minimization, constructs hyperplanes multidimensional feature space
Decision table	Constructs rules involving different combinations of attributes
Decision stump	A weak tree-based machine learning model consisting of a single-level decision tree
J48 (C4.5) decision tree	A decision tree model that identifies the splitting attribute based on information gain/gini impurity
Alternating decision tree	Tree consists of alternating prediction nodes and decision nodes, an instance traverses all applicable paths
Random tree	Considers a randomly chosen subset of attributes
Reduced-error pruning tree	Builds a tree using information gain/variance and prunes it using reduced-error pruning to avoid over-fitting
AdaBoost	Boosting can significantly reduce error rate of a weak learning algorithm
Bagging	Builds multiple models on bootstrapped training data subsets to improve model stability by reducing variance
Random subspace	Constructs multiple trees systematically by pseudo-randomly selecting subsets of features
Random forest	An ensemble of multiple random trees
Rotation forest	Generates model ensembles based on feature extraction followed by axis rotations



evaluation and the results combined to get predictions for all the instances in the dataset. This setting is called twofold cross validation, as the dataset is split into two parts. It can further be generalized to  $k$ -fold cross validation, where the dataset is randomly split into  $k$  parts.  $k - 1$  parts are used to build the model, and the remaining one part is used for testing. This process is repeated  $k$  times with different test splits, and the results are combined to get predictions for all the instances in the dataset using a model that did not see them while training. Leave-one-out cross validation (LOOCV) is a special case of the more generic  $k$ -fold cross validation, with  $k = N$ , the number of instances in the dataset. LOOCV is commonly used when the dataset is not very large. To predict the target attribute for each data instance, a separate predictive model is built using the remaining  $N - 1$  data instances, and the whole process is repeated for each data instance. The resulting  $N$  predictions can then be compared with the  $N$  actual values to calculate various quantitative metrics for accuracy. In this way, each of the  $N$  instances is tested using a model that did not see it while training, thereby maximally utilizing the available data for model building. Cross validation is a standard evaluation setting to eliminate any chances of over-fitting. Of course,  $k$ -fold cross validation necessitates building  $k$  models, which may take a long time on large datasets.

Comparative assessments of how close the models can predict the actual outcome are used to provide an evaluation of the models' predictive performance. Many binary classification performance metrics are usually used for this purpose such as accuracy, precision, recall/sensitivity, specificity, area under the ROC curve, etc.

1. **c-statistic (AUC):** The receiver operating characteristic (ROC) curve is a graphical plot of true-positive rate and false-positive rate. The area under the ROC curve (AUC or c-statistic) is one of the most effective metrics for evaluating binary classification performance, as it is independent of the probability cutoff and measures the discrimination power of the model.
2. **Overall accuracy:** It is the percentage of predictions that are correct. For highly unbalanced classes where the minority class is the class of

interest, overall accuracy by itself may not be a very useful indicator of classification performance, since even a trivial classifier that simply predicts the majority class would give high values of overall accuracy:

$$\text{Overall accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

where TP is the number of true positives (hits), TN is number of true negatives (correct rejections), FP is number of false positives (false alarms), and FN is number of false negatives (misses).

3. **Sensitivity (recall):** It is the percentage of positive labeled records that were predicted positive. Recall measures the completeness of the positive predictions:

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

4. **Specificity:** It is the percentage of negative labeled records that were predicted negative, thus measuring the completeness of the negative predictions:

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})}$$

5. **Positive predictive value (precision):** It is the percentage of positive predictions that are correct. Precision measures the correctness of positive predictions:

$$\text{Positive predictive value} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$

6. **Negative predictive value:** It is the percentage of negative predictions that are correct, thereby measuring the correctness of negative predictions:

$$\text{Negative predictive value} = \frac{\text{TN}}{(\text{TN} + \text{FN})}$$

7. **F-measure:** It is not too difficult to have a model with either good precision or good recall, at the cost of each other.  $F$ -measure

combines the two measures in a single metric such that it is high only if both precision and recall are high:

$$F - \text{measure} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}$$

### Deployment

After the predictive models have been constructed and properly evaluated, they need to be deployed appropriately to make the resulting healthcare insights available to various stakeholders at the point of care. For the lung cancer survival prediction project, the predictive models were incorporated in a web tool that allows users to enter patient attributes and get patient-specific risk values. More details about the lung cancer outcome calculator are described later in this chapter.

### Conditional Survival Prediction

Survival prediction from time of diagnosis can be very useful as we have seen till now, but for patients who have already survived a period of time since diagnosis, conditional survival is a much more clinically relevant and useful measure, as it tries to incorporate the changes in risk over time. Therefore, the above-described lung cancer survival prediction system was adapted to create additional conditional survival prediction models. Since 5-year survival rate is the most commonly used measure to estimate the prognosis of cancer, the conditional survival models were designed to estimate patient-specific risk of mortality after 5 years of diagnosis of lung cancer, given that the patient has already survived for 3 months, 6 months, 12 months, 18 months, and 24 months.

In order to construct a model for estimating mortality risk after 5 years of diagnosis of patients already survived for time  $T$ , only those patients were included in the modeling data that survived at least time  $T$ . Note that this is equivalent to taking the data used in the calculator to build 5-year survival prediction model, and removing the instances where the survival time was less than  $T$ . Thus, five new datasets were created for five different values of  $T$  (3 months, 6 months,

12 months, 18 months, and 24 months), and the same binary classification techniques were used to build five new models.

### Association Rule Mining

Association rule mining is useful to discover patterns in the data. In contrast with predictive modeling where one is interested in predicting the outcome for a given patient, here one is interested in bottom-up discovery of associations among the attributes. If a target attribute is specified, such association rule mining can help identify segments (subsets of data instances) in the data defined by specific attributes' values such that those segments have extreme average values of the target attribute. Note that this is tantamount to the inverse question of retrieval in databases, where one gives the segment definition in terms of attribute values, and the database system returns the segment, possibly along with the average value of the target attribute in that segment. However, such database retrieval cannot automatically discover segments with extreme average values of the target attribute, which is exactly what association rule mining can do. Let us take the example of the SEER dataset to make it clear. In this case, we have patient attributes including an outcome/target attribute (survival time). Let us say the average survival time in the data is  $t_{\text{avg}}$ . It would then be of interest to automatically discover from the data under what conditions – as defined by the combination of patient attribute/values – is the survival time  $t'_{\text{avg}}$  significantly greater or significantly lower than  $t_{\text{avg}}$ . Similarly, if the target attribute is nominal like 5-year survival (whether or not a patient survived for at least 5 years), and the fraction of survived patients in the entire dataset is  $f$ , then it would be interesting to find segments where this fraction  $f$  is significantly higher or lower than  $f$ .

### Illustrative Data Mining Results on SEER Data

We now present some examples of the results of above-described big data analytics on lung cancer EHR data from SEER. In Agrawal et al. (2012),

the SEER November 2008 Limited-Use Data files (SEER 2008) were used, which was released in April 2009. It had a follow-up cutoff date of December 31, 2006, i.e., the patients were diagnosed and followed up up to this date. Data was selected for the patients diagnosed between 1998 and 2001. Since the follow-up cutoff date for the SEER data in study was December 31, 2006, and the goal of the project was to predict survival up to 5 years, data of 2001 and before was used. Also, since several important attributes were introduced to the SEER data in 1998 (like RX Summ-Surg Site 98-02, RX Summ-Scope Reg 98-02, RX Summ-Surg Oth 98-02, Summary stage 2000 (1998+)), data of 1998 and after was used. There were a total of 70,132 instances of patients with cancer of the respiratory system between 1998 and 2001, and there were 118 attributes in the raw data from SEER.

The SEER-related preprocessing resulted in modification and splitting of several attributes, many of which were found to have significant predictive power. In particular, 2 out of 11 newly created (derived) attributes were within the top 13 attributes that were eventually selected to be used in the lung cancer outcome calculator. These were (a) the count of regional lymph nodes that were removed and examined by the pathologist and (b) the count of malignant/in situ tumors. These attributes were derived from “Regional Nodes Examined” and “Sequence Number-Central,” respectively, from raw SEER data, both of which had nominal values encoded within the same attribute, with the latter also encoding non-malignant tumors. After performing various steps of data transformation and feature selection, the data was reduced to 46,389 instances of lung cancer patients and 13 attributes (excluding the outcome attribute).

### Predictive Analytics

For predictive analytics, binary outcome attributes for 6-month, 9-month, 1-year, 2-year, and 5-year survival were derived from survival time. The dataset of 5-year survival was subsequently filtered to generate five new datasets for modeling conditional survival after 5 years of diagnosis,

given that the patient has already survived 3 months, 6 months, 12 months, 18 months, and 24 months.

Many predictive modeling techniques were found to give good accuracy measures that were statistically indistinguishable with the best accuracy. From among those, we chose the model based on alternating decision trees with additional logistic modeling on top for better calibration. Tenfold cross validation was used to estimate the accuracy of all the ten models. Table 3 presents the results for all the models (only accuracy and AUC included here for simplicity), along with the distribution of survived and not-survived patients in the data used to build the corresponding model.

### Association Rule Mining

For association rule mining analysis, all missing/unknown values were removed, since we are interested in finding segments with precise definitions in terms of patient attributes. The survival time (in months) was chosen as the target attribute for the Hotspot algorithm. The dataset had 13,033 instances, 13 input patient attributes, and 1 target attribute. The average survival time in the entire dataset ( $t_{avg}$ ) was 24.45 months. So it would be interesting to find segments of patients where the

**Table 3** Model classification performance (tenfold cross validation)

Model	% Survived	% Not survived	% Model accuracy	AUC
5 year	12.8	87.2	91.8	0.924
2 year	23.4	76.6	85.6	0.859
1 year	40.2	59.8	74.5	0.796
9 month	48.8	51.2	71.0	0.779
6 month	60.1	39.9	69.8	0.765
5 year  3 month	16.9	83.1	89.8	0.912
5 year  6 month	21.4	78.6	87.3	0.900
5 year  12 month	31.9	68.1	82.1	0.875
5 year  18 month	43.9	56.1	78.1	0.850
5 year  24 month	54.9	45.1	76.1	0.830

average survival time is significantly higher than or significantly lower than 24.45 months. Two independent analyses were performed to find segments in which average survival time was higher and lower than overall average survival, represented in the form of association rules. *Lift* of a rule/segment is a multiplicative metric that measures the relative improvement in the target (here survival time) as compared to the average value of the target across the entire dataset.

For association rule mining analysis on conditional survival data, a new dataset was constructed using only the cases in which the patient survived at least 12 months from the time of diagnosis. The conditional survival dataset had 6,788 instances, the same 13 input patient attributes, and 1 target

attribute. The average survival time in the conditional survival dataset was 42.54 months. So, the above analysis was repeated on the conditional survival dataset with  $t_{avg} = 42.54$ .

Tables 4 and 5 present the nonredundant association rules obtained with “higher” and “lower” mode, respectively. Tables 6 and 7 present the same for the conditional survival dataset.

### Lung Cancer Outcome Calculator

The web tool is available at <http://info.eecs.northwestern.edu/LungCancerOutcomeCalculator>, and uses the following 13 attributes:

**Table 4** Nonredundant association rules denoting segments where average survival time is significantly higher than 24.45 months

Segment description	Avg. survival time	Segment size	Lift
The tumor is well differentiated and localized, regional lymph nodes examined are between 4 and 17, age of the patient at time of diagnosis is less than 79, current tumor is patient’s first or second tumor, and resection of lobe/bilobectomy is performed by the surgeon	68.18	100	2.79
The tumor is localized, age of patient is between 39 and 52, number of regional lymph nodes examined is between 1 and 14, and resection of lobe/bilobectomy is performed by the surgeon	68.11	100	2.79
Tumor is well differentiated, number of regional lymph nodes examined is less than 15, resection of lobe/bilobectomy is performed, and regional lymph nodes are removed	66.83	101	2.73
Tumor is localized, age of patient is between 41 and 52, tumor is confined to one lung, and resection of lobe/bilobectomy is performed	66.26	111	2.71
Patient is born in Hawaii, patient’s age is less than 76, there is no lymph node involvement, and resection of lobe/bilobectomy is performed	64.98	106	2.66
Tumor is localized, patient is born in Hawaii, patient’s age is less than 83, and surgery is performed	63.96	101	2.62
Tumor is well differentiated, number of lymph nodes examined is between 7 and 18, there is no lymph node involvement, and patient’s age is less than 81	63.86	101	2.61
Tumor is localized, patient is born in Connecticut, tumor is confined to one lung, number of lymph nodes examined is greater than two, and resection of lobe/bilobectomy is performed	63.10	103	2.58
Tumor is well differentiated, there is no lymph node involvement, patient’s age is less than 76, and intrapulmonary/ipsilateral hilar/ipsilateral peribronchial nodes are removed	62.16	100	2.54
Tumor is localized (confined to one lung), patient is born in Hawaii and is less than 82 years old	60.38	101	2.47
Tumor is localized (confined to one lung), patient is born in Hawaii, and cancer is confirmed by positive histology	60.18	103	2.46
Tumor is localized, patient is born in California, and resection of lobe/bilobectomy is performed by the surgeon	58.71	100	2.40

**Table 5** Nonredundant association rules denoting segments where average survival time is significantly lower than 24.45 months

Segment description	Avg. survival time	Segment size	Lift
Tumor has metastasized and is poorly differentiated, lymph nodes are involved in metastasis, and no lymph nodes are removed	5.21	100	4.69
Tumor has metastasized and is poorly differentiated, no surgery is performed, and the patient is born in Hawaii	5.67	110	4.31
Tumor has metastasized, no surgery is performed, cancer is confirmed by positive histology, and patient is born in Hawaii	5.73	128	4.26
Tumor has metastasized, surgery is contraindicated and not performed, and cancer is confirmed by positive histology	5.78	132	4.23
Pleural effusion has taken place, tumor is poorly differentiated, subcarinal/carinal/mediastinal/tracheal/aortic/pulmonary ligament/pericardial lymph nodes are involved, and no surgery is performed	7.53	205	3.25
Pleural effusion has taken place, cancer is confirmed by positive cytology, surgery is not recommended and hence not performed	8.60	112	2.84

**Table 6** Nonredundant association rules denoting segments in the conditional survival dataset where average survival time is significantly higher than 42.54 months

Segment description	Avg. survival time	Segment size	Lift
Tumor is well differentiated and localized, patient's age is less than 71, less than 13 regional lymph nodes are examined, and resection of lobe/bilobectomy is performed	72.92	104	1.71
Tumor is well differentiated and localized (confined to one lung), patient's age is less than 71, surgery is performed, less than eight regional lymph nodes are examined	72.50	103	1.70
Tumor is well differentiated, patient's age is less than 84, regional lymph nodes are removed, no lymph node involvement, no radiation therapy, and resection of lobe/bilobectomy is performed	71.95	100	1.69
Tumor is localized (confined to one lung), patient's age is between 41 and 52, surgery is performed, and resection of lobe/bilobectomy is performed	69.66	105	1.64
Tumor is well differentiated, patient's age is less than 79, no lymph node involvement, between 5 and 9 regional lymph nodes are examined	68.44	100	1.61
Tumor is localized (confined to one lung), patient's age is less than 77, patient is born in Connecticut, and resection of lobe/bilobectomy is performed	67.99	119	1.60
Patient's age is less than 76, patient is born in Hawaii, no lymph node involvement, and resection of lobe/bilobectomy is performed	67.81	101	1.59
Patient's age is less than 75, patient is born in California, no lymph node involvement, and resection of lobe/bilobectomy is performed	65.37	102	1.54
Tumor is localized, no regional lymph nodes are removed, and resection of lobe/bilobectomy is performed	62.14	102	1.46

1. **Age at diagnosis:** Numeric age of the patient at the time of diagnosis of lung cancer.
2. **Birth place:** The place of birth of the patient. There are 198 options available to select for this attribute (based on the values observed in the SEER database).
3. **Cancer grade:** A descriptor of how the cancer cells appear and how fast they may grow and spread. Available options are well-differentiated, moderately differentiated, poorly differentiated, undifferentiated, and undetermined.
4. **Diagnostic confirmation:** The best method used to confirm the presence of lung cancer. Available options are positive histology, positive cytology, positive microscopic

**Table 7** Nonredundant association rules denoting segments in the conditional survival dataset where average survival time is significantly less than 42.54 months

Segment description	Avg. survival time	Segment size	Lift
Tumor is undifferentiated and has metastasized, subcarinal/carinal/mediastinal/tracheal/aortic/pulmonary ligament/pericardial lymph nodes are involved, no regional lymph nodes are removed, and no surgery is performed	17.18	100	2.48
Tumor is spread, surgery not recommended, patient is born in Iowa	20.28	137	2.10
Tumor is spread and undifferentiated, surgery not recommended, subcarinal/carinal/mediastinal/tracheal/aortic/pulmonary ligament/pericardial lymph nodes are involved, and cancer is confirmed by positive histology	20.35	124	2.09
Pleural effusion has taken place, and tumor is poorly differentiated	22.96	101	1.85

confirmation (method unspecified), positive laboratory test/marker study, direct visualization, radiology, other clinical diagnosis, and unknown if microscopically confirmed.

5. **Farthest extension of tumor:** The farthest documented extension of tumor away from the lung, either by contiguous extension (regional growth) or distant metastases (cancer spreading to other organs far from primary site through bloodstream or lymphatic system). There are 20 options available to select for this attribute. The original SEER name for this attribute is "EOD extension."
6. **Lymph node involvement:** The highest specific lymph node chain that is involved by the tumor. Cancer cells can spread to lymph nodes near the lung, which are part of the lymphatic system (the system that produces, stores, and carries the infection-fighting cells). This can often lead to metastases. There are eight options available for this attribute. The original SEER name for this attribute is "EOD Lymph Node Involv."
7. **Type of surgery performed:** The surgical procedure that removes and/or destroys cancerous tissue of the lung, performed as part of the initial work-up or first course of therapy. There are 25 options available for this attribute, like cryosurgery, fulguration, wedge resection, laser excision, pneumonectomy, etc. The original SEER name for this attribute is "RX Summ-Surg Prim Site."
8. **Reason for no surgery:** The reason why surgery was not performed (if not). Available options are surgery performed, surgery not recommended, contraindicated due to other conditions, unknown reason, patient or patient's guardian refused, recommended but unknown if done, and unknown if surgery performed.
9. **Order of surgery and radiation therapy:** The order in which surgery and radiation therapies were administered for those patients who had both surgery and radiation. Available options are no radiation and/or surgery, radiation before surgery, radiation after surgery, radiation both before and after surgery, intraoperative radiation therapy, intraoperative radiation with other radiation given before/after surgery, and sequence unknown but both surgery and radiation were given. The original SEER name for this attribute is "RX Summ-Surg/Rad Seq."
10. **Scope of regional lymph node surgery:** It describes the removal, biopsy, or aspiration of regional lymph node(s) at the time of surgery of the primary site or during a separate surgical event. There are eight options available for this attribute. The original SEER name for this attribute is "RX Summ-Scope Reg 98-02."
11. **Cancer stage:** A descriptor of the extent to which the cancer has spread, taking into account the size of the tumor, depth of penetration, metastasis, etc. Available options are in situ (noninvasive neoplasm), localized (invasive neoplasm confined to the lung), regional (extended neoplasm), distant (spread neoplasm), and unstaged/unknown. The original SEER name for this attribute is "Summary Stage 2000 (1998+)."

12. **Number of malignant tumors in the past:** An integer denoting the number of malignant tumors in the patient’s lifetime so far. This attribute is derived from the SEER attribute “Sequence Number-Central,” which encodes both numeric and categorical values for both malignant and benign tumors within a single attribute. As part of the preprocessing, the original SEER attribute was split into numeric and nominal parts, and the numeric part was further split into two attributes representing number of malignant and benign tumors, respectively.
13. **Total regional lymph nodes examined:** An integer denoting the total number of regional lymph nodes that were removed and examined by the pathologist. This attribute was derived by extracting the numeric part of the SEER attribute “Regional Nodes Examined.”

Figure 3 shows a screenshot of the lung cancer outcome calculator. This calculator is widely accessed from more than 15 countries, including many medical schools and hospitals. A previous version of this calculator was presented in Agrawal et al. (2011a). The current calculator incorporates faster models as described in this chapter and has a redesigned interface. It allows the user to enter values for the above-described 13 attributes and get patient-specific risk. For all the ten models, it also shows the distribution of survived and not-survived patients in the form of pie charts. Upon entering the patient attributes on the website, the patient-specific risk calculated by all the ten models is depicted along with the healthy and sick patient risk, which are essentially the median risk of death of patients who actually survived and did not survive, respectively, as calculated by the corresponding model. It generates bar charts corresponding to each of the ten models, and each of them has three bars. The middle bar denotes the patient-specific risk, and the left (right) bars denote the healthy (sick) patient risk. The patient-specific risk is thus put in context of the healthy and sick patient risk for an informative comparison.

Any data-driven tool like this in the field of healthcare has a disclaimer about its use, stating

that it is meant to complement and not replace the advice of a medical doctor. Many such calculators are becoming popular in healthcare.

### Other Applications of Big Data Analytics in Healthcare

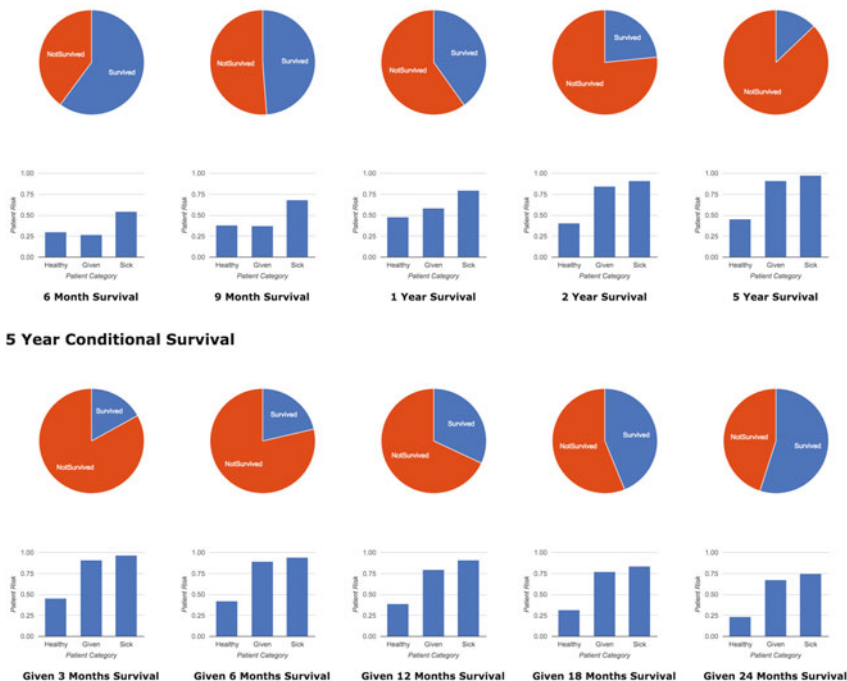
We will conclude with a sampling of some other applications of big data in healthcare. There has been abundant work on mining electronic health records in addition to what is described in this chapter. Some of these include mining data from a particular hospital (Mathias et al. 2013), American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) (Agrawal et al. 2013b), and United Network for Organ Sharing (UNOS (Agrawal et al. 2013c).

Apart from electronic health records, a very important source of healthcare data is social media. We are in the midst of a revolution in which, using social media, people interact, communicate, learn, influence, and make decisions. This data includes multi-way communications and interactions on social media (e.g., Facebook, Twitter), discussion forums, and blogs in the area of healthcare, public health, and medicine. The emergence and ubiquity of online social networks have enriched this data with evolving interactions and communities at mega-scale, and people are turning to social media for various kinds of healthcare guidance and knowledge, including proactive and preventive care. Patients with like conditions – often chronic conditions, such as flu, cancer, allergy, multiple sclerosis, diabetes, arthritis, ALS, etc. – find patients with the same condition on these networking sites and in public forums. And these virtual peers can very much become a key guiding source of data unlike in the past, when all information emanated from physicians. This big data, being produced in social media domain, offers a unique opportunity for advancing, studying the interaction between society and medicine, managing diseases, learning best practices, influencing policies, identifying best treatment, and, in general, empowering people. It thus has numerous applications in public health informatics, and we are already seeing

# Lung Cancer Outcome Calculator

**Disclaimer:** The outcome calculator results are estimates based on data consisting of a large number of lung cancer records. All results are provided for informational purposes only, in furtherance of the developers' educational mission, and are not meant to replace the advice of a medical doctor. The developers may not be held responsible for any medical decisions based on this outcome calculator.

Welcome to our online lung cancer outcome calculator. The calculator is based on data obtained from Surveillance Epidemiology and End Results (SEER) of the National Cancer Institute which is an authoritative repository of cancer statistics in the United States. The data contains lung cancer records of nearly 50000 patients. The calculator estimates the risk of mortality after 6 months, 9 months, 1 year, 2 year, and 5 years of diagnosis, using a small non-redundant subset of 13 patient attributes which were carefully selected using attribute selection techniques. The graph shows the five risk values obtained for specific attribute values, which are shown below the graph. To obtain risk values for a new set of attribute values, please change the attribute values below and click on the submit button.



For a given time interval  $T$ ,  
**Healthy patient risk** - Median risk of death of patients who survived after time  $T$ , as calculated by our calculator.  
**Patient risk** - This corresponds to the risk of death of a patient after time  $T$ , calculated based on the provided values of the patient attributes.  
**Sick patient risk** - Median risk of death of patients who did not survive after time  $T$ , as calculated by our calculator.

Age	75	Reason for no surgery	Surgery performed
Birth Place	Afghanistan	Order of surgery and radiation therapy	No radiation and/or surgery
Cancer grade	Grade I (well-differentiated)	Scope of regional lymph node surgery	No regional lymph nodes removed
Diagnostic confirmation	Positive histology	Cancer stage	In situ (Noninvasive neoplasm)
Farthest extension of tumour	In situ (Noninvasive/intraepithelial)	Number of malignant tumors in the past	0
Lymph node involvement	No lymph node involvement	Total regional lymph nodes examined:	0
Surgery of primary site	No surgery		

Developed by [Ankit Agrawal](#) and [Alok Choudhary](#)

**Publications:**

A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "Lung cancer survival prediction using ensemble data mining on SEER data," Scientific Programming, vol. 20, no. 1, pp. 29-42, 2012. [url]  
 A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "A lung cancer outcome calculator using ensemble data mining on SEER data," In Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics (BIOKDD), 2011, pp. 1-9. [url]  
[Center for Ultra-scale Computing and Information Security \(CUCIS\)](#), EECS Department, Northwestern University, Evanston, IL 60208, USA

**Fig. 3** Screenshot of the lung cancer outcome calculator. (Available at <http://info.eecs.northwestern.edu/LungCancerOutcomeCalculator>)



several studies in this domain (Lee et al. 2013, 2015; Xie et al. 2013).

Technological advances in sensors, micro- and nano-electronics, advanced materials, mobile computing, etc., have had an immense impact toward enabling future Internet of things (IoT) applications in several fields including healthcare. We are currently witnessing a rapid adoption of wearable devices under the IoT paradigm for a variety of healthcare applications (Andreu-Perez et al. 2015). These wearable and implantable sensors along with smartphones that are ubiquitously used all over the world form another source of healthcare big data and provide unprecedented opportunities for continuous healthcare monitoring and management.

The field of genomics is another area where big data analytics can play an important role. It is well recognized that in genomics and life sciences, almost everything is based on complex sequence-structure-function relationships, which are far from being well understood. With genomic sequencing becoming progressively easier and affordable, we have arrived at a point in time where huge amounts of biological sequence data have become increasingly available, thanks to the advent of next-generation sequencing (NGS). Functional interpretation of genomic data is the major task in fundamental life science. Research results in this area in turn feed research in other important areas such as cell biology, genetics, immunology, and disease-oriented fields. There has been a lot of work in bioinformatics on sequence data in terms of computationally mining the genomic sequences for interesting insights such as homology detection (Agrawal and Huang 2009, 2011). Furthermore, biological sequencing data also ushers an era of personal genomics enabling individuals to have their personal DNA sequenced and studied to allow more precise and personalized ways of anticipating, diagnosing, and treating diseases on an individual basis (precision medicine). Genome assembly and sequence mapping techniques (Huang and Madan 1999; Misra et al. 2011) form the first step of this process by compiling the overlapping reads into a single genome. While it is a fact that personalized medicine is becoming more and more common, it

is nonetheless in its infancy, and we are still far from realizing the dream of personalized medicine by optimally utilizing the flood of genomic data that we are able to collect now. Clearly, computational sequence analysis techniques are critical to unearth the hidden knowledge from such genomic sequence data, and big data analytics is expected to play a big role in that. For further reading on big data analytics in genomics, the following articles are recommended (Howe et al. 2008; ODriscoll et al. 2013; Marx 2013).

---

## Summary

Big data has become a very popular term denoting huge volumes of complex datasets generated from various sources at a rapid rate. This big data potentially has immense hidden value that needs to be discovered by means of intelligently designed analysis methodologies that can scale for big data and all of that falls in the scope of big data analytics. In this chapter, we have looked at some of the big data challenges in general and also what they mean in context of healthcare. As an example on big data mining in healthcare, some recent works dealing with the use of predictive analytics and association rule mining on lung cancer data from SEER were discussed, including a lung cancer outcome calculator that has been deployed as a result of this analytics. Finally, we also briefly looked at a few other healthcare-related areas where big data analytics is playing an increasingly vital role.

---

## References

- Agrawal A, Choudhary A. Association rule mining based hotspot analysis on seer lung cancer data. *Int J Knowl Discov Bioinform (IJKDB)*. 2011a;2(2):34–54.
- Agrawal A, Choudhary A. Identifying hotspots in lung cancer data using association rule mining. In: 2nd IEEE ICDM workshop on biological data mining and its applications in healthcare (BioDM); 2011b. p. 995–1002.
- Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the fourth paradigm of science in materials science. *APL Mater*. 2016;4 (053208):1–10.

- Agrawal A, Huang X. Psiblast pairwisestat: reordering psi-blast hits using pairwise statistical significance. *Bioinformatics*. 2009;25(8):1082–3.
- Agrawal A, Huang X. Pairwise statistical significance of local sequence alignment using sequence-specific and position-specific substitution matrices. *IEEE/ACM Trans Comput Biol Bioinformatics*. 2011;8(1):194–205.
- Agrawal A, Misra S, Narayanan R, Polepeddi L, Choudhary A. A lung cancer outcome calculator using ensemble data mining on seer data. In: Proceedings of the tenth international workshop on data mining in bioinformatics (BIOKDD), New York: ACM; 2011. p. 1–9.
- Agrawal A, Misra S, Narayanan R, Polepeddi L, Choudhary A. Lung cancer survival prediction using ensemble data mining on seer data. *Sci Program*. 2012;20(1):29–42.
- Agrawal A, Patwary M, Hendrix W, Liao WK, Choudhary A. High performance big data clustering. IOS Press; 2013a. p. 192–211.
- Agrawal A, Al-Bahrani R, Merkow R, Bilimoria K, Choudhary A. “Colon surgery outcome prediction using acs nsqip data.” In: Proceedings of the KDD workshop on Data Mining for Healthcare (DMH); 2013b. p. 1–6.
- Agrawal A, Al-Bahrani R, Raman J, Russo MJ, Choudhary A. Lung transplant outcome prediction using unos data. In: Proceedings of the IEEE big data workshop on Bioinformatics and Health Informatics (BHI); 2013c. p. 1–8.
- Andreu-Perez J, Leff DR, Ip H, Yang G-Z. From wearable sensors to smart implants – toward pervasive and personalized healthcare. *IEEE Trans Biomed Eng*. 2015;62(12):2750–62.
- Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *Ann Intern Med*. 2015;162(1):55–63.
- Ganguly AR, Kodra E, Agrawal A, Banerjee A, Boriah S, Chatterjee S, Chatterjee S, Choudhary A, Das D, Faghmous J, Ganguli P, Ghosh S, Hayhoe K, Hays C, Hendrix W, Fu Q, Kawale J, Kumar D, Kumar V, Liao WK, Liess S, Mawalagedara R, Mithal V, Oglesby R, Salvi K, Snyder PK, Steinhaeuser K, Wang D, Wuebbles D. Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques. *Nonlinear Process Geophys*. 2014;21:777–95.
- Hey T, Tansley S, Tolle K, editors. The fourth paradigm: data-intensive scientific discovery. Redmond: Microsoft Research; 2009.
- Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, Pierre SS, et al. Big data: the future of biocuration. *Nature*. 2008;455(7209):47–50.
- Huang X, Madan A. Cap3: a dna sequence assembly program. *Genome Res*. 1999;9(9):868–77.
- Lee K, Agrawal A, Choudhary A. Real-time disease surveillance using twitter data: demonstration on flu and cancer. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD); 2013. p. 1474–77.
- Lee K, Agrawal A, Choudhary A. Mining social media streams to improve public health allergy surveillance. In: Proceedings of IEEE/ACM international conference on Social Networks Analysis and Mining (ASONAM); 2015. p. 815–22.
- Magill SS, Edwards JR, Bamberg W, Beldavs ZG, Dumyati G, Kainer MA, Lynfield R, Maloney M, McAllister-Holod L, Nadle J, Ray SM, Thompson DL, Wilson LE, Fridkin SK. Multistate point-prevalence survey of health care-associated infections. *N Engl J Med*. 2014;370(13):1198–208.
- Marx V. Biology: the big challenges of big data. *Nature*. 2013;498(7453):255–60.
- Mathias JS, Agrawal A, Feinglass J, Cooper AJ, Baker DW, Choudhary A. Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. *J Am Med Inform Assoc*. 2013;20:e118–24. JSM and AA are co-first authors.
- Misra S, Agrawal A, Liao W-k, Choudhary A. Anatomy of a hash-based long read sequence mapping algorithm for next generation dna sequencing. *Bioinformatics*. 2011;27(2):189–95.
- ODriscoll A, Daugelaitė J, Sleator RD. Big data, hadoop and cloud computing in genomics. *J Biomed Inform*. 2013;46(5):774–81.
- Ries LAG, Eisner MP. Cancer of the lung. In: Ries LAG, Young JL, Keel GE, Eisner MP, Lin YD, Horner M-J, eds. SEER survival monograph: Cancer survival among adults: U.S. SEER program, 1988–2001, Patient and Tumor Characteristics. NIH Pub. No. 07–6215. Bethesda, Md: National Cancer Institute, SEER Program; 2007:73–80.
- SEER, Surveillance, epidemiology, and end results (seer) program ([www.seer.cancer.gov](http://www.seer.cancer.gov)) limited-use data (1973–2006). National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch; 2008. Released April 2009, based on the November 2008 submission.
- Xie Y, Honbo D, Choudhary A, Zhang K, Cheng Y, Agrawal A. Voxsup: a social engagement framework. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD) (Demo paper). ACM; 2012. p. 1556–9.
- Xie Y, Chen Z, Zhang K, Cheng Y, Honbo DK, Agrawal A, Choudhary A. Muses: a multilingual sentiment elicitation system for social media data. *IEEE Intell Syst*. 2013a;99:1541–672.
- Xie Y, Chen Z, Cheng Y, Zhang K, Agrawal A, WK Liao, Choudhary A. Detecting and tracking disease outbreaks by mining social media data. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI); 2013b. p. 2958–60.
- Xie Y, Palsetia D, Trajcevski G, Agrawal A, Choudhary A. Silverback: scalable association mining for temporal data in columnar probabilistic databases. In: Proceedings of 30th IEEE International Conference on Data Engineering (ICDE), Industrial and Applications Track; 2014. p. 1072–83.



# Health Services Data: Managing the Data Warehouse: 25 Years of Experience at the Manitoba Centre for Health Policy

# 2

Mark Smith, Leslie L. Roos, Charles Burchill, Ken Turner, Dave G. Towns, Say P. Hong, Jessica S. Jarmasz, Patricia J. Martens, Noralou P. Roos, Tyler Ostapyk, Joshua Ginter, Greg Finlayson, Lisa M. Lix, Marni Brownell, Mahmoud Azimae, Ruth-Ann Soodeen, and J. Patrick Nicol

## Contents

<b>Introduction</b> .....	21
Who We Are .....	21
What We Do .....	21
Our Data Is Our Strength .....	21
<b>Privacy</b> .....	24
<b>Repository Tools</b> .....	25
Glossary .....	25
Concept Dictionary .....	25
<b>Characteristics of Administrative Data</b> .....	26
<b>Data Documentation</b> .....	27
<b>Applying for Access</b> .....	30
Repository Documentation .....	31
<b>The Data Management Process</b> .....	31
Step 1: Formulate the Request and Receive the Data .....	31
Step 2: Become Familiar with the Data Structure and Content .....	33
Step 3: Apply SAS® Programs .....	34
Step 4: Evaluate Data Quality .....	35
Step 5: Document the Data .....	36

Patricia J. Martens: deceased.

M. Smith (✉) · L. L. Roos · C. Burchill · K. Turner · D. G. Towns · S. P. Hong · J. S. Jarmasz · N. P. Roos · M. Brownell · R.-A. Soodeen · J. P. Nicol  
Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada  
e-mail: [Mark\\_Smith@cpe.umanitoba.ca](mailto:Mark_Smith@cpe.umanitoba.ca);  
[Leslie\\_Roos@cpe.umanitoba.ca](mailto:Leslie_Roos@cpe.umanitoba.ca); [Charles\\_Burchill@cpe.umanitoba.ca](mailto:Charles_Burchill@cpe.umanitoba.ca); [Ken\\_Turner@cpe.umanitoba.ca](mailto:Ken_Turner@cpe.umanitoba.ca);  
[Dave\\_Towns@cpe.umanitoba.ca](mailto:Dave_Towns@cpe.umanitoba.ca); [Say\\_PhamHong@cpe.umanitoba.ca](mailto:Say_PhamHong@cpe.umanitoba.ca);  
[Jessica\\_Jarmasz@cpe.umanitoba.ca](mailto:Jessica_Jarmasz@cpe.umanitoba.ca);  
[Noralou\\_Roos@cpe.umanitoba.ca](mailto:Noralou_Roos@cpe.umanitoba.ca);  
[Marni\\_Brownell@cpe.umanitoba.ca](mailto:Marni_Brownell@cpe.umanitoba.ca);  
[Ruth-Ann\\_Soodeen@cpe.umanitoba.ca](mailto:Ruth-Ann_Soodeen@cpe.umanitoba.ca)

Step 6: Release the Data .....	36
Percent of Time Spent on Each Data Management Activity .....	36
Summary .....	37
<b>Data Quality Evaluation Tool for Administration Data .....</b>	<b>37</b>
Completeness and Correctness .....	37
Assessing Consistency .....	37
Referential Integrity .....	38
Trend Analysis .....	38
Assessing Agreement .....	40
Assessing Crosswalk Linking .....	40
Summary .....	40
<b>Advantages of Using a Population-Based Registry .....</b>	<b>41</b>
Expanding Capabilities into Social Policy Research .....	42
Using Place-of-Residence Data .....	42
Constructing Reliable Social Measures .....	43
Identifying Siblings and Twins .....	43
Beyond Health Research .....	43
<b>Summing Up .....</b>	<b>44</b>
<b>References .....</b>	<b>44</b>

## Abstract

The “Data Repository” at the Manitoba Centre for Health Policy (MCHP) is a cornerstone of the organization and one of the three “pillars” on which it stands (the other two being Research Program and Knowledge Translation). For 25 years, MCHP has maintained

one of the most extensive collections of government administrative, survey, and clinical data holdings in the world, including everything from hospital and medical claims to child welfare services and educational enrolment and outcomes. Over 70 different government and clinical databases flow into the organization on an annual basis. This chapter outlines how the data are collected, organized, documented, managed, and accessed in a privacy protecting fashion for use by researchers in Canada, North America, and around the world. The research conducted by MCHP, which is located in the Rady Faculty of Health Sciences at the University of Manitoba, in addition to being relevant to policy and government decision makers is regularly published in leading academic journals. The chapter concludes with a discussion of the relative strengths of using a population-based longitudinal registry and some of the challenges faced in organizing and using available data for research purposes.

P. J. Martens  
Winnipeg, MB, Canada

T. Ostapyk  
University Advancement, Carleton University, Ottawa,  
ON, Canada  
e-mail: [Tyler.Ostapyk@Carleton.ca](mailto:Tyler.Ostapyk@Carleton.ca)

J. Ginter  
Montreal, QC, Canada  
e-mail: [joshua.ginter@gmail.com](mailto:joshua.ginter@gmail.com)

G. Finlayson  
Finlayson and Associates Consulting, Kingston, ON,  
Canada  
e-mail: [finlayson.consulting@cogeco.ca](mailto:finlayson.consulting@cogeco.ca)

L. M. Lix  
Department of Community Health Sciences, University of  
Manitoba, Winnipeg, MB, Canada  
e-mail: [lisa.lix@umanitoba.ca](mailto:lisa.lix@umanitoba.ca)

M. Azimae  
ICES Central, Toronto, ON, Canada  
e-mail: [Mahmoud.Azimae@ices.on.ca](mailto:Mahmoud.Azimae@ices.on.ca)

## Introduction

### Who We Are

The Manitoba Centre for Health Policy (MCHP) is a research organization located within the Department of Community Health Sciences, Max Rady College of Medicine, Rady Faculty of Health Sciences, at the University of Manitoba (see Fig. 1). MCHP maintains the unique Population Health Research Data Repository (the Repository) that is used by researchers to describe and explain patterns of health care as well as profiles of illness, and to explore other factors that influence health such as socioeconomic status (income, education, employment, social status, etc.). This chapter provides an overview of MCHP, concentrating on the acquisition and preparation of data, and the management of the Repository to support research and to protect the privacy and confidentiality of Manitobans. The chapter that follows concentrates on MCHP's research production as well as the policy and program impacts of those products over the past 25 years.

### What We Do

MCHP's mission is to conduct world-class population-based research to support the development of evidence-informed policy, programs, and services that maintain and improve the health and well-being of Manitobans (see Fig. 2).

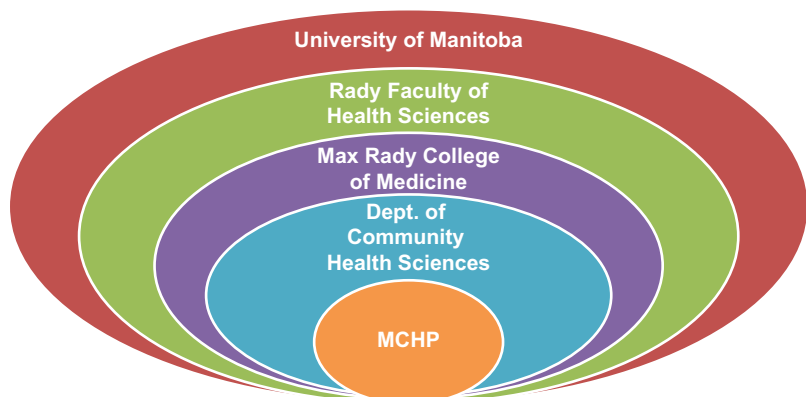
## Our Data Is Our Strength

MCHP was the first research unit of its kind in Canada. It continues to be recognized for its comprehensive and ever-expanding, linkable population-based data repository; its collaborative models of working with government and health regions; and for the outstanding caliber of its research (Jutte et al. 2011; Wolfson 2011). The Repository (see Fig. 3) is unique in terms of its comprehensiveness, degree of integration, and orientation around an anonymized population registry.

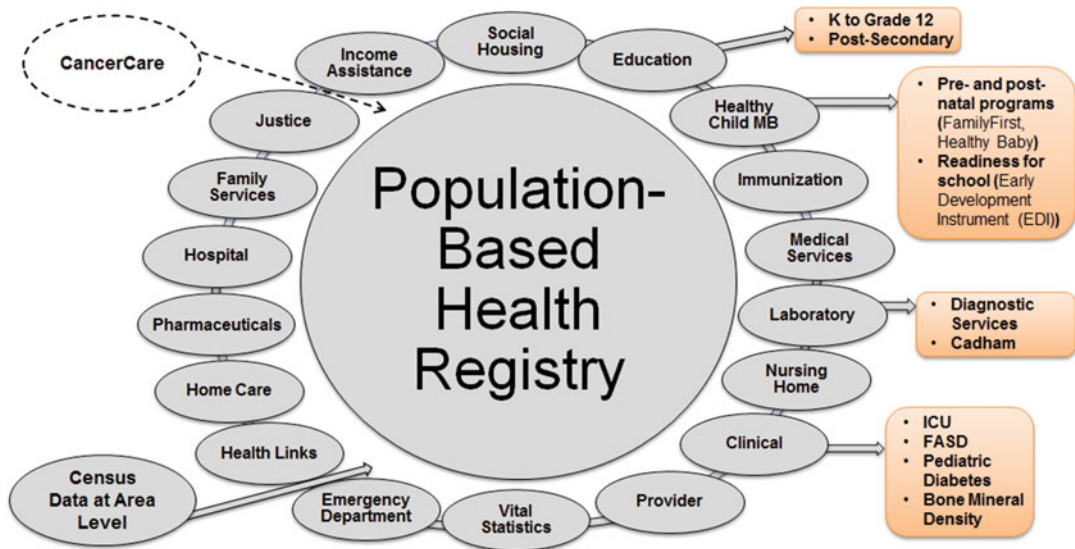
All the data files in the Repository are de-identified: names, addresses, phone numbers, and real personal health identification numbers (PHINs) are removed before files are transferred to MCHP by the data provider. MCHP complies with all laws and regulations governing the protection and use of personal information. Strict policies and procedures are implemented to protect the privacy and security of anonymized data.

Information in the Repository comes mainly from Manitoba Health and other provincial government departments. The ability to link files and track healthcare use from more than 70 databases, some of which include data as far back as 1970, allows researchers to investigate the health of Manitobans across a wide spectrum of indicators. The data can tell us about Manitobans' visits to the doctor, hospital stays, home care and nursing-home use, pharmaceutical prescriptions, etc. It is continually expanding into new areas such as education (kindergarten through grade 12 and

**Fig. 1** MCHP's location within the University of Manitoba



**Fig. 2** MCHP’s mission, goals, vision, and values



**Fig. 3** The Population Health Research Data Repository Data not yet a part of the registry but is currently being acquired is represented by a dotted line

some post-secondary), social housing, laboratory diagnostic information, in-hospital pharmaceuticals, and justice. Additional area-level data such as the Canadian census indicator of average household income, available from Statistics

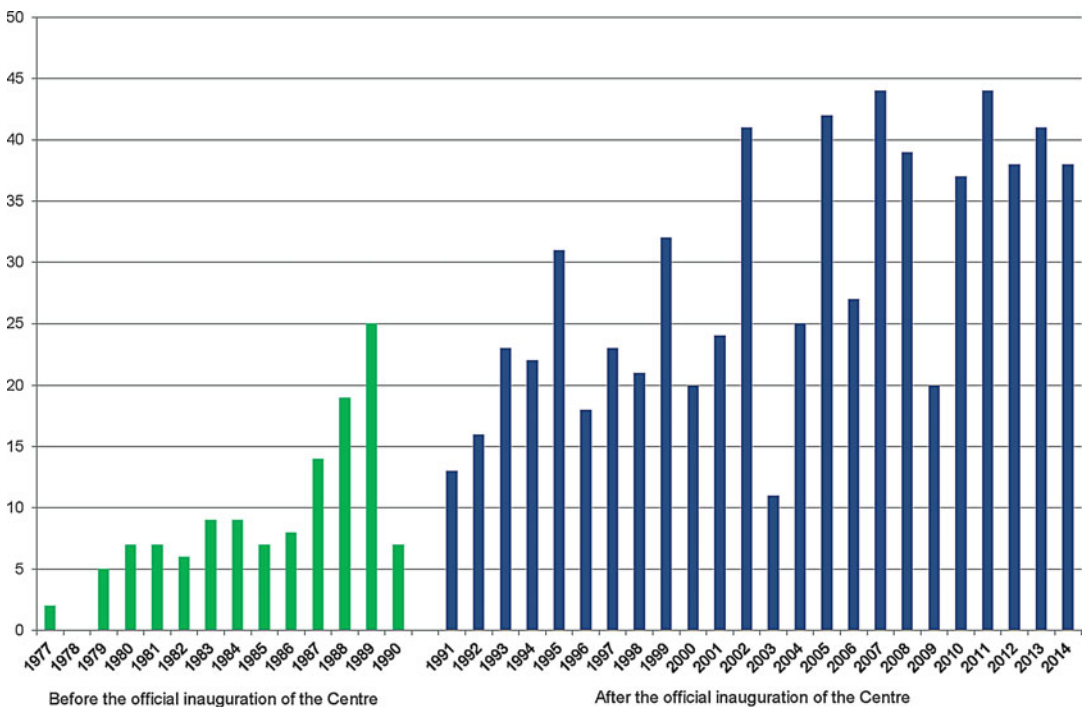
Canada through the Data Liberation Initiative, are also stored in the repository and available for linkage by postal code.

Some examples of how the data in the repository have been used in the past include:

- MCHP’s research into Manitoba’s aging population has helped estimate future needs for nursing-home beds, so regions can begin strategically to add services over the coming decades (Chateau et al. 2012).
- The results from MCHP’s report *Population Aging and the Continuum of Older Adult Care in Manitoba*, published in February 2011, were used by the Manitoba Government to invest \$216 million to add more home care support; a new rehabilitation program for seniors after surgery; as well as new personal care homes (Doupe et al. 2011).
- A report released in 2010 found that women enrolled in Manitoba’s Healthy Baby Prenatal Benefit program had fewer low birth weight babies and fewer preterm births among other measurable improvements, lending substantial support for the program (Brownell et al. 2010).
- Other MCHP reports document comparative health status and the use of health and social

services for groups such as Manitoba’s Franco-phone and Métis populations or for individual regional health authorities (RHAs) (Chartier et al. 2012; Fransoo et al. 2013; Martens et al. 2010).

MCHP personnel interact extensively with government officials, healthcare administrators, and clinicians to develop a topical and relevant research agenda. The strength of these interactions, along with the application of rigorous academic standards, enables MCHP to make significant contributions to the development of health and social policy. MCHP undertakes five major research projects every year under contract with Manitoba Health. In addition, MCHP investigators secure external funding by competing for research grants. Research completed at MCHP is widely published and internationally recognized (see Fig. 4). MCHP researchers collaborate with a number of highly respected scientists from Canada, the United States, Europe, and Australia.



**Fig. 4** Number of documented publications in peer-reviewed journals arising from the use of MCHP Data, 1977–2014

### Privacy

Ensuring privacy and confidentiality of data regarding individuals is a priority. MCHP protects data against loss, destruction, or unauthorized use. The data MCHP receives is de-identified so researchers and data analysts never know the identity of the individuals in the data. A detailed process has been developed whereby information from trustees can be transferred to MCHP in de-identified and scrambled form (see Fig. 5). Our principles and procedures for ensuring confidentiality go beyond using de-identified data. As a custodian of sensitive information, MCHP adheres to the rules for privacy and protection of personal information outlined in the province’s Freedom of Information and Protection of Privacy Act (FIPPA) and the Personal Health Information Act (PHIA). MCHP implements many security safeguards in its data network, including restricted access, two-factor authentication, and file encryption. Every project requires review and approval from the University of Manitoba’s Health Research Ethics Board (HREB), the Health Information Privacy Committee (HIPC), and relevant data providers. MCHP’s commitment to privacy also includes mandatory accreditation sessions for

everyone who has access to MCHP data (see **Applying for Access** below for details).

Under PHIA entities that collect data are called trustees. Briefly, demographic data (identifying/personal information) – including items such as name, address, and phone number – and an internal reference number are sent from the trustee to Manitoba Health, where the identifying information is used to lookup or verify an existing PHIN for each client. This process involves deterministic and probabilistic data linkage. The PHIN is then encrypted and attached to each record, and the identifying information is removed. At the same time, the trustee sends the reference number and the program data to MCHP. When the encrypted PHIN is received by MCHP, the reference number is used to link it to the program data (Fig. 5). Consequently, no single organization has all of the pieces of the linkage puzzle: the trustee does not have access to the scrambled PHIN, Manitoba Health does not have access to the program data, and MCHP does not have access to the identifying information.

At MCHP, files are stored separately until all approvals for a project are received and then they are linked. Once a research project is complete, the code and data are retained for up to 7 years, but

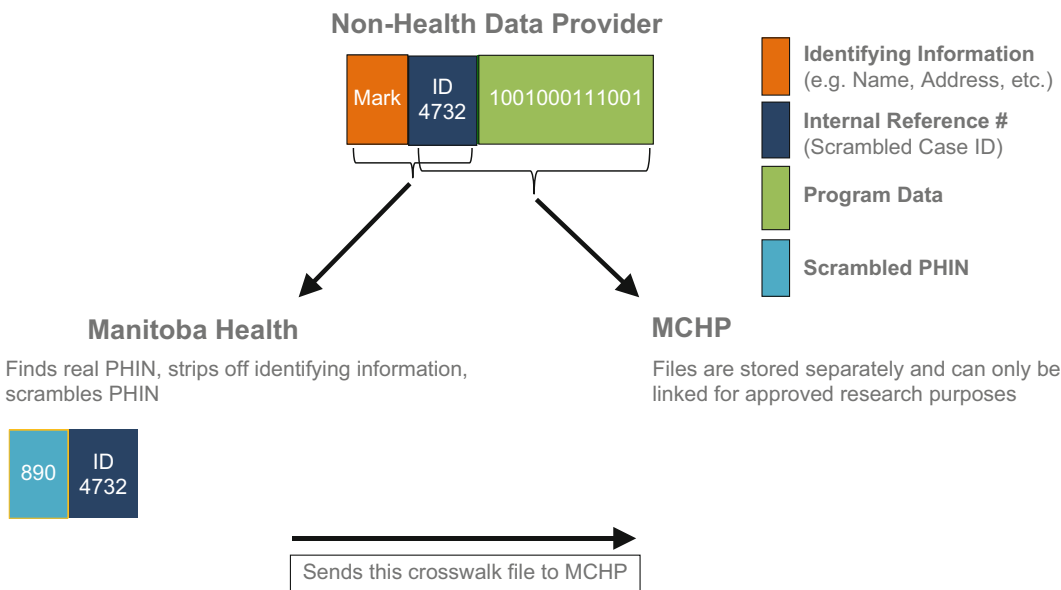


Fig. 5 De-identification process diagram



the archived data cannot be accessed without appropriate approvals.

MCHP also implements small number disclosure control. Non-zero values that are less than six are suppressed in final reports. This helps to ensure that the privacy and confidentiality of individuals is retained while allowing individual level data to be used for research purposes.

---

## Repository Tools

MCHP has developed a number of web-based resources that document the historical use of information stored in the repository. Much of this “corporate knowledge” is captured in two resources: the MCHP Glossary and the Concept Dictionary.

## Glossary

The MCHP Glossary is a compilation of short definitions for key terms used in MCHP publications. It documents terms commonly used in population health and health services research and consists of over 2,300 entries. Each glossary term contains a brief definition (and its source), links to related entries in the glossary and concept dictionary, and links to pertinent external sites and reports.

## Concept Dictionary

The MCHP Concept Dictionary contains detailed operational definitions and SAS<sup>®</sup> program code for variables or measures developed from administrative data. Because data are often complicated to work with and government decisions about definition, collection, and availability of data can change over time, having these resources available helps to communicate historical learning and reduce the probability of future error. Some examples of the many types of concepts that have been developed include:

- **Health:** Charlson Comorbidity Index, Suicide and Attempted Suicide, the John Hopkins

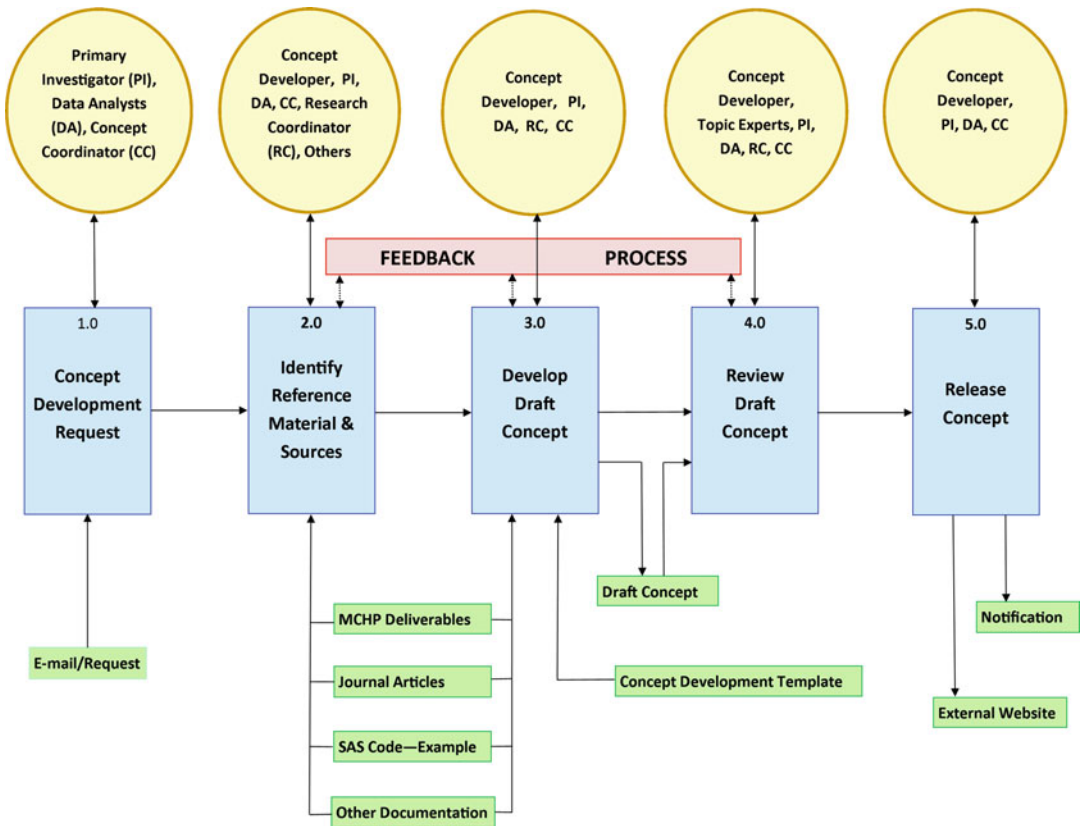
Adjusted Clinical Group<sup>®</sup> (ACG<sup>®</sup>) Case-Mix System, Complications and Comorbidities, Teenage Pregnancy, Diagnoses and Procedures

- **Education:** High School Completion, Indices of Educational Achievement, Curriculum Level
- **Statistics:** Intra-class Correlation Coefficient (ICC), Sensitivity and Specificity, Prevalence and Incidence, General Estimating Equations (GEE)
- **Data Management:** Record Linkage, Common Exclusions, Duplicate Records
- **Geographic Analysis:** Regional Health Authorities (RHAs), Winnipeg Community Areas (WCAs)
- **Costing:** Hospital Stays, Prescriptions, Physician Services, Home Care
- **Socioeconomic Status:** Income Quintiles, Socioeconomic Factor Index (SEFI)
- **Social:** Family Structure, Income Assistance (IA), Residential Mobility

Developing new concepts involves collaboration between the research team, a concept developer, and the Concept Dictionary Coordinator. As shown in Fig. 6, the process involves five steps: (1) A request for the development (or update) of a concept; (2) Identification of reference materials and sources; (3) Development of a draft; (4) Review of the draft involving feedback and revisions; and (5) Publication of the concept on the MCHP website.

The contents of a concept typically include:

- A descriptive title
- The date it was first available or last updated
- An introduction to the topic
- A description of the methods used, including data sources, background information, steps in developing the concept, validation of the process, and hyperlinks to additional information such as findings and results in publications
- The concept’s limitations or any cautions related to its use
- An example of the SAS<sup>®</sup> code and/or formats associated with the concept
- Links to related terms in the MCHP Glossary or Concept Dictionary



**Fig. 6** Concept development process

- Links to additional supporting material (both internal and external), and a list of references for the concept

An example concept is listed below in Fig. 7

The MCHP Glossary and Concept Dictionary are available on-line at: [http://umanitoba.ca/faculties/health\\_sciences/medicine/units/community\\_health\\_sciences/departamental\\_units/mchp/resources/concept\\_dictionary.html](http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departamental_units/mchp/resources/concept_dictionary.html)

### Characteristics of Administrative Data

Because administrative data are collected primarily for purposes other than research, care is required to ensure accurate results. Potential limitations include clinically imprecise coding, absence of

key data on processes and outcomes, and the inconsistent recording of provider information. On the other hand, the administrative data housed in the Repository yields a number of advantages for conducting high-quality research, including:

- **Population based:** The entire population of the province is covered by the Manitoba Health Services Insurance Plan. Nonparticipation is minimal since residents are not required to pay premiums to register for insured benefits.
- **Unique identifiers:** Use of a consistent set of identifiers (with identification numbers of both program recipients and providers scrambled to ensure confidentiality) permits researchers to build histories of individuals across time and across government programs. For example, individuals who are discharged from hospital

## Concept: Income Quintiles - Child Health Income Quintiles

### Concept Description

Last Updated: 2014-07-21

#### Introduction

An income quintile is a measure of neighbourhood socioeconomic status that divides the population into 5 income groups (from lowest income to highest income) so that approximately 20% of the population is in each group. At MCHP, we have created income quintiles for two distinct population groups: **urban** (Winnipeg and Brandon) and **rural** (other Manitoba areas).

This concept contains information in the following sections:

- data sources required for generating income quintiles;
- the steps for creating Income Quintile groups;
- postal codes that cannot be ranked and the Income Not Found group;
- a comparison of methodologies for the Standard Income Quintile versus the Child Health Income Quintile;
- urban and rural income quintile considerations;
- income quintile values and ranges;
- SAS code and format files; and
- a variation on the Child Health Income Quintile Methodology.

#### Data Sources

In order to generate Income Quintiles for socioeconomic analyses, the following data sources are required. The purpose of each data source is explained:

1. the [Manitoba Health Insurance Registry Data](#) is used to generate a population file for a selected year.
2. the public-use [Census data](#) files developed by Statistics Canada are used to identify the average household income value by Dissemination Area (DA).
3. the [Postal Code Conversion File \(PCCF\)](#) developed by Statistics Canada is used to identify and verify postal codes that fit within Dissemination Areas (DA).

#### Steps for Creating Income Quintiles Groups

Income Quintile information can be created for each year using the following general steps:

1. generate the population file for a selected year.
2. remove the postal codes that cannot be ranked. (see the next section for more information on why postal codes cannot be ranked).
3. attach the average household income value from the Census files to the population file using the Postal Code Conversion File (PCCF).
4. rank the population by Urban/Rural geographical location and by average household income.
5. form the 20% population income quintile groups based on the average household income values.

The data will be coded into one of the following groups:

- Urban Income Quintiles: U1 (lowest) to U5 (highest);
- Rural Income Quintiles: R1 (lowest) to R5 (highest); or
- Income Not Found (NF).

If the postal code cannot be ranked to an income quintile the value will be **Not Found (NF)**.

**Fig. 7** A screenshot of the “Income Quintiles” concept. Available at: <http://mchp-appserv.cpe.umanitoba.ca/viewConcept.php?conceptID=1161>

can be linked to the medical claims file in order to determine whether adverse events are being treated in physicians’ offices.

- **Longitudinal:** Migration into and out of the province as well as mortality can be traced from 1970 onward. Tracking groups of subjects through time can determine if individuals receiving a given intervention truly have no adverse outcomes or if adverse events are not showing up because the individual has left the province or has died.

Some of the key characteristics and research importance of these attributes are detailed in Table 1.

MCHP has created a series of tools to document the content of the data files, the process of gaining access to the data, and techniques for working with the data.

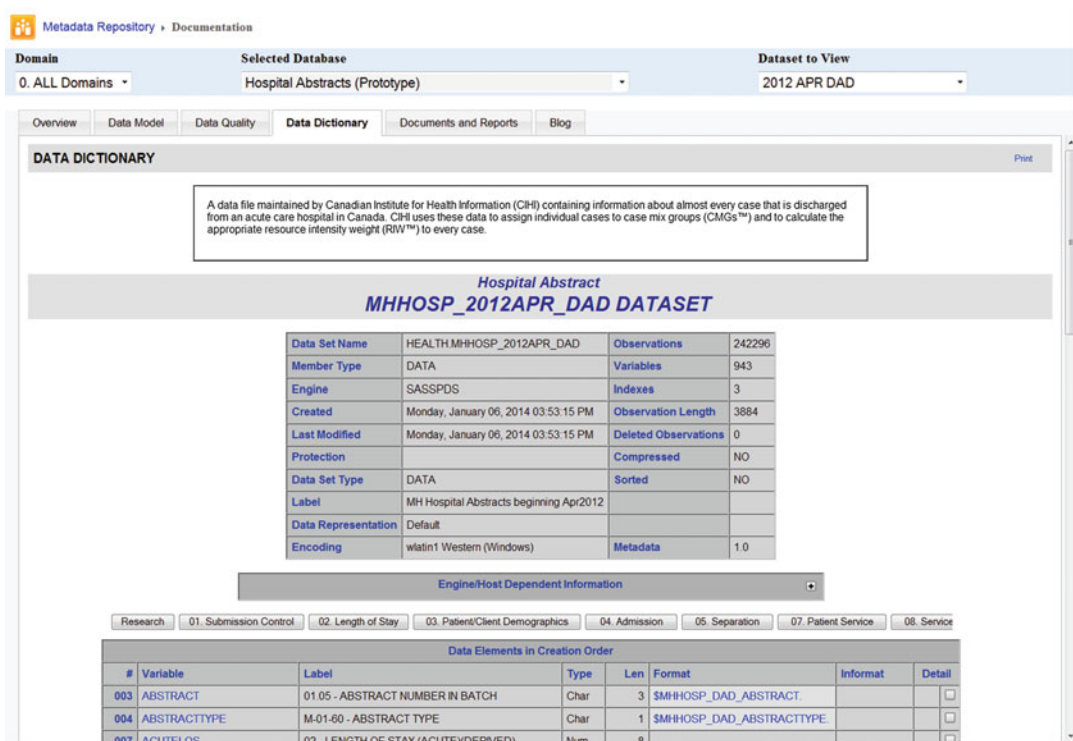
---

## Data Documentation

The MCHP Metadata Repository, currently available to internal users and users at remote access sites, organizes all of the Repository’s documentation. This tool provides a consistent set of documentation components for each group of data files. Components, displayed in the form of six

**Table 1** Manitoba research registry: key characteristics and research relevance (Roos 1999)

Characteristics	Research relevance
Very large N	Many physical and statistical controls are feasible; rare events can be analyzed; statistical power is high
Population based for an entire province	Heterogeneity along many variables is provided
Longitudinal data (going back over 30 years)	Many types of longitudinal designs are possible; important variables can be measured more reliably
Loss to follow-up specified	Follow-up critical for cohort studies is accommodated
Place of residence (according to postal code) at any point in time	Length of exposure to geographic areas can be quantified; measures of mobility and migration can be defined; small area variation analyses can be developed
Family composition at any point in time	Nonexperimental designs estimating the importance of different social variables and controlling for unmeasured background characteristics are facilitated



**Fig. 8** The MCHP metadata repository data dictionary intranet page

tabbed pages, include an Overview, Data Model, Data Quality Report, Data Dictionary, Additional Documents and Reports, and a Blog. See Fig. 8.

- 1. Overview** – A standardized data description summarizes the data, information on the data provider, purpose and method of data collection, years of available data, size of data files,

geographical parameters, data caveats, access requirements, and links to concepts to assist users working with the data. These descriptions provide users a sense of the extent, purpose, scope, and subject of a given database. They can also act as a first stop for researchers attempting to assess the feasibility of an administrative data project. The following list (see Table 2)

**Table 2** Standard headers used to describe all databases housed in the Repository

Header name	Description
<b>Summary</b>	A brief summary of the data, often used in grant applications, requests for data, and report glossaries. These serve as a very basic and general introduction to the data
<b>Source agency</b>	Data provider. Frequently the same agency from which access permission is required
<b>Type</b>	A conceptual category (domain) that is indicative of the type of record included in the file (e.g., administrative or survey)
<b>Purpose</b>	Provides a brief overview of why the data is collected by the source agency. What use it serves in the originating organization
<b>Scope</b>	The scope of the database; who or what is in, and who or what is not. May also include geographic, age, or program scope
<b>Data collection method</b>	A brief description of the original data collection process at the source
<b>Size</b>	General estimates of numbers of rows (records or observations) and columns (fields or variables)
<b>Data components</b>	The separate tables or sections that make up the data set
<b>Data level</b>	The level at which researchers can effectively and reliably study the data (e.g., individual or aggregate)
<b>Data years</b>	Range of data years and whether acquired by calendar, fiscal, or academic year
<b>Data highlights</b>	Key characteristics applicable for typical analyses
<b>Data cautions</b>	Obvious issues with the data of potential importance to researchers or useful for assessing project feasibility
<b>Access requirements</b>	Who to apply to in order to gain access to the data Direct links to the source agency's contact info or website are also included when appropriate
<b>More information</b>	Links to other sources of information such as the glossary, data dictionary, concept dictionary, provider's webpage, etc
<b>Previous and potential studies</b>	List of, and links to, MCHP deliverables and other reports or projects using the data
<b>References</b>	Any references used in the description/overview
<b>Date modified</b>	The date the overview was last modified

shows the standard headers used to describe all databases housed in the Repository.

Before the overview is published the data provider and a selection of users who frequently work with the data review the document for accuracy and completeness.

Overviews have also been sent to external organizations, such as Thomson and Reuter's "Data Citation Index," that include these documents in their integrated search systems. This facilitates the introduction of the data to external researchers, allows users to track and discover publications using a specific MCHP dataset, and increases the reach of the work produced by MCHP.

2. **Data Model** – A data model is created to display the structure of data files and how they are linked together in the Repository (see Fig. 9).

3. **Data Quality Report** – The usability of each field is addressed when data files are stored in the Repository and evaluations are summarized in a report available in the metadata repository. The data quality framework guiding this effort is available on MCHP's external website. A complete description of the data quality process is provided in the document *A Data Quality Evaluation Tool for Administration Data* available online and from MCHP Data Quality Framework [http://umanitoba.ca/faculties/health\\_sciences/medicine/units/communty\\_health\\_sciences/departmental\\_units/mchp/protocol/media/Data\\_Quality\\_Framework.pdf](http://umanitoba.ca/faculties/health_sciences/medicine/units/communty_health_sciences/departmental_units/mchp/protocol/media/Data_Quality_Framework.pdf).

4. **Data Dictionary** – The data dictionary identifies the files and tables held in the Repository. It provides detailed descriptions of individual data elements to assist users in their extraction,

### Social Housing – Tenant Management System – Client Related Tables

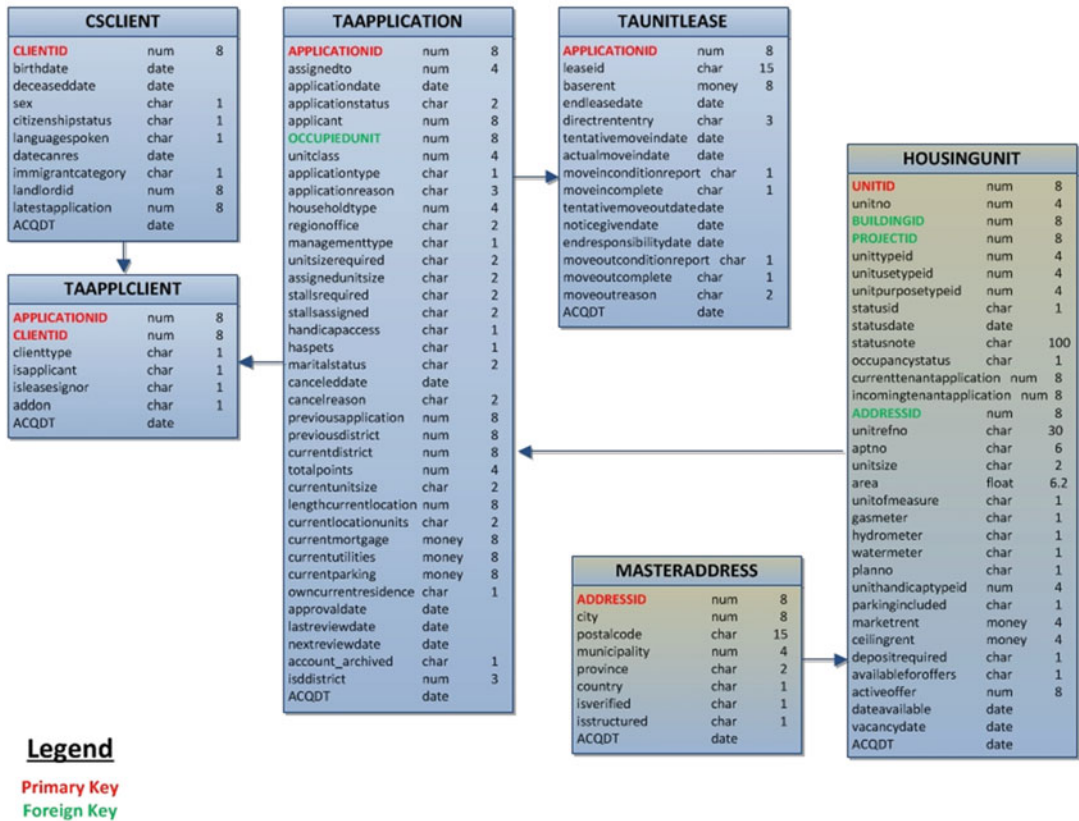


Fig. 9 Example of a data model diagram: social housing

management, and understanding of the data. Data file names and locations, field names, field definitions, descriptive labels, formats, a list of responses and frequencies for categorical variables or means, and distributions for numeric variables are provided in a web-based format.

- Documentation Directory**– Original information from the data provider, project documentation, and links to relevant concepts are stored with other documents that may be helpful in the interpretation of data files, such as training manuals, annual reports, and validation studies.
- Blog** – The blog component is a communication tool for analysts and researchers’ interested in communicating information about the data as it is discovered.

### Applying for Access

Under the Personal Health Information Act of Manitoba (PHIA), MCHP acts as a custodian of the data housed in the Repository. Access is based on the principle that the data is owned by the organization contributing the data – the data steward. Data-sharing agreements, negotiated with the provider, spell out the terms of use once the information is housed at MCHP. In addition, the research proposal process, administrative reporting requirements, and data use and disclosure requirements have been documented and are available on the MCHP website on the Applying for Access page: [http://umanitoba.ca/faculties/health\\_sciences/medicine/units/community\\_health\\_sciences/departamental\\_units/mchp/resources/access.html](http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departamental_units/mchp/resources/access.html)

As the number of data files and users has grown ensuring a common prerequisite level of knowledge has become increasingly important. An accreditation process established in April 2010 provides a consistent overview of MCHP and its data access and use policies and procedures. The accreditation material covers the MCHP mission (see Fig. 2), available data in the Repository, and the requirements for data use and publication of results. Accreditation is required for all researchers, students, and personnel working on approved projects. Once the initial accreditation session is completed, an online accreditation refresher module is available and must be completed annually. Accreditation information is also available for public access at: [http://umanitoba.ca/faculties/health\\_sciences/medicine/units/community\\_health\\_sciences/departmental\\_units/mchp/resources/accreditation.html](http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departmental_units/mchp/resources/accreditation.html)

## Repository Documentation

More general summaries of the Repository contents are produced in several formats:

### 1. Dataflow Diagram

The dataflow diagram illustrates the flow of data from its original source into the Repository. A reduced-scale version is shown in Fig. 10.

### 2. Data lists – several lists are maintained, each serving different purposes:

a. *Population Health Research Data Repository List* – a searchable and filterable list that indicates the years of available data, the source agency for each database, and provides links to individual data descriptions. An illustration of the interface is provided in Fig. 11.

b. *Data Years Chart* – Displays the years of available data for each file, with links to data descriptions. Figure 12 provides an example of the list.

### 3. Data Repository Slides – PowerPoint slides commonly used by researchers that describe or provide a representation of the MCHP data Repository (see Fig. 3).

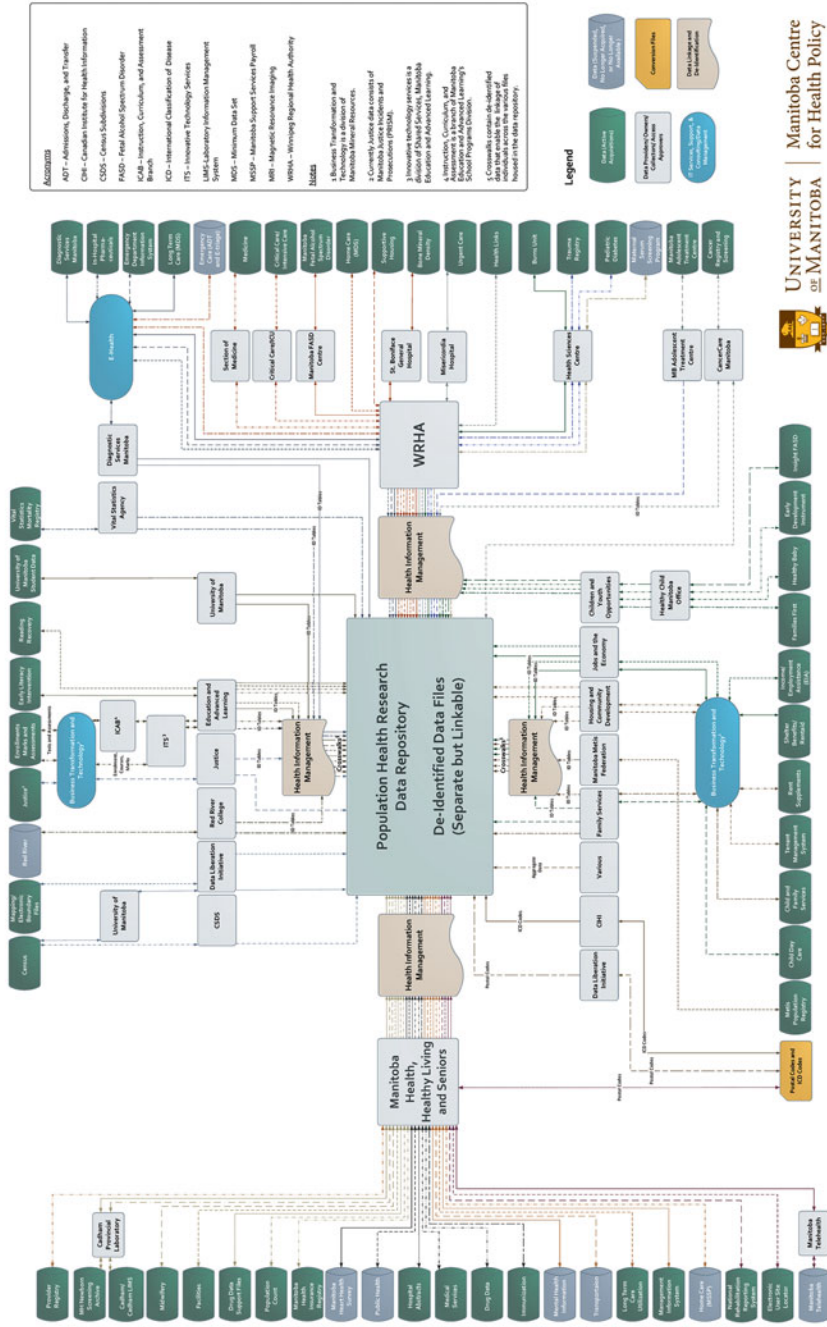
## The Data Management Process

MCHP's six-step data management process (see Fig. 13) describes how data are transferred from a source agency, processed, and brought into the Repository in order to be used for research purposes.

### Step 1: Formulate the Request and Receive the Data

A data-sharing agreement must be in place before any data can be received from the source agency. MCHP works in consultation with the source agency and the University of Manitoba's Office of Legal Counsel to produce an agreement. The data-sharing agreement defines policies and practices about data confidentiality, privacy, legislative and regulatory requirements, data transfer, and ongoing use of the data for research purposes. Data-sharing agreements are of two types: agreements for data added to the Repository at regular intervals (typically annually), and agreements for data provided for a single research project. For data added to the Repository at regular intervals, MCHP assumes responsibility for overseeing its use. This involves ensuring that appropriate policies and procedures governing use are established, documented, and enforced. For data added only for one specific project – called project-specific data – the principal investigator of the project assumes responsibility for overseeing the use of the data.

Once a data-sharing agreement is produced, a data management analyst is assigned to work with the source agency to facilitate the transfer. Initially this involves meeting with representatives from the source agency to acquire background information, documentation, data model diagrams, data dictionaries, documentation about historical changes in the data (including changes in program scope, content, structure, and format), existing data quality reports, and other information relevant to the description or use of the data. This information is used to: (a) develop a formal data request; (b) enhance the metadata repository, which contains database documentation; and (c) prepare the Data Quality Report. The analyst asks the source agency for reports or publications that document



**Fig. 10** A screenshot of the MCHP dataflow diagram available online. The full-scale diagram is available online at: [http://umanitoba.ca/faculties/health\\_sciences/medicine/units/community\\_health\\_units/mchp/protocol/media/dataflow\\_diagram.pdf](http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_units/mchp/protocol/media/dataflow_diagram.pdf)



The screenshot shows the 'Population Health Research Data Repository Data List' page. At the top, there is a navigation bar with the University of Manitoba logo and 'FACULTY OF MEDICINE Manitoba Centre for Health Policy'. A search bar is located in the top right corner. Below the navigation bar, there is a sidebar with various links such as 'Home', 'About MCHP', 'Data Repository', 'Applying for Access', 'Accreditation', 'Concept Dictionary and Glossary', 'Data Descriptions', 'Data Quality', 'Documentation', 'Study Design and Methods', 'Acknowledgments and Disclaimers', 'Research Resources', 'Updates', 'Research', 'News & Events', 'Knowledge Translation', 'Privacy & Confidentiality', and 'Contact'.

The main content area is titled 'Population Health Research Data Repository Data List'. It includes a navigation menu with 'Data List', 'Data Years', 'Data Diagram', 'Dataflow Diagram', and 'Repository Growth'. Below this is a dropdown menu set to 'All' and a 'Print' button. The data is presented in a table with columns for 'Name', 'Years Available', and 'Data Provider'.

Name	Years Available	Data Provider
<b>Health Data</b>		
<b>Admission, Discharge, and Transfer and E-Triage</b>	ADT: 1999/2000 to the end of 2010 for adults (children's records are only available in 2006/07, and only for 9 months)  E-triage: 2004/05 to 2010/11 for adults and children (children's records are not linkable to other data sets, unless they appear in ADT).	Winnipeg Regional Health Authority
<b>Bone Mineral Density</b>	2000/01 to 2012/13	Winnipeg Regional Health Authority St. Boniface Hospital

On the right side of the page, there is a text block: 'The MCHP data repository sends out a quarterly update identifying new research resources, data, and documentation at MCHP. If you're interested, please enter your email address in the space below. You can update your information or unsubscribe at any time.' Below this is a registration form with fields for 'Email', 'First Name', 'Last Name', and 'Business'. A 'Submit' button is at the bottom of the form. A note states: 'This email list is privately maintained for MCHP by iContact.' At the bottom right, there are social media icons for Twitter and Facebook with the text 'Find us. Twitter and Facebook.'

**Fig. 11** A Screenshot of how the Population Health Research Data Repository Searchable List Appears on the Website. The list is available online at: [http://umanitoba.ca/faculties/health\\_sciences/medicine/units/community\\_health\\_sciences/departmental\\_units/mchp/resources/repository/datalist.html](http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departmental_units/mchp/resources/repository/datalist.html)

[faculties/health\\_sciences/medicine/units/community\\_health\\_sciences/departmental\\_units/mchp/resources/repository/datalist.html](http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departmental_units/mchp/resources/repository/datalist.html)

the entities in the data, such as people, places, events, or activities (e.g., annual reports). This information is used to assess the accuracy and validity of the files that are brought into the Repository. Available financial data, such as annual budgets and total expenditures for specific programs, are also requested if available.

The initial data request encompasses historical documentation; that is, information that may have gone through multiple revisions over time, particularly in response to health system changes. The initial data request may in fact be a series of requests, one for each generation of source data. Future requests for updates may refer to the most recent generation only. All changes in coding methods, program constraints, and accounting measures are documented and incorporated into the metadata repository.

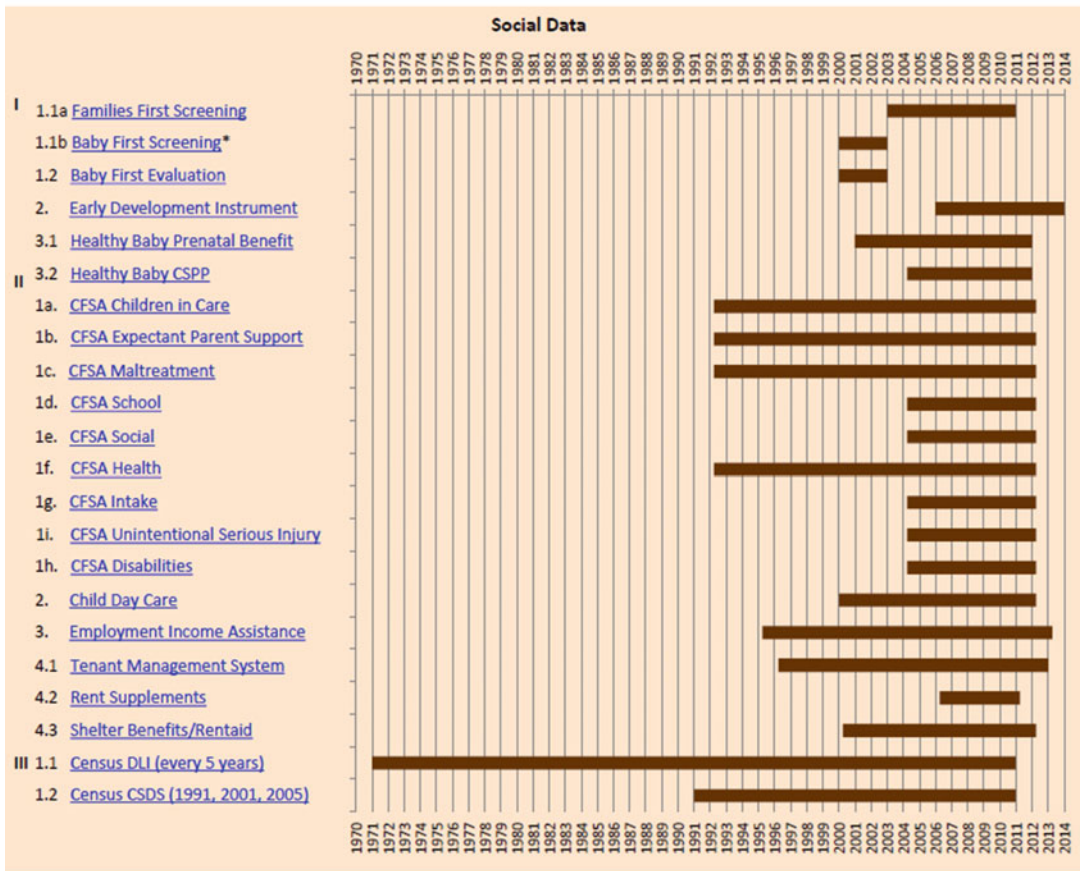
A sample data file is often prepared by the source agency and transferred to MCHP at the same time as the initial documentation transfer. Ideally, the sample consists of a random anonymized subset of the original data.

Once the documentation and sample data file have been evaluated, a formal data request is prepared and sent to the source agency. The data are then shipped to Manitoba Health for de-identification and data linkage (described above, under **Privacy**).

## Step 2: Become Familiar with the Data Structure and Content

Once MCHP receives the data, a data management analyst reviews the documentation and the organization of files and structures. While data in the Repository are usually organized to reflect the structure of the original source data, sometimes the files must be reorganized to permit addressing questions about different units of analysis that comprise the data, including persons, places, objects, events, and dates.

Tasks undertaken in the process of becoming familiar with the data structure and content include:



<sup>1</sup> Data years are displayed according to calendar year (ending Dec. 31). Not all repository data are acquired at the same regular intervals and some repository data are obtained by fiscal year (April 1-March 31). Available fields and coding rules may differ from year to year for many data files. Examples include changes in legislation, ICD coding classifications, changes in CMG coding, and the introduction of new computer or software resources over time. The numbering corresponds with the Data Repository Data List.

<sup>2</sup> Urgent care data are not available for the years 2005/06 to 2009/10.

<sup>3</sup> These data are not available in their current form beyond the last year shown in this chart. There may be alternative sources for some of the discontinued information (for example EDIS replaces ADT).

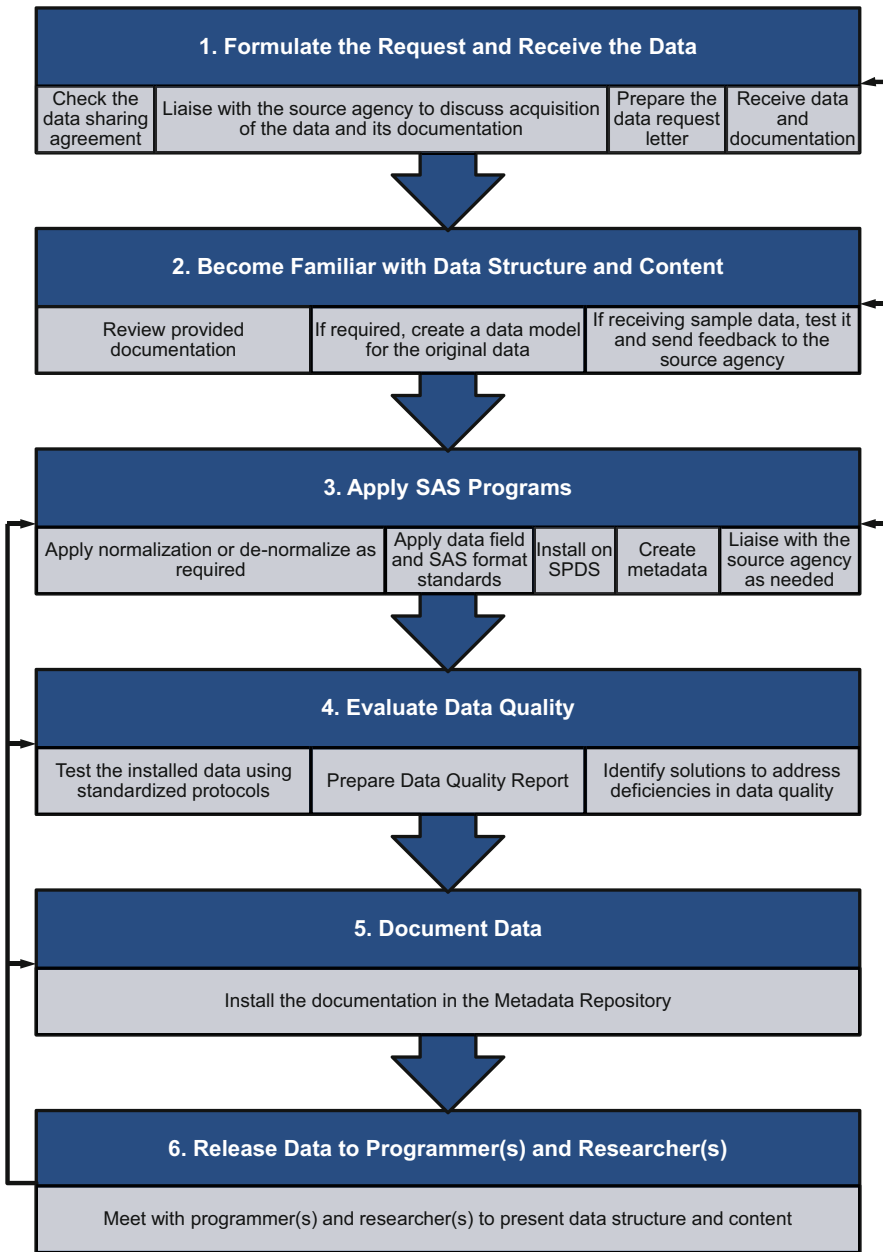
**Fig. 12** A screenshot of the website and the years of social data available. Also available online at: [http://umanitoba.ca/faculties/health\\_sciences/medicine/units/community\\_health\\_sciences/departamental\\_units/mchp/protocol/media/Available\\_Years.pdf](http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departamental_units/mchp/protocol/media/Available_Years.pdf)

1. Standardizing unique record identifiers. If the PHIN is missing, then a unique “placeholder” value is created by MCHP analysts.
2. Standardizing dates of events and correcting incomplete dates, where possible.
3. Standardizing frequently used demographic data elements, including sex and postal code.
4. Identifying and restricting access to data elements not normally made available to researchers without special permission. Examples include registration numbers and hospital chart numbers.

5. Reorganizing and converting files to a different file format, if necessary.

**Step 3: Apply SAS® Programs**

MCHP uses SAS® for analysis, which performs optimally with data files that have been denormalized (SAS Institute Inc. 2006). Denormalization is a process of adding redundant information to a data file to reduce the processing time required for analysis. Standardized formats are applied to selected fields, such



**Fig. 13** The six-step data management process

as date fields. Once a data file has been prepared for research use, the SAS Scalable Performance Data Server (SPDS) is used to sort and create indices and other design elements appropriate for the most commonly used applications. During this process, standard naming conventions for data files are applied. SAS® is

then used to create a summary of the contents for documentation purposes.

**Step 4: Evaluate Data Quality**

A Data Quality Report is produced for each dataset in the Repository. This report is housed

in the metadata repository, which provides a single point of access for all documentation concerning a data file. The structure and contents of the Report, and the framework guiding the development of the report, are described below under **Data Quality Evaluation Tool for Administration Data**.

**Step 5: Document the Data**

Data dictionaries, which contain information about the name, contents, and format of each field, are created and stored in the metadata repository. The data dictionaries can be used to conduct an initial review of data quality; a cursory review can identify problems such as missing data, incompleteness of labels and descriptors, problems with ranges in numeric values, and/or integrity of data linkage keys.

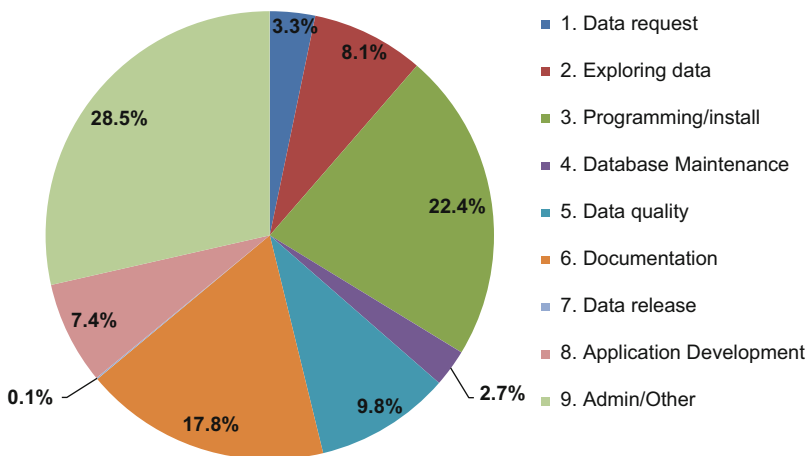
Before the data are stored in the Repository, the data dictionaries are subjected to an initial assessment of accuracy and completeness. If deficiencies are identified, the analyst will investigate them through further contacts with the source agency, Manitoba Health, or MCHP personnel.

**Step 6: Release the Data**

If the data files and documentation appear ready, the data can be released internally for use. Release may be informal, in which case analysts are simply notified that the new data and documentation are available for use, or more formal, involving presentations to data analysts and researchers. The latter is useful when a new data source is valuable for multiple research projects, if substantial changes have occurred to existing data or when the source agency has introduced a new data-capture process or system. New and updated datasets are also announced in the MCHP quarterly newsletter *Research Resources Update*, published online at: [http://umanitoba.ca/faculties/health\\_sciences/medicine/units/community\\_health\\_sciences/departmental\\_units/mchp/resources/repository/rupdate.html](http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departmental_units/mchp/resources/repository/rupdate.html)

**Percent of Time Spent on Each Data Management Activity**

Now that MCHP has developed a methodological approach to acquiring and installing data, time spent in each of the various categories of activity can be tracked. Figure 14 shows staff



\*There are seven Data Analysts in the Data Management Work

**Fig. 14** Average percent of time spent on each data management activity\* for 2014

time spent in each category accumulated over a 1 year period. It was instructive to realize that about one third of staff time is spent in either administrative (meetings and presentations, general communication, training) or application development activities. The latter includes such things as the development of data quality macros and tools to implement the metadata repository. As Fig. 14 shows, programming data to be stored (programming/installing data) and documenting data are two of the largest areas of activity, followed by data quality assessments and exploring data on arrival at the center. The smallest areas of activity involve requesting data and performing revisions to existing data (database maintenance). MCHP continues to monitor time spent on each activity in order to track fluctuations over time. At the moment, MCHP does not have a formal data release process; therefore, no time is accruing in that activity. A dissemination strategy will be developed in the coming year.

## Summary

The six-step data management process used at MCHP follows standards and practices observed in other similar initiatives as well as recommendations developed by organizations maintaining repositories of anonymized personal health information for research purposes (for examples, see (Daas et al. 2008; Holman et al. 1999; Lyman et al. 2008)). MCHP's process also reflects some of the more unique aspects of the political and social environment in which it operates, including relationships with source agencies, the software platform on which the Repository is maintained, and provincial health privacy legislation.

---

## Data Quality Evaluation Tool for Administration Data

Data collected for administrative purposes are not always of the best quality for research, and poor quality data may lead to false conclusions.

To determine the quality of data coming into MCHP an evaluation tool was developed (see Fig. 15). This tool was implemented using SAS<sup>®</sup> software and is specifically designed to assess the following characteristics of administrative data:

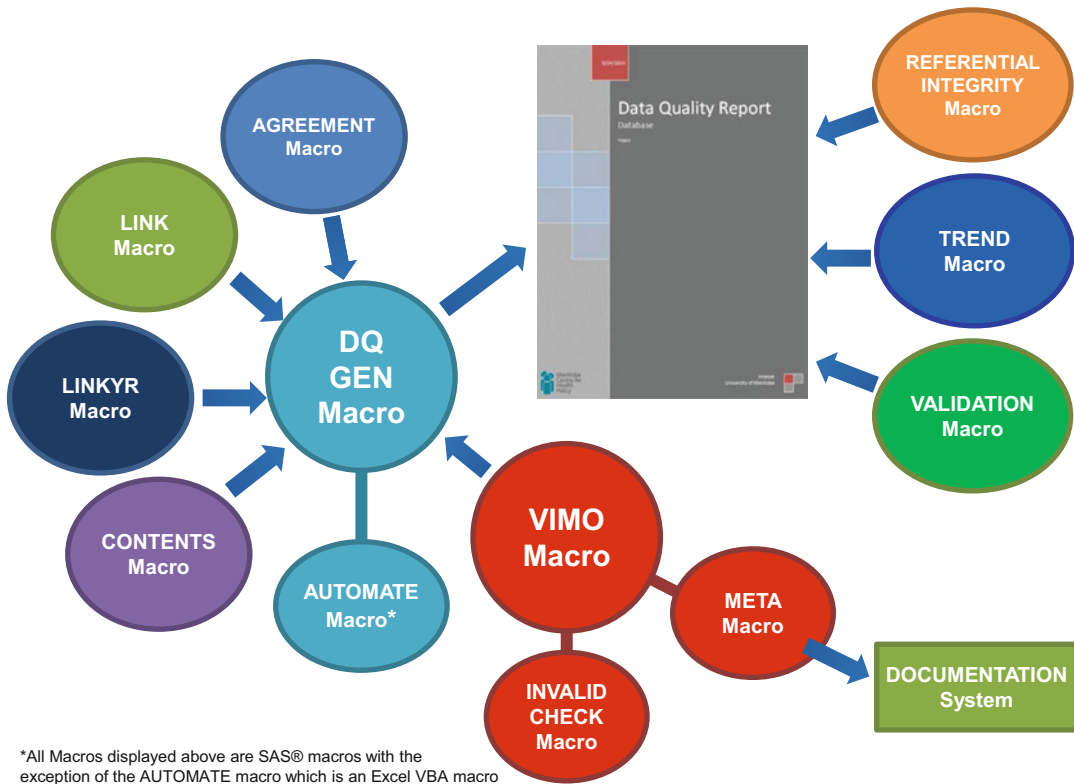
- Completeness and correctness
- Consistency
- Referential Integrity
- Trends in the data
- Crosswalk linkage assessment
- Agreements using kappa statistics

## Completeness and Correctness

Completeness refers to the magnitude of missing values; such values are identified and reported for all data elements. The assessment of correctness includes the fraction of data elements that are valid, invalid (e.g., categorical variables that do not match a reference list, out of range numerical variables, invalid dates such as a living person born in the 1800s), missing data, and outliers for all numeric variables. The process of checking the large number of files that flow into the repository at MCHP would be infeasible if not for the ability to automate the process. Completeness and correctness can be evaluated using an automated set of SAS macros developed at MCHP called **META**, **INVALID CHECK**, and **VIMO**. These macros produce the VIMO table (see Fig. 16) that documents the percentage of valid, invalid, missing, and outlier data. Fields with invalid values are flagged and the total number of invalid records is automatically noted in the comment column.

## Assessing Consistency

Consistency refers to the intra-record relationship among data elements. For example, hospital admission dates must precede hospital separation (discharge) dates. Consistency can be assessed using MCHP's **VALIDATION** macro which is based on predefined consistency criteria. Each record is checked for consistency, and the results are summarized as a table showing the total number of



**Fig. 15** Generating a data quality report

inconsistent records. For example, the validation macro can be used to check for inconsistencies in reporting the pregnancy indicator (see Fig. 17).

### Referential Integrity

In a relational database, referential integrity refers to the quality of linkages existing between database tables. Typically one table contains a unique identifier known as the primary key, which may be a single attribute or a set of attributes that uniquely identify each record. Other tables will contain foreign keys, and each foreign key value must reference a primary key value in the primary table. The **REFERENTIAL INTEGRITY** macro (see Fig. 18) checks for the number of primary keys having duplicate or missing values as well as the total number of foreign key values that do not reference a valid primary key (orphaned values).

### Trend Analysis

A macro has been developed to perform a trend analysis for core data elements. For example, no change across years may indicate a data quality problem if the data are expected to trend naturally upward or downward due to policy, social, or economic changes. Fields such as the diagnosis and treatment of a specific cancer can be assessed over a number of years. The macro plots frequency counts across a specified time period. This macro also fits a set of common regression models and chooses the best-fit model based on the minimum root mean square error (RMSE). With the best regression model selected, studentized residuals with the current observation deleted are calculated. Aggregated observations with absolute studentized residuals greater than  $t(0.95, n - p - 1)$  are flagged as potential outliers indicating an unusual change

Dataset Label: dataset label      Records: 10000  
 Dataset Name: dataset name      Period: yyyy

Legend (Potential Data Quality Problems) :

None or Minimal < 5%	Moderate 5-30%	Significant > 30%	Unknown or N/A
-------------------------	-------------------	----------------------	-------------------

= No variance or 100% missing value  
 = Min, Max values based on valid range

Type	Variable Name	Variable Label	Valid	Invalid	Missing	Outlier	Min	Max	Mean	Median	STD	Comment
Num	VAR1	variable1	100.00		.00							
	VAR2	variable2	100.00		.00							
	VAR3	variable3	94.75		4.76	.49	0.83	10.00	8.67	9.23	1.48	
	VAR4	variable4	70.77		29.23	.00	1.00	99.00	38.63	2.08	46.06	
	VAR5	variable5	95.09		4.70	.21	0.00	10.00	8.13	9.01	1.96	
	VAR6	variable6	100.00		.00	.00	0.00	0.00	.00	.00	.00	
	VAR7	variable7	85.91		.00	14.09	0.00	110.00	6.10	.01	22.99	
Char	VAR8	variable8	99.32	68	.00		-1 0 1	Observed Values				-1 (68 Invalid Obs. in total)
	VAR9	variable9	.00	100.00			23 01 21 25 19 07 16 09 26 28 08 10 27 30 18 17 29 22 31 12 11 03 15 14 13 02 04 05 06 24 20					
	VAR10	variable10	93.41		6.59		15 24 75 76 78 79 80 81 83 84 85 86 88 89 90 91 92 94 97 98					
	VAR11	variable11	100.00		.00		100 102 103 104 130 132 137 138 146 148 217 229 233 234 236 237 238 239 112 77 101 231 113 82 74 87 227 235 226 232					
	VAR12	variable12	100.00		.00		2011					
	VAR13	variable13	28.02	.02	71.96		2001-03-28	2006-03-13				1582-10-14 ( 2 Invalid Obs. in total )
	VAR14	variable14	99.61		.39		2003-06-28	2006-11-04				
	VAR15	variable15	87.74	12.26	.00		02JAN2001:03:13:36	01APR2006:22:26:52				1226 invalid obs. out of [01JAN2001:23:59:59, 01APR2006:23:59:59] range
	VAR16	variable16	100.00		.00		0:00:02	23:59:48				

Fig. 16 A VIMO table generated by macros which documents the percentage of valid, invalid, missing, and outlier data elements

Validation Check for Data Consistency

Count	Error Message	Condition
1	pregnant man	sex = '1' and preg = '1'
1	pregnant woman with age >= 70	sex = '2' and preg='1' and age >= 70

Fig. 17 The validation macro demonstrates inconsistencies

Key: CLIENT\_VISIT\_GUID

PRIMARY TABLE	DUPLICATE	MISSING	TOTAL RECORDS
WRHA_EDIS_CLIENT_2007JAN	1 (x3)		1,098,981

FOREIGN TABLE	ORPHAN VALUES	TOTAL RECORDS
WRHA_EDIS_STATUS_2007JAN	399	2,987,150
WRHA_EDIS_PROVIDER_2007JAN	400	6,133,612
WRHA_EDIS_NACRS_2007JAN	188	586,504

Fig. 18 Output from the referential integrity macro

for a particular year of data. Typical output is illustrated in Fig. 19. Variations in expected trends are typically used as indicators that further exploration is necessary.

**Assessing Agreement**

Since many of the MCHP data linkages are based on probabilistic matching, rates of agreement for sex and date of birth between the incoming data and MCHP’s population-based longitudinal health registry are evaluated using kappa statistics and the **AGREEMENT** macro.

**Assessing Crosswalk Linking**

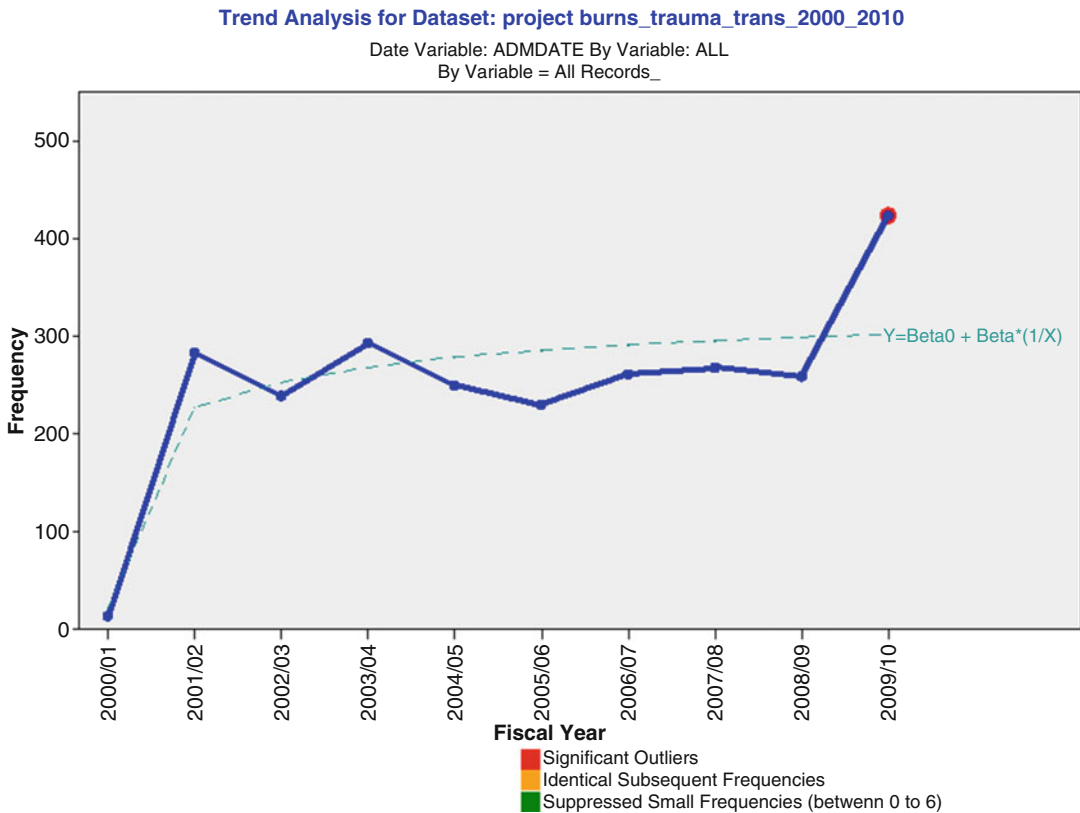
Before data arrive at Manitoba Health, they are first sent to Manitoba Health for data linkage. Manitoba Health then removes all personal and

identifying information, and the data are sent to MCHP with an encrypted PHIN that can be linked with MCHP databases for research purposes. The viability of linking incoming data with other MCHP databases can be assessed using the **LINK** and **LINKYR** macros (see Fig. 20).

**Summary**

The data quality report generated by these macros is useful in several ways. First, potential data quality issues are flagged so that researchers and data analysts are aware of potential pitfalls when performing data analyses. Second, sharing the report with data providers draws their attention to potential issues so that action can be taken to improve the quality of data over time. Third, it provides a useful starting point for discussing the data, both to new users who may have no idea of the content as well as among data acquisition staff so that they can spot





**Fig. 19** Typical output from the trend analysis macro

discrepancies and anomalies in the data and correct or document them before the data is released.

Anyone interested in implementing the Data Quality assessment tools developed at MCHP can download the source code, examples, and documentation at [http://umanitoba.ca/faculties/health\\_sciences/medicine/units/community\\_health\\_sciences/departmental\\_units/mchp/resources/repository/data\\_quality.html](http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departmental_units/mchp/resources/repository/data_quality.html). This software is freely available for use under a GNU General Public License.

### Advantages of Using a Population-Based Registry

As illustrated in Fig. 3, a central component of the Repository is an anonymized population-based registry: a longitudinal registry of individuals covered by the provincial health insurance plan. It

provides an opportunity for preparing data, improving quality, and understanding error through linkage to files with independent information on relevant variables. For example, comparing date of death from the Manitoba Health Insurance Registry with the date recorded in the governments Vital Statistics files allows for error correction.

The population-based registry has been critically important for many studies since 1977 (Roos et al. 1977). Besides using the registry for computing geographically-based rates, individuals have been located within families to determine the health and health-service use of particular ethnic groups (Martens et al. 2005, 2011) and the registry has been critical for longitudinal studies, being used for relatively short-term follow-up of surgical outcomes and multi-year birth-cohort research (Brownell et al. 2014; Oreopoulos et al. 2008; Roos et al. 1992).

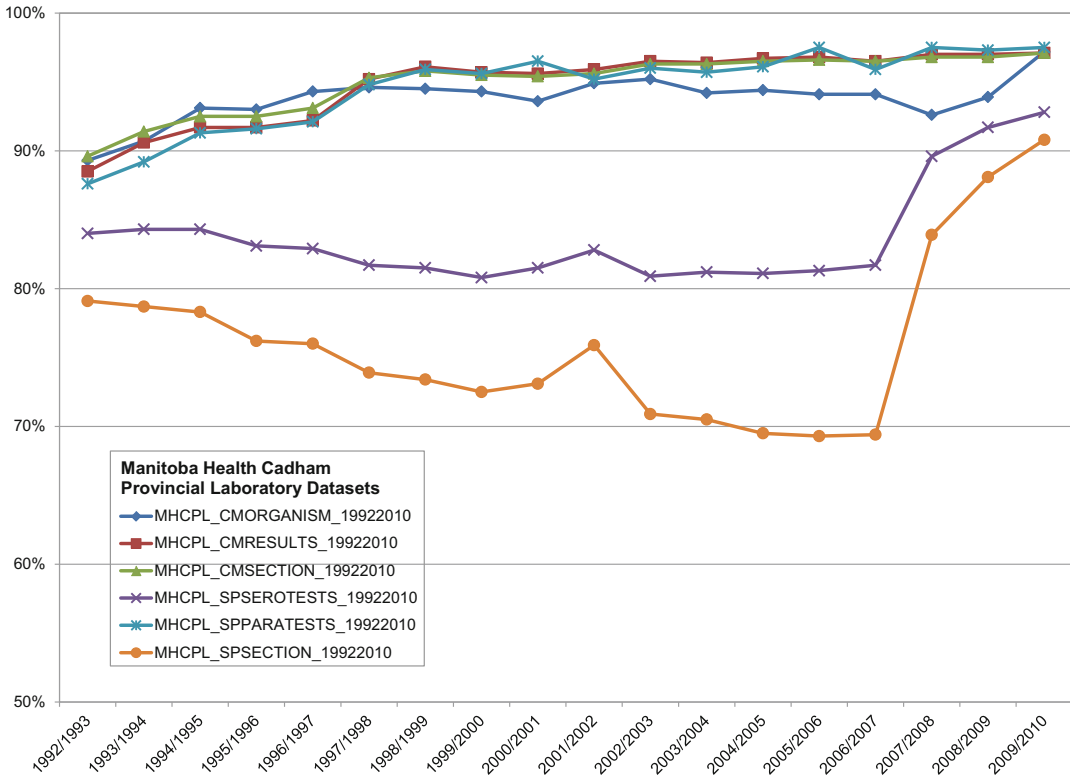


Fig. 20 Output from the LINKYR macro

### Expanding Capabilities into Social Policy Research

Canada’s population-based data on families, neighborhoods, and schools are increasingly being used to study individuals cost-effectively in their social context. However, considerable preparatory work is necessary to move from health research to social research. This work includes:

- Building reliable social measures applicable across studies. In some cases this may mean tracking family composition over time (marriages, marriage break-up, remarriage, family size, and ages of children).
- Identifying siblings and twins to facilitate more complex research methodologies.

### Using Place-of-Residence Data

Postal codes allow users of the MCHP research registry to infer the location of individuals at any specified date after 1970. Semi-annual updates allow capturing information on individuals who have notified Manitoba Health of a change in residence or provided an updated address as part of an encounter with the healthcare system. Consequently, research designs that involve linking individuals to their neighborhood context are relatively easy to implement.

- Developing scales to measure new outcomes such as educational achievement at the population level.
- Using place-of-residence data (at any given point in time) to calculate the number of moves, number of years in certain neighborhoods (poor vs. wealthy), upward and downward mobility (as defined by neighborhood median income).

Variables pertaining to residential mobility, years living in a neighborhood with particular characteristics and so on, can be generated. Such recording of “exposure” is methodologically superior to relying on cross-sectional variables from one point in time. Variables measured at various times over relatively long periods may help resolve disagreements as to when, in the early life course, different factors might be occurring. To treat periods in the life course separately, counts of years in a neighborhood or in a particular social situation can be generated for different intervals (e.g., ages 0–1, 2–4, 5–9, etc.).

## Constructing Reliable Social Measures

The MCHP research registry’s capacity to place a Manitoba resident within a family structure at any given time permits development of a number of social variables:

- Number of years on income assistance
- Average household income
- Number of children in the family
- Birth order (particularly being first-born)
- Mother’s marital status at birth of first child
- Number of years living in a single-parent family
- Age of mother at birth of first child
- Family structure (or number of years in different types of families)
- Number of family structure changes (parental separations, remarriages)
- Number of years living with a disabled parent
- Number of household location moves
- Immigrant status
- Neighborhood characteristics

The longitudinal nature of the data and the repeated measurements that it facilitates constitute a real strength. For example, since the short-term effects of income assistance and welfare reciprocity differ from long-term effects, being able to differentiate between the two can be critical. A few survey-based studies have counted years in particular types of families or neighborhoods in an effort to quantify the impact of various social environments, but longitudinal administrative data may

make collecting this data more efficient and reliable (Roos et al. 2008).

One of the advantages in using administrative social variables is that they can help adjust for differences in family background. For example, in one study it was found that nine variables (gender, income assistance, receiving services/children in care, family structure, number of siblings, birth order, mother’s age at first birth, residential mobility, and the neighborhood-based Socioeconomic Factor Index) accounted for as much variance in the Manitoba Language Arts achievement test as a similar sized set of variables from survey data (Roos et al. 2008, 2013). That is, administrative data were as good at predicting the outcome as were the survey data.

## Identifying Siblings and Twins

Birth cohorts, siblings, and twins may be defined from one or more sources of administrative data. In Manitoba, two hospital separation (discharge) abstracts – one for the mother and one for the infant – are produced for each in-hospital birth. These records can be checked against each other and against the Manitoba Health Insurance Registry. Two siblings with the same birth date are designated as twins.

Sibling and twin studies are important because omitted variables and measurement error that occur in studies that do not examine siblings and twins are likely to bias the coefficients attached to measured variables. The effects of certain variables such as birth weight may be overestimated when other variables associated with the family are not appropriately controlled for. In Canada, the availability of statistical power (a large N), heterogeneity – e.g., wide variations in population characteristics across areas – and place-of-Residence data enhances the possibility for sophisticated research designs employing siblings and twins (Roos et al. 2008).

## Beyond Health Research

Since the importance of early childhood conditions are known to be significant, yet childhood disease is

relatively rare, information on school performance and income assistance can provide a window on well-being during childhood and adolescence. Attention to the socioeconomic gradient over the early life course builds on the hypothesis that the relatively affluent will disproportionately take advantage of and benefit from health and educational programs. In other words, wealthy individuals are more likely to be exposed to and take advantage of new initiatives and opportunities that make them healthier and low-income people are less likely to do so.

---

## Summing Up

The Population Health Research Data Repository housed at MCHP is one of the most established and comprehensive provincial repositories of health and social data in Canada. Currently, more than 200 research projects are being conducted using these data. In addition to the policy-relevant research produced in the form of deliverables to the Manitoba government (discussed in the next chapter) numerous high-quality academic papers are published in areas of health services research and population health. Increasingly, studies are focusing on the social determinants of health as more social data becomes available (some of these are listed in Roos et al. 2008).

Research units like MCHP that house large databases accessed by many investigators and graduate students can benefit from the creation of web-based research resources to compile and disseminate common organizational knowledge. Creating a single point of access to the knowledge generated from a wide range of projects is important for ensuring a high level of productivity and methodological excellence.

---

## References

Brownell M, Chartier M, Au W, Schultz J. Evaluation of the healthy baby program. 2010. [http://mchp-appserv.cpe.umanitoba.ca/reference/MCHP-Healthy\\_Baby\\_Full\\_Report\\_WEB.pdf](http://mchp-appserv.cpe.umanitoba.ca/reference/MCHP-Healthy_Baby_Full_Report_WEB.pdf). Accessed 29 May 2013.

Brownell MD, Nickel NC, Chateau D, et al. Long-term benefits of full-day kindergarten: a longitudinal

population-based study. *Early Child Dev Care*. 2014;185:291–316.

Chartier M, Finlayson G, Prior H, McGowan K-L, Chen H, de Rocquigny J, Walld R, Gousseau M. Health and healthcare utilization of francophones in Manitoba. 2012. [http://mchp-appserv.cpe.umanitoba.ca/reference/MCHP\\_franco\\_report\\_en\\_20120513\\_WEB.pdf](http://mchp-appserv.cpe.umanitoba.ca/reference/MCHP_franco_report_en_20120513_WEB.pdf). Accessed 29 May 2013.

Chateau D, Doupe M, Walld R, Soodeen RA, Ouelette C, Rajotte L. Projecting personal care home bed equivalent needs in Manitoba through 2036. 2012. [http://mchp-appserv.cpe.umanitoba.ca/reference/MCHP\\_pch\\_days\\_report\\_WEB.pdf](http://mchp-appserv.cpe.umanitoba.ca/reference/MCHP_pch_days_report_WEB.pdf). Accessed 29 May 2013.

Daas PJH, Arends-Tóth J, Schouten B, Kuijvenhoven L, Statistics Netherlands. Quality framework for the evaluation of administrative data. 2008. <http://www.pietdaas.nl/beta/pubs/pubs/21Daas.pdf>

Doupe M, Fransoo R, Chateau D, Dik N, Burchill C, Soodeen R-A, Bozat-Emre S, Guenette W. Population aging and the continuum of older adult care in Manitoba. 2011. [http://mchp-appserv.cpe.umanitoba.ca/reference/LOC\\_Report\\_WEB.pdf](http://mchp-appserv.cpe.umanitoba.ca/reference/LOC_Report_WEB.pdf). Accessed 29 May 2013.

Fransoo R, Martens P, The Need to Know Team, Prior H, Burchill C, Koseva I, Bailly A, Allegro E. The 2013 RHA indicators atlas. 2013. [http://mchp-appserv.cpe.umanitoba.ca/reference/RHA\\_2013\\_web\\_version.pdf](http://mchp-appserv.cpe.umanitoba.ca/reference/RHA_2013_web_version.pdf). Accessed 20 Nov 2013.

Holman CDJ, Bass AJ, Rouse IL, et al. Population-based linkage of health records in Western Australia: development of a health services research linked database. *Aust N Z J Public Health*. 1999;23:453–9.

Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. *Annu Rev Public Health*. 2011;32:91–108.

Lyman JA, Scully K, Harrison JH. The development of health care data warehouses to support data mining. *Clin Lab Med*. 2008;28:55–71.

Martens PJ, Sanderson D, Jebamani L. Mortality comparisons of First Nations to all other Manitobans: a provincial population-based look at health inequalities by region and gender. *Can J Public Health*. 2005;96:S33–8.

Martens PJ, Bartlett J, Burland E, Prior H, Burchill C, Huq S, Romphf L, Sanguins J, Carter S, Bailly A. Profile of metis health status and healthcare utilization in Manitoba: a population-based study. 2010. [http://mchp-appserv.cpe.umanitoba.ca/reference/MCHP-Metis\\_Health\\_Status\\_Full\\_Report\\_\(WEB\)\\_update\\_aug11\\_2011.pdf](http://mchp-appserv.cpe.umanitoba.ca/reference/MCHP-Metis_Health_Status_Full_Report_(WEB)_update_aug11_2011.pdf). Accessed 29 May 2013.

Martens PJ, Bartlett JG, Prior HJ, et al. What is the comparative health status and associated risk factors for the Metis? A population-based study in Manitoba, Canada. *BMC Public Health*. 2011;11:814.

Oreopoulos P, Stabile M, Walld R, et al. Short, medium, and long term consequences of poor infant health: an analysis using siblings and twins. *J Hum Resour*. 2008;43:88–138.

Roos NP. Establishing a population data-based policy unit. *Med Care*. 1999;37:JS15–26.

- Roos NP, Roos LL, Henteleff PD. Elective surgical rates: do high rates mean lower surgical standards? *N Engl J Med.* 1977;297:360–5.
- Roos LL, Fisher ES, Brazauskas R, et al. Health and surgical outcomes in Canada and the United States Summer. *Health Aff (Millwood).* 1992;11 (Summer):56–72.
- Roos LL, Brownell M, Lix L, et al. From health research to social research: privacy, methods, approaches. *Soc Sci Med.* 2008;66:117–29.
- Roos LL, Hiebert B, Manivong P, et al. What is most important: social factors, health selection, and adolescent educational achievement. *Soc Indic Res.* 2013;110:385–414.
- SAS Institute Inc. SAS data integration studio 3.3: user’s guide. 2006. <http://support.sas.com/documentation/onlinedoc/etls/usage33.pdf>. Accessed 12 Aug 2014.
- Wolfson M. A shining light in Canada’s health information system. *Healthcare Policy.* 2011;6:8–13.



# Health Services Data, Sources and Examples: The Institute for Clinical Evaluative Sciences Data Repository

# 3

Karey Iron and Kathy Sykora

## Contents

<b>Introduction</b> .....	48
<b>Strengths and Challenges of Using Health Administrative Data for Health Services Research</b> .....	48
<b>The ICES Data Repository</b> .....	49
Privacy, Data Governance, and Access to Data at ICES .....	51
<b>Record Linkage and Desensitizing the Data for Research</b> .....	52
<b>Data Documentation, Metadata, and Data Quality Assessment</b> .....	53
Data Quality Assessment in the Literature .....	53
New Data, New Uses, and New Ideas .....	57
<b>References</b> .....	59

## Abstract

Under the 1982 *Canada Health Act*, health services deemed essential for all residents are universally paid for by the provinces. Canadian provinces, and others around the world, routinely collect data that allow them to *administer* health services provided to their populations. Generally, this spectrum of *health administrative data* includes information about people and their use of the health system, such as physicians' billing claims, hospital discharges, emergency and ambulatory care,

home care, complex continuing and long-term care, and claims for prescription drugs, for example. When linked to each other, these highly comprehensive data may be used to answer health system and research questions such as: Are those who require care getting the care they need? Is the care provided timely and based on evidence? What organizational aspects of the healthcare system could improve care? This chapter describes the uses of health administrative data for research, its benefits and limitations compared to traditional research data, the concept of linking datasets for health services research, emerging data quality scientific methods, and the caveats in interpreting administrative data. Issues of data governance and privacy, data documentation, and quality assessment are presented. These

K. Iron (✉)  
College of Physicians and Surgeons of Ontario, Toronto,  
ON, Canada  
e-mail: [kiron@cpso.on.ca](mailto:kiron@cpso.on.ca)

K. Sykora  
Toronto, ON, Canada

concepts will be illustrated through the example of the data held in the Institute for Clinical Evaluative Sciences (ICES) Data Repository in Ontario.

---

## Introduction

Under the 1982 *Canada Health Act*, health services deemed essential for all residents must be paid for by the provinces and territories. In order to manage, administer, and pay for health services for their populations, the provinces and territories routinely collect information about health system transactions. Generally, this spectrum of health administrative data includes information about people and their use of the health system, physicians' billing claims, hospital discharges, emergency and ambulatory care, home care, complex continuing and long-term care, and claims for publicly funded prescription drugs, to name a few. Other large and routinely collected datasets are also generated and used by various organizations throughout the health system to understand how health services are being used. Examples include public health program information, agency-level client information, population-based registries and surveys, electronic medical records, and, most recently, large genomic biobank data. The power of these data is amplified when they are linked to each other to understand the whole picture of healthcare delivery. According to Friedman et al. (2005), when these data are used to generate "health statistics," they create "fundamental knowledge about the health of populations" that inform the health system, "influences on health" that guide policy decisions, and "interactions among those influences" that guide program development and clinical care (Friedman et al. 2005). For example, linked data may answer health system, population-based, and clinical research questions such as: Are patients getting the care they need? Is the care timely and based on optimal evidence? How might the system be better organized to optimize care? Is the care provided equitable across the province?

This chapter focuses on the following areas:

- Strengths and challenges of using health administrative data for health services research
- Privacy and data governance
- Record linkage and desensitizing the data for research
- Data documentation and data quality assessment
- New data, new uses, and new ideas

These concepts will be illustrated through the example of the data held in the Data Repository at the Institute for Clinical Evaluative Sciences (ICES), a research organization in Toronto, Ontario, Canada, that collects and manages a large data repository that is used to generate evidence to improve health and the health system in Ontario.

---

## Strengths and Challenges of Using Health Administrative Data for Health Services Research

In Ontario and elsewhere in Canada, health administrative data are used not only for managing the health system but also for health services research, policy development, and healthcare planning. Since most residents are eligible for healthcare, the data reflect full coverage of publicly funded service transactions. The data represent actual encounters with the healthcare system and are therefore population based, free from recall bias, readily available, consistent over time, and are inexpensive to collect and use for secondary purposes compared to traditional research data. Generally, health administrative data are collected using standardized coding metrics, especially when the data are collected by a single source (such as a provincial health authority or ministry). Using one dataset alone is useful for health system surveillance and monitoring, but the real power of using administrative data lies in the ability to link multiple datasets at the individual person level and across healthcare sectors. In his seminal work, Dunn describes record linkage as follows:

Each person in the world creates a book of life. This book starts with birth and ends with death. Its pages are made of the records of principal events in life. Record linkage is the name given to the process of assembling the pages into a volume. (Dunn 1946)

The linkage of data enables researchers to answer questions based on information from different parts of the healthcare system. Without linkage, we can look at hospitalization data and ask: “How many people were admitted to hospital with a heart attack and what hospital care did they receive?” But with linked data, we can answer more involved questions, such as: “Of the people who were hospitalized with a heart attack, who received appropriate follow-up with a specialist? Who was prescribed the appropriate medication on a follow-up basis? What were their comparative mortality rates 5 or 10 years later?”

Linked data also allows for the creation of algorithms that generate cohorts of people with similar health conditions (such as diabetes, asthma, congestive heart failure, or opioid use) and/or healthcare experiences (such as mammography or hip replacement). These algorithms can be enriched when linked data, such as physician claims and hospital inpatient records, are used. Typically, algorithms are validated by primary data collection from medical charts at physician offices or in hospitals. Validated algorithms applied to annual or updated administrative data provide an efficient way to generate cohorts that would otherwise be very expensive to collect over time.

Using administrative health data for research has some challenges, however. Since the data are collected for administrative purposes, they are observational and therefore usually retrospective. They usually do not contain the clinical or sociodemographic detail (such as smoking, socioeconomic status, or medical test results) necessary to answer some research questions or to account for potential confounders of health outcomes. Administrative data may be prone to misclassifying individuals assigned to disease-based cohorts without adequate physician or hospital chart-abstracted person-level record validation. Finally, special legal authorities, privacy laws, and permissions are required to collect and

access these datasets because even when the identifiers in these records are encoded, in rare cases, individual linked records could potentially identify individuals if proper methodologies and access controls are not employed. As noted by Chamberlayne et al., “The ethical issues surrounding access to a resource made up of linked data are more complex than those pertaining to access to a single data source” (Chamberlayne et al. 1998).

Comprehensive and routinely updated documentation, or metadata, is required to fully understand the rationale for the original collection of each variable – documentation is elusive at best and not always available to researchers. Comprehensive metadata is necessary to develop an accurate analytic plan, to assess face validity, and to ensure a reasonable interpretation of the data once analyzed. Currently, there are methodologies in the emerging field of “data quality science” to better standardize the assessment of administrative health data quality and to understand whether the data are “fit” to answer the intended research questions (Lix et al. 2012).

---

## The ICES Data Repository

The Institute for Clinical Evaluative Sciences (ICES) in Ontario, Canada, is a not-for-profit research institute and the steward of a secure and accessible data repository that allows for the development of evidence that makes “policy better, health care stronger and people healthier” (from ICES website [www.ices.on.ca](http://www.ices.on.ca); March 2014). ICES is funded primarily by the Ontario Ministry of Health and Long-Term Care with special initiative funds and investigator-driven peer-reviewed grants. As of April 2014, there were approximately 180 affiliated faculty from around Ontario and about 160 staff whose expertise includes data linkage and analysis, biostatistics, health informatics, epidemiology, project management, research administration, information technology, and database development and support. ICES science is organized across clinical program areas: cancer, cardiovascular, primary care and population health, chronic disease and



pharmacotherapy, health system planning and evaluation, kidney, dialysis and transplantation, and mental health.

Most of the ICES staff are located at ICES Central on the campus of Sunnybrook Health Sciences Centre in Toronto, Ontario, and other affiliated ICES scientists and staff are located across the province: Downtown Toronto, Queen's University in Kingston, the University of Ottawa, Western University in London, and new sites developing at McMaster University in Hamilton and at the Northern Ontario School of Medicine in Thunder Bay.

ICES is the steward of a large comprehensive and linkable data repository used for research and evaluation. The ICES Data Repository consists primarily of health administrative data that are created in the day-to-day interactions with the healthcare system – billings of physicians to the Ontario Health Insurance Plan (OHIP), drug claims to the Ontario Drug Benefit (ODB) Program, discharge summaries of hospital stays (DAD) and emergency department visits (NACRS), and much more. With almost complete health services data coverage of the annual Ontario population from 1991 across most publicly funded healthcare sectors, ICES scientists, analysts, and staff apply scientific methods to advance the evidence for improvements in health and healthcare. The collection and use of these administrative data is authorized by ICES' designation as one of four prescribed entities in Ontario under the Personal Health Information Protection Act 2004 (PHIPA, s.45) – this means that ICES may collect and use personal health information for the purposes of evaluating and monitoring the health system, with adequate data governance permissions and controls.

The ICES Data Repository has the following attributes:

- Individual level: The data reflect people and their health and healthcare experiences, similar to data repositories in British Columbia, Manitoba, Quebec, Nova Scotia, and Newfoundland.
- Longitudinal: Like other jurisdictions, the ICES Data Repository includes most healthcare

experiences over time. The ICES Repository goes back to 1991 and in some cases, earlier.

- Population based: In 2013, there were over 13 million people in Ontario, and since most of the people who are eligible for healthcare are represented, this makes the ICES Repository the largest repository of its type in Canada.
- Comprehensive health sector data: Much of the administrative data in the ICES Repository represent publicly funded physician, hospital and health-based community care, as well as claims for prescription drugs for people aged 65 and over. Population and condition-specific registries are also included, where available. In some provincial data repositories, such as at the Manitoba Centre for Health Policy at the University of Manitoba, additional government administrative data outside the health sector, such as education and social support, are routinely included. At ICES, discussions to broaden the collection and use of data beyond the health sector have begun.
- Desensitized and linkable with coded identifiers: Individuals in the Repository are uniquely identified with an ICES-specific key number (IKN) which is obtained by encoding the Ontario health card number using a proprietary encoding algorithm. ICES in-house professionals replace any direct identifiers attached to the incoming data with a unique IKN that is used to link person-level records from one dataset to another. This in-house expertise that spans informatics and research has allowed for the easy integration of data with high data quality standards.
- Easy to use: All data are in an SAS format and ready to use in an analytic environment – these data are linkable to each other using a unique person-level identifier and ready to use after appropriate data access approvals. Having the data repository organized in this manner creates efficiencies for research as the data are already in record-level format.
- Secure and privacy protected: Ontario privacy legislation (Personal Health Information Protection Act – PHIPA 2004) allows for ICES to

collect direct identifiers from data custodians for the purpose of assigning an IKN to each data record. ICES' privacy policies, practices, and procedures and our prescribed entity status under PHIPA allow ICES to function with the approval of the Ontario Information and Privacy Commissioner (IPC). A full review of ICES privacy and security policies and procedures is undertaken every 3 years, with the approval letter from the IPC published on the ICES website (more detail on this below). Expert information and technology staff are on site to ensure the security and smooth maintenance of the research platform.

- Professional data management: Data quality and informatics experts apply the highest data quality standards and are leading in developing metadata and other documentation for the analysts and scientists to use.

The comprehensive collection in the ICES Data Repository is the basis of population-based examination of groups of people with particular health conditions (such as diabetes or cancer) or people who have had similar health services experiences (such as hip or knee surgery) or how the health system is working (performance indicators or continuity of care) and outcomes (length of hospital stay, emergency department visits, or death) over time.

The records in the ICES Data Repository include:

- Records of Ontarians' day-to-day interactions with the healthcare system: Physician claims submitted to the Ontario Health Insurance Plan, medical drug claims to the Ontario Drug Benefit Program, discharge summaries of hospital stays and emergency department visits, claims for home care, information about long-term care, and more.
- Special registry collections include Ontario Cancer Registry (Cancer Care Ontario), the Ontario Stroke Registry (ICES collection), Registry of the Cardiac Care Network, federal immigration information, an Ontario birth outcomes registry (Better Outcomes Registry and Network – BORN), and others.
- Derived chronic condition cohorts have been developed at ICES using linked data algorithms that have been validated by using primary data collection as a gold standard.
- Detailed clinical data has been extracted from electronic medical records and through ICES primary data collection projects.
- Population and demographic data through the Ministry of Health's Registered Persons Database (RPDB) is used to characterize study subjects and to generate denominators for rate calculation.
- Additional clinical data, agency client-level data, and research data collections that are linkable to longitudinal outcome data are included on a project-by-project basis.

A full listing of the data in the ICES Data Repository can be found on the ICES website.

### **Privacy, Data Governance, and Access to Data at ICES**

ICES is designated as a prescribed entity under the Ontario Personal Health Information Protection Act (PHIPA 2004 (s. 45[1] and O. Reg 329/04 section 18[1])). As a prescribed entity, health information custodians (HICs), such as healthcare practitioners, hospitals, laboratories, nursing homes, and community care access centers, including the Ministry of Health and Long-Term Care, may disclose personal health information (PHI) and associated information relating to their patients to ICES for purposes of "analysis or compiling statistical information with respect to the management of, evaluation or monitoring of, the allocation of resources to or planning for all or part of the health system, including the delivery of services" (PHIPA s.45(1)). Health Information Custodians and other data partners may also disclose personal health information and associated clinical data to ICES that are collected through approved research projects under the appropriate oversight of a Research Ethics Board (REB) and the authorities prescribed under PHIPA (s. 44(1)). As with all prescribed entities in Ontario, ICES security and privacy standard

operating procedures and policies are reviewed and approved by the Information and Privacy Commissioner of Ontario every 3 years.

The authority for ICES to hold and integrate data lies within detailed data sharing agreements or memoranda of understanding with every data partner. A data sharing agreement executed for every dataset integrated into the Repository outlines the legal authorities, the data collection and transfer methods, the desensitization procedures, and the use for each new dataset that ICES collects. The most comprehensive data sharing agreement is with the Ontario Ministry of Health and Long-Term Care, and this agreement outlines ICES' responsibility in using the Ontario health administrative data.

ICES' policies, practices, and procedures that prescribe the governance of the Repository overall and of each dataset at ICES are strictly followed – the use of the data at ICES is limited to the agreed-upon purpose and use defined in the data sharing agreement under which the data is authorized for ICES to collect.

### **Access to ICES Data**

Research at ICES is generally managed within clinical program areas: cancer, cardiovascular, population health and primary care, chronic disease and pharmacotherapy, health system planning and evaluation, kidney, dialysis and transplantation, and mental health and addictions. As well, ICES currently has four active satellite sites: ICES UofT at the University of Toronto, ICES Queen's in Kingston, ICES uOttawa, and ICES Western (ICES at McMaster University and ICES North at Lakehead/Laurentian University are being developed). Scientists and staff are affiliated with these programs. When a fully formed project is contemplated by an ICES scientist, the feasibility and rationale for its implementation is vetted by ICES program leads and management staff: Is the project aligned with the ICES mission? Can the question be answered with the data available (or new data collected)? What is the human resource capacity to implement the project – analyst and project management or coordination resources? Are there adequate funds to implement the project? After these criteria are vetted, a

privacy impact assessment (PIA) is completed by research teams outlining the project research protocol, the data being contemplated for the project, the output of the research, and the foreseeable privacy impacts or risks. The ICES privacy office reviews all privacy impact assessments and provides recommendations and final approval before any data can be accessed for projects. In some cases and according to data sharing agreements, the data custodian is notified or approves the use of their data for ICES projects and they receive a copy of reports that utilized their data. All ICES projects at a minimum undergo Research Ethics Board (REB) retrospective review – currently Sunnybrook Health Sciences Centre REB is the overseeing body for most ICES projects.

---

### **Record Linkage and Desensitizing the Data for Research**

The ICES Data Repository is continuously growing. Mostly, the data collected at ICES initially contains direct identifiers so that the records attributed to a unique individual can be assigned the correct ICES key number (IKN) and the direct identifiers removed. This process of desensitizing data for research at ICES may be facilitated by record linkage (also known as record matching) – a process by which records from two files are combined so that an individual's information from one file can be merged with the same individual's information from another file. For example, you may have one file of demographic data and another file of diagnostic patient information, and you want to combine and analyze them together. If both files contain a precise identifier that refers to the same person (such as health card number or social insurance number), the linkage task is relatively easy. This is called deterministic record linkage.

At ICES, not all individual-level data received contain Ontario health card numbers. Frequently, individuals are identified in the data records by their name, postal code, and other "soft" identifiers. Before data can be used for research, the IKN for these records must be found. Linkage to other fields may be used to match individuals from

different files. These, as listed below, come with some challenges.

Last name:

- Not unique between people (common names may be shared by numerous individuals)
- Subject to misspelling
- May change over time (e.g., at marriage)

First name:

- Similar issues as last names
- Nicknames may be used in one file and full names in the other

Date of birth:

- Subject to transcription and other errors
- Imprecise when supplied by someone other than the individual (e.g., family member at hospitalization)
- May be incomplete
- Not unique

Date of death:

- Similar issues as date of birth
- May only be applicable to a portion of the file

Location of personal residence such as postal code:

- Subject to change over time (as people move)
- Nonunique, in particular within families

To combine files that only contain imprecise direct identifiers such as those above, probabilistic record linkage (PRL) may be used. Another common term for PRL is “fuzzy matching.”

Probabilistic record linkage methodologies incorporate the relative frequencies of field values to compute their sensitivity and the positive predictive value and then combine these to form linkage weights for each pair of records. For example, if two records contain the same name, a greater weight is given if that name is rare in the population being studied. Conversely, two records sharing the same value that is quite common (e.g., birth year or female gender) may not contribute much to the linkage weight. Various encoding algorithms and string comparators are used to deal with alternate spellings, nicknames, and common transcription errors. Blocking is used to

reduce the total number of comparisons; and clerical review is applied to pairs that did not yield a conclusive weight.

The Registered Persons Database (RPDB) was described earlier in this chapter. ICES receives a number of RPDB files monthly and thus has a cumulative record of the names, postal codes, and other demographic information for all health card holders in Ontario over time. This file is an essential component of making files without HCNs useable for research.

Figure 1 illustrates the process of assignment of the ICES key number. Once an IKN is assigned to a record and the original direct identifiers are removed, that record is considered “desensitized” and can be (deterministically) linked to all other records in the ICES Data Repository that pertain to the same person. This facilitates the creation of analytic datasets that are prepared to answer specific research questions.

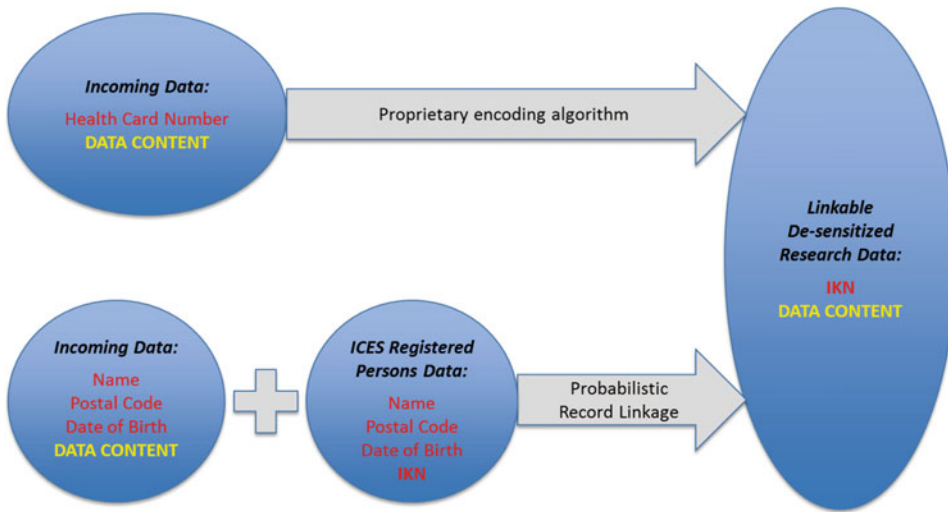
Other institutions that may not have the equivalent of the RPDB may find other solutions. For example, Chamberlayne et al. (1998) describe the creation of a Linkage Coordinating File (LCF) at the Centre for Health Services and Policy Research at the University of British Columbia. This file was created by applying probabilistic record linkage to data from various sources and contains personal identifiers and a unique person-level index. The file can be used to facilitate the linkage of other person-level files, in a way similar to the RPDB.

---

## Data Documentation, Metadata, and Data Quality Assessment

### Data Quality Assessment in the Literature

There are many frameworks and evaluation strategies for data quality, and many are created for specific purposes and types of data. Data quality assessment has been defined as “the whole of planned and systematic procedures that take place before, during and after data collection to be able to guarantee the quality of data in a database. . .for its intended use” (Arts et al. 2002).



**Fig. 1** Process for assignment of ICES key number at ICES, with and without Ontario health card number

Holt and Jones suggested that “data quality is not so much an absolute property of a statistical estimate but is related to the purpose for which the estimate is used” (Holt and Jones 1999, p. 24). When using administrative data, it is difficult to “guarantee” data quality; however, a robust assessment focusing on the linked data’s intended use and purpose will at least characterize the quality in an interpretable way.

Generally, the following domains and questions need to be examined when assessing data quality:

- Accuracy: Do the data reflect the truth?
- Validity: Do the data reflect what they were designed to reflect?
- Completeness: Do the data include all records that are collected? Have the fields been well populated?
- Comprehensiveness and coverage: Do the data cover 100 % of the intended population? Alternatively, do they constitute a representative sample?
- Reliability: Are the data reproducible?
- Timeliness: Is there a short lag between data collection and use?
- Linkability: Can the data be connected to other data to reflect healthcare system complexity?
- Privacy: Do the data adhere to jurisdictional privacy laws? Are there appropriate and auditable privacy preserving procedures and practices? Has the risk been sufficiently reduced by removing sensitive information?
- Usability: Are the data organized, accessible, and provided in a format that can be easily used?
- Currency: What is the time lag between the time period reflected in the data and the time that data are ready for use?

A number of organizations have developed data quality frameworks to assess the data in their repositories. For example, the Canadian Institute for Health Information (CIHI) framework includes dimensions of relevance, timeliness, usability, accuracy, and comparability within an envelope of planning, implementing, and assessing (CIHI 2009).

Researchers at the Manitoba Centre for Health Policy have developed a data quality framework that has been broadly adopted by ICES (Azimae et al. 2013). In that framework, dimensions of data quality are divided between those that can be assessed at the database level, versus those that can be assessed at the research level. In particular they described database-specific data quality dimensions as:

- Accuracy: Completeness (rate of missing values) and correctness (invalid codes, invalid dates, out of range, outliers, and extreme observations)
- Internal validity: Internal consistency, stability across time, and linkability
- External validity: Level of agreement with literature and available reports
- Timeliness: Currency of posted data, time to acquisition, and time to release for research purposes
- Interpretability: Availability, quality and ease of use of documentation, policies and procedures, format libraries, metadata, and data model diagrams

Data quality should also be assessed within a specific research project, where conclusions may be drawn about the accuracy and reliability of data, measurement error, bias, and agreement with other databases and other sources. According to Roos and others (Roos et al. 1989, 2005), data quality assessment can be carried out through comparisons of linked information across datasets used for the project.

In 2007 ICES researchers undertook an environmental scan of data quality assessments (Iron and Manuel 2007). Their conclusions from the environmental scan were:

- Quality should be routinely and systematically evaluated for all generally-used data.
- Data quality is contextual
- The evaluation and interpretation of data quality depends on the purpose for which the data are being used.
- The constructs of accuracy and validity are often confused.
- Accuracy (or truth) is an elusive construct and should not be expected.
- The most common ways to evaluate validity are concordance, comparability and inter-database reliability.
- Linked data, where available, should be used to evaluate data quality (when primary data collection is not feasible).
- There is a need for more investigation into evaluating data quality.
- The relevance of every data quality assessment requires full discussion. (p. 8)

They proposed an end-to-end data quality framework for projects using linked data. The

Quality Assessment of Administrative Data (QuAAD) framework leverages the traditional data quality framework and adds a number of domains that aim to help to develop data partnerships and improvements for health data:

- Context: What is the purpose of the project and data evaluation? Who are the key stakeholders and who is using the data? What is the purpose of the data collection? What is the political environment?
- Issues: Who is the target population and where do they live? What are the outcomes of the project – for example, quality of care, appropriateness, timeliness, mortality, and service use? What are the predictors of and influences that may affect the outcomes of the project, such as system characteristics?
- Data and sources: What data are being used? Who are the data custodians? What data elements are being used? Will the data be linked? What are the authorities for data use?
- Measurement: These are the usual data quality indicators: e.g., timeliness, reliability, completeness.
- Appraisal: Summary of data quality; stakeholder report; identification of data improvement opportunities
- Implementation: If opportunities or gaps are identified, how will these be addressed? (Discussions and next steps with data custodians)

### Data Documentation, Data Quality Assessment, and Metadata at ICES

Currently at ICES, a systematic and holistic approach is taken to documenting each dataset and assessing its completeness, correctness, stability, and linkability. Figure 2 summarizes the approach taken. Metadata information (such as the description of datasets, variables, and valid values) is extracted from the data repository into a metadata repository. This information in turn is used to produce the data dictionary, as well as data quality assessments. By utilizing a “single source of truth,” consistency between the data, the data dictionary, and the data quality assessments is assured.

### Data Documentation at ICES

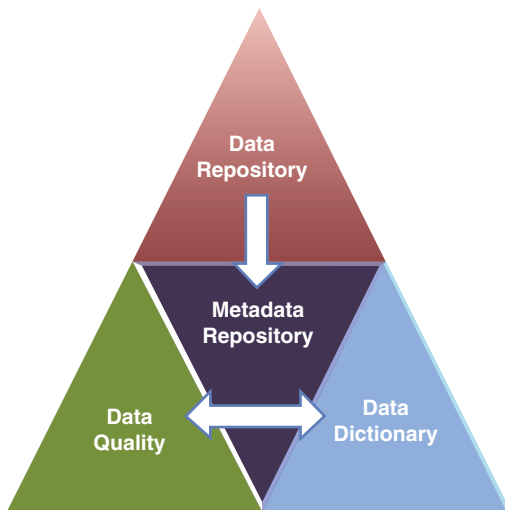
A detailed data dictionary is an essential tool in research. The *IBM Dictionary of Computing* defines a data dictionary as “a centralized repository of information about data such as meaning, relationships to other data, origin, usage, and

format” (ACM 1993). The information in the data dictionary should contain, at a minimum:

- The name and brief description of each file
- A list of fields and their description
- For each field, a list and description of valid values
- Unstructured or semi-structured comments with additional information

The files in the ICES Data Repository are stored as SAS datasets. ICES leverages certain features of the SAS software to create data dictionaries that correspond dynamically to the data files. In particular:

- Dataset labels are used to describe the contents of each file.
- Similarly, variable labels describe each field.
- A central format catalogue contains descriptions of all valid discrete values of each field of all datasets.



**Fig. 2** Holistic approach to data documentation and data quality assessment at ICES

Internal experts for each of the datasets have been identified. These individuals enrich the data dictionary with their insights and experience by means of comments.

A dot-net application displays this information in a user-friendly online data dictionary. Additional information, such as the expected date of the next update, is added manually. Most of the information is displayed for the public on the ICES website. Certain fields (such as the name of the internal expert) are only available internally. Figure 3 is an excerpt from the ICES Data Dictionary describing the variable admission date (ADMDATE) in the OHIP data library.

This approach to data documentation at ICES has a number of advantages over the more

Label:	Date of hospital admission
Type Length:	NUM:
Available to:	CHP1991, CHP1992, CHP1993, CHP1994, CHP1995, CHP1996, CHP1997, CHP1998, CHP1999, CHP2000, CHP2001, CHP2002, CHP2003, CHP2004, CHP2005, CHP2006, CHP2007, CHP2008, CHP2009, CHP2010, CHP2011, CHP2012, CHP2013, CHPLAB1991, CHPLAB1992, CHPLAB1993, CHPLAB1994, CHPLAB1995, CHPLAB1996, CHPLAB1997, CHPLAB1998, CHPLAB1999, CHPLAB2000, CHPLAB2001, CHPLAB2002, CHPLAB2003, CHPLAB2004, CHPLAB2005, CHPLAB2006, CHPLAB2007, CHPLAB2008, CHPLAB2009, CHPLAB2010, CHPLAB2011, CHPLAB2012, CHPLAB2013, CHPLAB2014
UNIX Access Group:	LEVEL2
Format:	DATE
Value:	
Notes:	<ul style="list-style-type: none"> <li>• Filled in by billing physician's office.</li> <li>• Mandatory: No</li> <li>• Only used if the patient is admitted to the hospital, but is not necessarily filled in, even so. If it is provided, it is likely to be accurate. Date is missing in approximately 90% of the records.</li> </ul>
Links:	

**Fig. 3** Excerpt of ICES Data Dictionary (Source: ICES Data Dictionary <https://datadictionary.ices.on.ca/Applications/DataDictionary/Variables.aspx?LibName=OHIP&MemName=&Variable=ADMDATE>)

traditional manual methods. Since information is based on actual data elements, there is internal consistency between the data and the documentation. The process of creating a data dictionary for a new dataset is automated and quick, so that a data dictionary can be made available immediately at the same time the data is posted. And finally, if errors are discovered, they are corrected in both the data and the documentation.

### Data Quality Assessment at ICES

ICES' holistic approach to assessing the data quality includes a variety of tools that are used to assess and document data quality, including:

- All the data elements in a dataset are displayed in a "VIMO report," which summarizes the valid, *invalid*, *missing*, and *outlier* rates. Examples of invalid values are listed. Simple descriptive statistics are also displayed and frequencies or histograms are linked to each field. ID variables are highlighted, and their uniqueness status is described.
- A trend analysis of the number of observations over time is performed, and the results are displayed graphically.
- The percent of records that are linkable to the rest of the ICES Data Repository is displayed over time.
- Missing values over time are presented visually, so substantial changes can be easily detected.
- Content experts are identified for each of the datasets. These content experts are expected to be familiar with the data quality assessment for their dataset and detect any issues that need to be addressed.
- All data users participate in a data blog, in which questions and issues are discussed and, when appropriate, acted upon.

Figure 4 illustrates of a VIMO assessment of a client intake dataset. Variable names are hyperlinked to additional univariate descriptions. For example, for numeric values, a histogram is presented, and for nonunique ID variables, frequencies of the number of records per ID are displayed.

### New Data, New Uses, and New Ideas

Health administrative data, particularly in the context of universal healthcare coverage, present a tremendous opportunity to conduct health and healthcare research. Linkable population-based data, with the appropriate privacy and security safeguards, are a resource for examining population- and disease-based cohorts, trends in health services utilization, prevalence and incidence trends, and effects of policy and system changes, among others. Expertise and care must be applied to use such data effectively and optimally.

Administrative data are also collected outside the health sector for managing social programs or educational systems. As with many similar data repositories in Canada and around the world, ICES is exploring the expansion of its linkable data holdings to include non-health administrative data from across the provincial and federal government and social service agencies. For example, a new research program at ICES focusing on mental health and addictions (MHA) was launched in 2013 where the need for integrating community addictions and mental health agency data with health data is critical to understanding prevention, early detection, and timely and sustained appropriate care which in many cases is done in a community setting outside the medical model. Although much of the routine health data to support this program already exists at ICES, a comprehensive evaluation of the full spectrum of MHA care requires linkable person-level data that are generated from education, social support, youth justice and child and youth services sectors for example.

Around the world, discussions about linking biobank and genomic data, electronic medical record data, and other large data collections with each other and with administrative data are propelling the field of big data repositories and analytics into new and uncharted paradigms. Innovative data collection tools, dynamic and privacy-protecting record linkage models, data use, and governance frameworks and technologies are quickly advancing to keep up with the amount and the scope of data being generated and the research and private sector demands that depend on linking disparate datasets.



**VIMO for dataset INTAKE (N=5,168,489)**

ID variables (2)				
Variable Name	Variable Label	% Valid	% Missing	Comments
IKN	ICES Key Number	99.90	0.01	Not unique
Refid	Encrypted ID	100.00	0.00	Unique

Numeric Variables (3)										
Variable Name	Variable Label	% Valid	% Invalid	% Missing	% Outlier	Min	Max	Mean	Median	STD
AGE	Age in years	100.00		0.00	0.00	0.00	999.00	47.80	52.90	29.40
ABS_CC	Agressive Behaviour Scale	76.24		14.72	9.04	0.00	12.00	1.10	0.00	2.20
CPS_CC	Cognitive Performance Scale	86.78		13.22	0.00	0.00	6.00	2.90	3.00	2.20

Character Variables (6)						
Variable Name	Variable Label	% Valid	% Invalid	% Missing	Values	Comments
AssessReas	Reason for Assessment	100.00		0.00	A, N	
LHIN	Local health integration network (LHIN) at admission	100.00		0.00	01, ..., 14	
LivArr	Living arrangement at admission	98.15		1.85	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	
RefBy	Source of referral	100.00		0.00	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	
ProgType	Program request	100.00		0.00	09, 10, 11, 12, 17, 18, 20	Invalid codes: 22, 65
Sex	Client's gender	100.00		0.00	F, M, O, U	

Date Variables (3)				
Variable Name	Variable Label	% Valid	% Missing	Date Range
AssessStartDate	Assessment Start Date	77.66	22.34	10-Oct-1997 to 31-Aug-2014
AssessEndDate	Assessment End Date	90.15	1.85	11-Nov-1997 to 21-Sep-2014
RefDate	Referral Date	72.57	27.43	04-Jan-1997 to 30-Sep-2014

**Fig. 4** Example of VIMO (valid, invalid, missing, outlier) data quality assessment at ICES

## References

- ACM. IBM dictionary of computing, 10th edn; 1993. <http://www-03.ibm.com/ibm/history/documents/pdf/glossary.pdf>
- Arts D, de Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc* 2002;9:600–11. <http://jamia.bmj.com/content/9/6/600.full>. Accessed 17 June 2014.
- Azimaee M, Smith M, Lix L, Ostapyk T, Burchill C, Hong SP. MCHP data quality framework. Winnipeg: Manitoba Centre for Health Policy, University of Manitoba; 2013. [www.umanitoba.ca/faculties/medicine/units/community\\_health\\_sciences/departmental\\_units/mchp/protocol/media/Data\\_Quality\\_Framework.pdf](http://www.umanitoba.ca/faculties/medicine/units/community_health_sciences/departmental_units/mchp/protocol/media/Data_Quality_Framework.pdf)
- Canadian Institute for Health Information. The CIHI data quality framework, 2009. Ottawa: CIHI; 2009. [http://www.cihi.ca/CIHI-ext-portal/pdf/internet/DATA\\_QUALITY\\_FRAMEWORK\\_2009\\_EN](http://www.cihi.ca/CIHI-ext-portal/pdf/internet/DATA_QUALITY_FRAMEWORK_2009_EN)
- Chamberlayne R, Green B, Barer ML, Hertzman C, Lawrence WJ, Sheps SB. Creating a population-based linked health database: a new resource for health services research. *Can J Public Health*. 1998;89(4):270–3.
- Dunn HL. Record Linkage. *Am J Public Health Nation Health*. 1946;36:1412.
- Friedman D, Hunter E, Parrish II G. Defining health statistics and their scope. In: Friedman DJ, Hunter EL, Gibson Parrish II R, editors. *Health statistics: shaping policy and practice to improve the population's health*. New York: Oxford University Press; 2005. p. 16.
- Holt T, Jones T. Quality work and conflicting quality objectives. In: *Quality work and quality assurance within statistics*. Eurostat Proceedings; 1999. p. 15–24. [http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/DGINS%20QUALITY%20Q98EN\\_0.pdf](http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/DGINS%20QUALITY%20Q98EN_0.pdf). Accessed 17 June 2014.
- ICES. website [www.ices.on.ca](http://www.ices.on.ca)
- Iron K, Manuel DG. Quality assessment of administrative data (QuAAD): an opportunity for enhancing Ontario's health data. ICES investigative report. Toronto: Institute for Clinical Evaluative Sciences; 2007.
- Lix LM, Neufeld SM, Smith M, Levy A, Dai S, Sanmartin C, Quan H. Quality of administrative data in Canada: a discussion paper 2012. Background for science of data quality invitational workshop. Ottawa; 2012. [http://www.usaskhealthdatalab.ca/wp-content/uploads/2012/09/Workshop\\_Background-Documents-updated-AppendixD.pdf](http://www.usaskhealthdatalab.ca/wp-content/uploads/2012/09/Workshop_Background-Documents-updated-AppendixD.pdf)LisaLix
- Personal Health Information Protection Act (PHIPA). 2004. [http://www.e-laws.gov.on.ca/html/statutes/english/elaws\\_statutes\\_04p03\\_e.htm](http://www.e-laws.gov.on.ca/html/statutes/english/elaws_statutes_04p03_e.htm)
- Roos LL, Sharp SM, Wajda A. Assessing data quality: a computerized approach. *Soc Sci Med*. 1989;28(2):175–82.
- Roos LL, Gupta S, Soodeen RA, Jebami L. Data quality in a data rich environment: Canada as an example. *Can J Aging*. 2005;24 Suppl 1:153–70.



# Health Services Data: The Centers for Medicare and Medicaid Services (CMS) Claims Records

# 4

Ross M. Mullner

## Contents

<b>Introduction</b> .....	62
<b>Major Healthcare Programs</b> .....	63
Medicare .....	63
Medicaid .....	63
Children’s Health Insurance Program .....	64
<b>Information and Data Products</b> .....	64
<b>Information Products</b> .....	64
Publications .....	65
Data Navigator .....	65
Interactive Dashboards .....	65
<b>Data Products</b> .....	66
Medicare and Medicaid Public Use Data File .....	66
Chronic Conditions Data Warehouse .....	71
Medicare Current Beneficiary Survey .....	72
Medicare Qualified Entity Program .....	73
<b>Conclusion</b> .....	73
<b>References</b> .....	74

### Abstract

The US Centers for Medicare and Medicaid Services (CMS) is the largest purchaser of healthcare in the nation – serving almost 123 million people, more than one in three Americans. CMS is responsible for administering and overseeing three of the nation’s largest ongoing

healthcare programs: Medicare, Medicaid, and the Children’s Health Insurance Program (CHIP). The Medicare program provides government-sponsored health insurance for people 65 or older and under age 65 with certain diseases and disabilities. The Medicaid program, which is a joint state-federal program, provides healthcare for the poor. CHIP is a grant program that provides health insurance to targeted low-income children in families with incomes above Medicaid eligibility levels. CMS sponsors many data and information initiatives for health

R. M. Mullner (✉)  
 Division of Health Policy and Administration, School of Public Health, University of Illinois, Chicago, IL, USA  
 e-mail: [rmullner@comcast.net](mailto:rmullner@comcast.net)

services researchers, policymakers, educators, students, and the general public. In 2014, CMS established the Office of Enterprise Data and Analytics (OEDA) to better oversee and coordinate its large portfolio of data and information. The office also funds the privately run Research Data Assistance Center (ResDAC), which provides training and technical assistance to individuals requesting the agency's data files. CMS information products include an online research journal *Medicare and Medicaid Research Review* (MMRR); other publications including *Medicare and Medicaid Statistical Supplement*, *Statistics Reference Booklet*, and *CMS Fast Facts*; a data navigator; and several interactive dashboards. Its data products include numerous Medicare and Medicaid public use data files, the Chronic Conditions Data Warehouse (CCW), the Medicare Current Beneficiary Survey (MCBS) files, and the Medicare Qualified Entity (QE) Program. Many examples of CMS' information and data products are highlighted and discussed.

---

## Introduction

The Centers for Medicare and Medicaid Services (CMS) is a major agency within the US Department of Health and Human Services (DHHS). CMS (previously known as the Health Care Financing Administration or HCFA) is responsible for administering and overseeing three of the nation's largest ongoing healthcare programs: Medicare, Medicaid, and the Children's Health Insurance Program (CHIP). In addition, CMS is responsible for implementing various provisions of the Patient Protection and Affordable Care Act (ACA) of 2010, including the construction of an insurance exchange or marketplace, consumer protections, and private health insurance market regulations. In 2015, CMS through its various programs served almost 123 million people, more than one in three Americans, making it the single largest purchaser of healthcare in the United States.

CMS' stated mission is "as an effective steward of public funds, CMS is committed to

strengthening and modernizing the nation's health care system to provide access to high equality care and improved health at lower cost" (CMS 2015).

Headquartered in Baltimore, Maryland, with other offices in Bethesda, Maryland, and Washington, DC, ten regional offices located throughout the nation, and three antifraud field offices, CMS employs about 5,900 federal employees. CMS employees in Baltimore, Bethesda, and Washington, DC, develop healthcare policies and regulations, establish payment rates, and develop national operating systems for programs. Regional office employees provide services to Medicare contractors; accompany state surveyors to hospitals, nursing homes, and other facilities to ensure health and safety standards; and assist state CHIP and Medicaid programs. CMS employees also work in offices in Miami, Los Angeles, and New York, cities known to have high incidences of healthcare fraud and abuse.

Operationally, CMS consists of 15 major divisions, including seven centers: Center for Strategic Planning, Center for Clinical Standards and Quality, Center for Medicare, Center for Medicaid and CHIP Services, Center for Program Integrity, Center for Consumer Information and Insurance Oversight, and Center for Medicare and Medicaid Innovation.

CMS also has a number of operational offices. One office that will increasingly play an important role in data and information initiatives is the Office of Enterprise Data and Analytics (OEDA). Established in 2014 and managed by CMS' first chief data officer (CDO), the OEDA is tasked with overseeing improvements in the agency's data collection and dissemination activities. It will work to better harness CMS' vast data resources to guide decision-making, promoting greater access to the agency's data to increase higher-quality, patient-centered care at lower costs. The OEDA also manages the CMS-funded Research Data Assistance Center (ResDAC) at the University of Minnesota, which conducts education and training programs and provides assistance to researchers who want to access the agency's data files (Brennan et al. 2014). In 2015, CMS' budget totaled an estimated \$602 billion (CMS 2015).

## Major Healthcare Programs

### Medicare

Established in 1965, Medicare (Title XVIII of the Social Security Act) is a federal health insurance program for people 65 or older, those under age 65 with certain disabilities, people of any age with end-stage renal disease (permanent kidney failure requiring dialysis or a kidney transplant), and individuals with amyotrophic lateral sclerosis (ALS) commonly known as Lou Gehrig's disease. In 2015, a total of 55.2 million individuals were enrolled in Medicare in the nation.

Medicare consists of four separate parts: Medicare Part A (Hospital Insurance), Medicare Part B (Medical Insurance), Medicare Part C (Medicare Advantage plans), and Medicare Part D (Medicare Prescription Drug Coverage).

Medicare Part A provides insurance coverage for hospital inpatient care (covering stays in a semiprivate room, meals, general nursing and other hospital services, and supplies), skilled nursing facility care (covering up to 100 days in a semiprivate room, skilled nursing and rehabilitation services, and other services and supplies, following a hospital stay), home health care services (covering part-time or intermittent skilled nursing care, physical therapy, speech language pathology, and occupational therapy), and hospice care (covering drugs for pain relief and medical and support services).

Medicare Part B provides insurance coverage for necessary medical services (covering physician services, outpatient medical and surgical services and supplies, diagnostic tests, and durable medical equipment (DME)), clinical laboratory services (covering blood tests, urinalysis, and other screening tests), home health care services (covering part-time or intermittent skilled nursing care, physical therapy, speech language pathology, and occupational therapy), and outpatient hospital services (covering hospital services and supplies).

Medicare Part A and B are known as Original Medicare. Most healthcare services provided to beneficiaries enrolled in Original Medicare are paid for on a fee-for-service basis. Most Medicare

beneficiaries enroll in both Part A and Part B. Often beneficiaries do not pay premiums for Part A, because they have worked for 40 quarters and paid into the Social Security System. But, they do have to pay monthly premiums for Part B.

Medicare Part C, or Medicare Advantage, is a managed care program. Medicare Advantage plans combine Medicare Part A and Part B and often provide additional benefits that Original Medicare does not cover such as dental, hearing, vision care, and prescription drug coverage. Depending upon the particular managed care plan, Medicare Advantage can cost beneficiaries less and provide more benefits than Original Medicare. Medicare Advantage plans are run by private companies that contract with CMS to provide covered services. The types of plans include Health Maintenance Organizations (HMOs), Preferred Provider Organizations (PPOs), Private Fee-for-Service (PFFS) plans, and Special Needs Plans (SNPs) (Office of the Assistant Secretary for Planning and Evaluation 2014).

Medicare Part D is a voluntary outpatient prescription drug benefit for Medicare beneficiaries. The Medicare program provides the drug benefit through either Medicare Advantage plans or private standalone prescription drug plans approved by CMS. The prescription drug plans vary in terms of the type of drugs they cover and their coinsurance and deductible costs. In 2015, there were 1,001 prescription drug plans in the nation (Blumenthal et al. 2015; Henry J. Kaiser Family Foundation 2014).

### Medicaid

Established in 1965 along with the federal legislation that created the Medicare program, Medicaid (Title XIX of the Social Security Act) is a joint federal-state healthcare program for the poor. The federal government provides the states with matching contributions to help fund the various Medicaid programs. States design and administer their own Medicaid programs, determining eligibility standards, benefit packages, and payment rates under broad federal guidelines. As a result, state Medicaid programs vary greatly in size,

scope, and generosity. For example, a low-income individual may be eligible for Medicaid in one state, but not in another. In 2015, an estimated 66.7 million individuals were receiving Medicaid benefits in the nation.

Medicaid originally only provided healthcare services for certain categories of the poor such as pregnant women, children, parents with young children, the elderly, and blind and disabled individuals. The Affordable Care Act (ACA) of 2010 greatly expanded the Medicaid program to cover millions of uninsured Americans. Under the new law, many states have expanded their Medicaid programs to cover nearly all non-elderly poor adults (Henry J. Kaiser Family Foundation 2015; Orentlicher 2015).

Medicaid is a very important payer for infants and the elderly and younger individuals with significant disabilities. It pays for about half of all births in the nation. And Medicaid is the nation's only safety net for people who need long-term care services. About a third of Medicaid spending pays for personal assistance in nursing homes and at home for people who need help with the basic tasks of daily living (Feder and Komisar 2012).

Some individuals, known as dual eligible beneficiaries, receive both Medicaid and Medicare benefits. They are enrolled in Medicare Part A and/or Part B and receive some form of Medicaid benefits. In 2015, about 9.6 million individuals were dually eligible in the United States (Cohen et al. 2015; Henry J. Kaiser Family Foundation 2015).

### Children's Health Insurance Program

Established in 1992 and reauthorized several times, the state Children's Health Insurance Program (CHIP) (Title XXI of the Social Security Act) is a program that provides federal funds to states and matches state contributions to provide health insurance to children who do not qualify for Medicaid. Specifically, CHIP provides health insurance for children less than 19 years of age whose families are ineligible for Medicaid. While state benefit plans vary, all CHIP plans cover immunizations, prescription medications, routine physician visits, dental care, medically necessary

orthodontics, mental and behavioral health, hospitalizations, home health care, rehabilitation care, medical equipment, and laboratory and x-ray services. In 2015, about 6.2 million children were enrolled in CHIP (Ewing 2008; National Conference of State Legislature 2014).

---

## Information and Data Products

Each year CMS collects and processes enormous amounts of data. For just the Medicare program alone, CMS and its contractors process more than 1.3 billion claims a year and generate billions of other non-claims data, such as eligibility checks, queries from telephone contacts through its toll-free 1-800 MEDICAR(E) help line, patient experience surveys, and enrollment information. Additionally, CMS collects data on its Medicare and Medicaid Electronic Health Record (EHR) Incentive Programs and on health insurance exchanges or marketplaces coverage.

In the past, CMS tended to view the data and information it produced as only by-products of its operations. Today, however, the development, management, use, and dissemination of data and information resources have become one of CMS' core functions. To become more transparent and accountable, CMS is increasingly making more of its data and information available to researchers, policymakers, educators, students, and the general public. By releasing these resources, CMS is attempting to leverage its data and information to better evaluate and improve its programs, facilitate healthcare innovation, develop new products and analysis tools, and highlight actionable information for internal and external policy- and decision-makers (CMS 2012).

---

## Information Products

CMS produces many information products that are readily available to researchers and the general public. These products include numerous publications, a data navigator, and several interactive dashboards. Examples of some of the major information products are described below.

## Publications

For health services researchers and policy analysts, CMS publishes a peer-reviewed online journal, the *Medicare and Medicaid Research Review* (MMRR). The journal (previously titled the *Health Care Financing Review*) publishes research articles throughout the year on a continuous basis. The articles address various topics such as trends in Medicare, Medicaid, and CHIP, access and quality of care issues, healthcare insurance coverage, and payment for health services. It also includes CMS News and Data Briefs. Issues of MMRR, as well as the entire run of the *Health Care Financing Review* (Vols. 1–39; 1979–2009), can be accessed at: [www.ncbi.nlm.nih.gov/pmc/journals/2404](http://www.ncbi.nlm.nih.gov/pmc/journals/2404).

CMS publishes annual data in its *Medicare and Medicaid Statistical Supplement*. This comprehensive statistical supplement is updated on an ongoing basis by section as the data becomes available. Consisting of 14 chapters, including 115 tables and 67 charts, the supplement provides detailed tables on the personal healthcare expenditures for the entire US population; characteristics of the Medicare program including enrollment, program payments, cost sharing, utilization of short-stay hospitals, skilled nursing facilities, home health agencies, hospices, physician services, hospital outpatient services, end-stage renal disease services, managed care, and Medicare Part D; and characteristics of the Medicaid program including the number of persons served, their demographic characteristics, and the types of services they received. Current and past statistical supplements (2001 to the present) can be accessed at: [www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareMedicaidStatSupp/2013.html](http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareMedicaidStatSupp/2013.html).

CMS also publishes an abridged version of the statistical supplement entitled *CMS Statistics Reference Booklet*. This quick reference guide summarizes information about national healthcare expenditures and the Medicare and Medicaid programs. Published in June of each year, the booklet provides the most currently available information. Booklets are available online for 2003 through the most currently available complete calendar year, at: [www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareMedicaidStatSupp/2013.html](http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareMedicaidStatSupp/2013.html).

[Data-and-Systems/Statistics-Trends-and-Reports/CMS-Statistics-Reference-Booklet/2014.html](http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/CMS-Statistics-Reference-Booklet/2014.html).

The briefest statistical summary on annual CMS program and financial data is published in *CMS Fast Facts*. It includes summary information on total national health expenditures; Medicare enrollment, utilization, and expenditures; and the number of Medicaid recipients and payment by selected types of service. *CMS Fast Facts* can be found at: [www.cms.gov/fastfacts](http://www.cms.gov/fastfacts).

## Data Navigator

An important tool for finding CMS information and data is the agency's data navigator. The data navigator is an easy-to-use, menu-driven search tool that guides the user to CMS' information and data on the World Wide Web, including the agency's data housed on external websites such as the Henry J. Kaiser Family Foundation, the National Institute of Medicine, and the Health Indicators Warehouse. The navigator enables the user to organize data into categories, such as by CMS program, setting/type of care, topic, geography, and document type. It also contains a comprehensive glossary of terms, a list of frequently asked questions, and a place to subscribe for email updates. The CMS data navigator's address is: <https://dnav.cms.gov>.

## Interactive Dashboards

To make its information more accessible, CMS has developed several interactive dashboards. For example, the Medicare Geographic Variation Dashboard provides users with an easy-to-use, customizable tool to find, compare, and analyze state- and county-level variations in Medicare per capita costs. Data used in the dashboard are based on CMS claims data for Medicare beneficiaries enrolled in the fee-for-service programs during the 5-year period 2008–2012. Users of the dashboard can compare state and county Medicare costs to that of the nation and identify year-to-year trends compared to national trends over the same time period. Specifically, users can compare

Medicare's total per capita costs, inpatient per capita costs, post-acute care per capita costs, hospice per capita costs, physician/outpatient department per capita costs, durable medical equipment per capita costs, Medicare Part B drug per capita costs, outpatient dialysis facility per capita costs, and the total number of Medicare beneficiaries in the state or county. The dashboard can be found at: [www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Geographic-Variation/GV\\_Dashboard.html](http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Geographic-Variation/GV_Dashboard.html).

Another example is the Medicare Chronic Condition Dashboard, which presents information for 2012 on the prevalence, utilization, and Medicare spending for 17 chronic disease conditions. The conditions include Alzheimer's disease/dementia, arthritis, asthma, atrial fibrillation, autism spectrum disorders, cancer, chronic kidney disease, chronic obstructive pulmonary disease (COPD), depression, diabetes, heart failure, hyperlipidemia, hypertension, ischemic heart disease, osteoporosis, schizophrenia/psychoses, and stroke. The information is presented by geographic areas such as federal government region, state, county, and hospital referral region. Users of the dashboard can select specific categories by gender, age group, Medicare beneficiaries only, and for dual eligible beneficiaries (individual receiving both Medicare and Medicaid). The dashboard is located at: [www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Chronic-Conditions/CCDashboard.html](http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Chronic-Conditions/CCDashboard.html).

## Data Products

CMS produces many data products that are available to researchers as well as the general public. These data products include many Medicare and Medicaid public use data files, the Chronic Conditions Data Warehouse (CCW), the Medicare Current Beneficiary Survey (MCBS), and the Medicare Data Sharing Program.

## Medicare and Medicaid Public Use Data File

Many of CMS' Medicare and Medicaid data files may be very useful to health services researchers.

Some of these files because they contain specific patient and condition identifiable data are restricted and difficult to obtain; however, other de-identified files are readily available as public use data files, which are free and can be easily downloaded.

Table 1 presents a list of 23 CMS public use data files and systems and the years for which they are available. The files are divided into nine broad

**Table 1** List of CMS' public use data files and the years for which they are available

<b>Healthcare organization cost data files</b>
<i>1. Healthcare Cost Report Information System (HCRIS)</i>
Community Mental Health Centers, 2010–2015
Health Clinics, 2009–2015
Home Health Agencies, 1994–2014
Hospices, 1999–2015
Hospitals, 1996–2015
Renal Dialysis Facilities, 1994–2015
Skilled Nursing Facilities, 1996–2014
<b>Medicare claims data files</b>
<i>2. Basic Stand Alone (BSA) Medicare Claims Public Use Files (PUFs)</i>
Carrier Line Items PUF, 2008, 2010
Durable Medical Equipment (DME) Line Items PUF, 2008, 2010
Home Health Agency (HHA) Beneficiary PUF, 2008, 2010
Hospice Beneficiary PUF, 2008, 2010
Inpatient Claims PUF, 2008
Outpatient Procedures PUF, 2008, 2010
Prescription Drug Events PUF, 2008
Skilled Nursing Facility (SNF) Beneficiary PUF, 2008, 2010
Chronic Conditions PUF, 2008, 2010
Institutional Providers and Beneficiary Summary PUF, 2013
Prescription Drug Profiles PUF, 2008, 2010
<i>3. Data Entrepreneurs' Synthetic Public Use Files (DE-SynPUF), 2008–2010</i>
Beneficiary Summary
Carrier Claims
Inpatient Claims
Outpatient Claims
<b>Physician and supplier Medicare charges</b>
<i>4. Medicare Provider Utilization and Payment Data</i>
Medicare Physician and Other Suppliers, 2012
Medicare Provider Utilization and Payment Data: Inpatient, 2011–2012

(continued)



**Table 1** (continued)

Medicare Provider Utilization and Payment Data: Outpatient, 2011–2012
<b>Program evaluation and health outcomes</b>
5. <i>Consumer Assessment of Healthcare Providers and Systems (CAHPS)</i> , Varies
Hospital CAHPS
Home Health CAHPS
Fee-for-Service CAHPS
Medicare Advantage and Prescription Drug Plan CAHPS
In-Center Hemodialysis CAHPS
Hospice
6. <i>Healthcare Effectiveness Data and Information Set (HEDIS)</i> , 1997–2015
7. <i>Medicare Compare</i>
Dialysis Facility Compare, 2010–2014
Home Health Compare, 2003–2014
Hospital Compare, 2005–2014
Nursing Home Compare, 2002–2014
Physician Compare, 2010–2014
8. <i>Medicare Health Outcome Survey (HOS)</i> , Varies by Cohort Group 1998–2015
Base Line PUFs
Follow-Up PUFs
Analytic PUFs
<b>Medicare prescription drug program</b>
9. <i>Prescription Drug Plan Formulary and Pharmacy Network Files</i> , 2005 – Current
Beneficiary Cost File
Formulary File
Geographic Locator File
Pharmacy Network File
Plan Information File
Pricing File
Record Layout
<b>Medicare electronic medical records program files</b>
10. Medicare Electronic Health Record (ERH) Incentive Program Eligible
Professionals Public Use File (PUF), 2013
Eligible Professionals PUF
Eligible Hospitals PUF
<b>Medicaid data files</b>
11. <i>Medicaid Analytic Extract (MAX) Provider Characteristics File</i> , 2009–2011
12. <i>Medicaid/CHIP Environmental Scanning and Program Characteristics (ESPC) File</i> , 2005–2013
13. <i>Medicaid State Drug Utilization File</i> , 1991–2014
14. <i>Medicaid Statistical Information System (MSIS) Datamart</i>
MSIS State Summary Datamart, 1999–2012
MSIS Drug Utilization Datamart, 2004–2010

(continued)

**Table 1** (continued)

<b>Geographic regions and hospital service areas</b>
15. <i>Hospital Service Area File</i> , 1992–2013
16. <i>Medicare Geographic Variation Files</i> , 2007–2013
Hospital Referral Region (HRR) Report – All Beneficiaries
Hospital Referral Region (HRR) Report – Beneficiaries Under 65
Hospital Referral Region (HRR) Report – Beneficiaries 65 and Older
Hospital Referral Region (HRR) Table – All Beneficiaries
Hospital Referral Region (HRR) Table – Beneficiaries Under 65
Hospital Referral Region (HRR) Table – Beneficiaries 65 and Older
State/County Report – All Beneficiaries
State Report – Beneficiaries Under 65
State Report – Beneficiaries 65 and Older
State/County Table – All Beneficiaries
State Table – Beneficiaries Under 65
State Table – Beneficiaries 65 and Older
<b>Directories of providers and coding systems</b>
17. <i>Health Care Information System (HCIS) Data File</i> , 2009–2011
18. <i>Medicare Part B Summary Data Files</i>
Carrier File, 2005–2011
National File, 2000–2013
19. <i>National Provider Identifier (NPI) Downloadable File</i> , 2007 – Current
20. <i>Physician Supplier Procedure Summary Master File</i> , 1991–2013
21. <i>Provider of Services (POS) File</i> , 1991–2014
22. <i>Unique Physician Identification Number (UPIN) Directory</i> , 2003–2007
23. <i>Unique Physician Identification Number (UPIN) Group File</i> , 2005–2007

categories: healthcare organization cost data files, Medicare claims data files, physician and supplier Medicare charges, program evaluation and health outcomes, Medicare prescription drug program, Medicare electronic medical records program files, Medicaid data files, geographic regions and hospital service areas, and directories of providers and coding systems. The categories are discussed below, and individual files are highlighted.

**Healthcare Organization Cost Data Files**

Some of the most widely used CMS public use files are those containing Medicare cost reports.

Specifically, these reports are included in the Healthcare Cost Report Information System (HCRIS). The various files in HCRIS contain annual mandatory cost reports submitted to CMS from all healthcare facilities that accept Medicare funds. Nearly all of the nation's hospitals, skilled nursing homes, hospices, renal dialysis facilities, independent rural health clinics, and freestanding federally qualified health centers submit these reports. The cost reports consist of a series of forms that collect descriptive, financial, and statistical data to determine if the Medicare program over or underpaid the facility. These files are frequently used by health services researchers to examine various facility characteristics, calculate costs and charges, and determine the financial viability of the facility (Asper 2013; Holmes et al. 2013; Kane and Magnus 2001). More information on the various files can be found at: [www.resdac.org/cms-data/files/hcris](http://www.resdac.org/cms-data/files/hcris).

### Medicare Claims Data Files

Another widely used data source is the Medicare Claims Data Files. These files are part of the Basic Stand Alone (BSA) Medicare Claims Public Use Files (PUFs). It consists of 11 separate basic standalone public use files. Most of these files contain non-identifiable claims-specific data derived from a 5 % sample of all Medicare beneficiaries. The files are often used by health services researchers, and they are increasingly being used to conduct public health surveillance (Erdem and Concannon 2012; Stein et al. 2014; Erdem et al. 2014). Additional information on the files and how health services researchers use them can be found at: [www.academyhealth.org/Training/ResourceDetail.cfm?ItemNumber=7097](http://www.academyhealth.org/Training/ResourceDetail.cfm?ItemNumber=7097).

To encourage researchers to use the Medicare claims files, CMS has constructed the Data Entrepreneurs' Synthetic Public Use Files (DE-SynPUF). The DE-SynPUF allows researchers to develop and create software applications for Medicare claims data, train individuals to analyze claims data using the actual files, and support safe data mining innovations. Data contained in the DE-SynPUF is based on a 5 % sample of Medicare beneficiaries including beneficiary summary data, inpatient, outpatient, carrier, and prescription drug event claims. More

information can be found at: [www.resdac.org/event/webinar-introduction-data-entrepreneurs-synthetic-public-use-file-de-synpuf](http://www.resdac.org/event/webinar-introduction-data-entrepreneurs-synthetic-public-use-file-de-synpuf).

### Physician and Supplier Medicare Charges

The next category includes the Medicare Provider Utilization and Payment Data files. These files contain data on the services and procedures provided to Medicare beneficiaries by physicians and other healthcare professionals on an inpatient and outpatient basis. They also include all final-action physician/supplier Part B noninstitutional line items for the Medicare fee-for-service population. For more information on these files, go to [www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data](http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data).

### Program Evaluation and Health Outcomes

CMS offers researchers many program evaluation and health outcome public use data files. One such set of files is contained in the Consumer Assessment of Healthcare Providers and Systems (CAHPS). CAHPS consists of a family of various patient experience surveys. These surveys ask patients, or in some cases family members, about their experiences with, and ratings of, the care they received. The surveys in many cases are the only source of information on the care they received. CAHPS surveys have been developed for hospitals, home health, Medicare fee-for-service care, Medicare Advantage and Prescription Drug plans, in-center hemodialysis, and hospices. Results from the surveys are contained in various public use files. Copies of the CAHPS survey instruments can be found at: [www.cms.gov/Research-Statistics-Data-and-Systems/Research/CAHPS/index.html](http://www.cms.gov/Research-Statistics-Data-and-Systems/Research/CAHPS/index.html). And more information on the CAHPS public use data files can be found at: [www.resdac.org/cms-data/files/cahps-puf](http://www.resdac.org/cms-data/files/cahps-puf).

A number of other CMS public use files are also derived from CAHPS. Data from various CAHPS surveys are used to produce Medicare Compare files and related websites, which contain data on individual facilities and physicians. These files provide contact information, quality of care measures, lists of services offered, and a five-star rating system.

The Medicare Compare files are available for kidney dialysis facilities ([www.medicare.gov/dialysisfacilitycompare/](http://www.medicare.gov/dialysisfacilitycompare/)), home health care agencies ([www.medicare.gov/homehealthcompare/](http://www.medicare.gov/homehealthcompare/)), hospitals ([www.medicare.gov/hospitalcompare/search.html](http://www.medicare.gov/hospitalcompare/search.html)), skilled nursing facilities ([www.medicare.gov/nursinghomecompare/search.html](http://www.medicare.gov/nursinghomecompare/search.html)), and physicians ([www.medicare.gov/physicianscompare/search.html](http://www.medicare.gov/physicianscompare/search.html)). Many health services researchers have used these files to measure the quality of care provided at various healthcare facilities (Werner and Bradow 2006; Saunders and Chin 2013; Lutfiyya et al. 2013; Williams et al. 2014). More information on the public use files can be found at [www.resdac.org/cms-data/files/medicare-compare](http://www.resdac.org/cms-data/files/medicare-compare).

Another public use file dealing with quality of healthcare is the Healthcare Effectiveness Data and Information Set (HEDIS) public use file. CMS uses HEDIS to compare health plans providing Medicare and Medicaid services. HEDIS, which was developed by the independent not-for-profit National Committee for Quality Assurance (NCQA), is a widely used tool to measure the performance of health plans. It currently consists of 81 measures across five domains of care and service. HEDIS, which is used by more than 90 % of America's health plans, enables researchers to compare the performance of the plans. HEDIS has been used to compare different quality measures of care (Pugh et al. 2013; Bundy et al. 2012). Information on HEDIS and its performance measures can be found at: [www.ncqa.org/HEDISQualityMeasurement.aspx](http://www.ncqa.org/HEDISQualityMeasurement.aspx). And information on the public use file is available at: [www.resdac.org/cms-data/files/hedis-puf](http://www.resdac.org/cms-data/files/hedis-puf).

Lastly, the Medicare Health Outcome Survey (HOS) public use files provide a rich source of outcome data on Medicare beneficiaries enrolled in Medicare Advantage programs. The Medicare HOS consists of Base Line, Follow-Up, and Analytic Public Use Files. The survey, which measures quality improvement activities, health plan performance, and outcomes of care, is administered to cohorts of individuals who are repeatedly sampled over time. Results from the Medicare HOS have been used by health services researchers and quality improvement professionals to explore functional status measurement issues and identify

ways to improve healthcare practices (Haffer and Bowen 2004; Bowen 2012). More information can be found at [www.resdac.org/cms-data/file-family/Health-Outcomes-Survey-HOS](http://www.resdac.org/cms-data/file-family/Health-Outcomes-Survey-HOS).

### **Medicare Prescription Drug Program**

The next category includes the Prescription Drug Plan Formulary and Pharmacy Network Files. It consists of seven separate files: Beneficiary Cost File, Formulary File, Geographic Locator File, Pharmacy Network File, Plan Information File, Pricing File, and Record Layout. These files contain data on Medicare prescription drug plans and Medicare Advantage prescription drug plans. The various files are updated weekly, monthly, and quarterly. For more information see: [www.resdac.org/cms-data/files/pharmacy-network](http://www.resdac.org/cms-data/files/pharmacy-network).

### **Medicare Electronic Medical Records Program Files**

CMS encourages the greater use of electronic medical records by all healthcare providers. It has established an incentive program that provides payments to hospitals and healthcare professionals to adopt, implement, upgrade, or demonstrate the use of electronic health record technology. As of February 2015, more than 438,000 healthcare providers received funds for participating in the program. To identify eligible hospitals and professionals, CMS has constructed the Medicare Electronic Health Record (ERH) Incentive Program Eligible Professional Public Use File (Wright et al. 2014). More information on the program and the files can be obtained at: [www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/DataAndReports.html](http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/DataAndReports.html).

### **Medicaid Data Files**

The next category identifies four CMS Medicaid public use files. The Medicaid Analytic Extract (MAX) Provider Characteristics File contains data on state Medicaid programs including the number of individuals enrolled, demographic characteristics (age, gender, ethnicity, and race), basis of eligibility (aged, disabled, children, and adults), and maintenance assistant status (medically needy, poverty, waiver, and other). However,

after several years of data collection, the files were discontinued. They were last updated in 2011. The MAX files have been used by researchers to study medical adherence to drugs (Rust et al. 2013) and the maternal and infant outcomes of multistate Medicaid populations (Palmsten et al. 2014). A chartbook summarizing 2010 MAX data is also available (Borck et al. 2014). For more information about the public use files, see [www.resdac.org/cms-data/files/max-provider-characteristics](http://www.resdac.org/cms-data/files/max-provider-characteristics).

The second public use file is the Medicaid/CHIP Environmental Scanning and Program Characteristics (ESPC) File. This file was created by CMS to encourage cross-state analysis of Medicaid programs. It is now part of CMS' Environmental Scanning and Program Characteristics (ESPC) Database. The ESPC can be linked to the Medicaid Analytic Extract (MAX) files and other Medicaid data. More information can be found at: [www.resdac.org/cms-data/files/medicaidchip-esp](http://www.resdac.org/cms-data/files/medicaidchip-esp).

Another public use file is the Medicaid State Drug Utilization File. This file contains data for covered outpatient drugs paid for by state Medicaid agencies since the start of the federal Drug Rebate Program in 1990. Currently, all states and the District of Columbia participate in the program, as well as about 600 drug manufacturers. For more information see: [www.resdac.org/cms-data/files/medicaid-state-drug-utilization](http://www.resdac.org/cms-data/files/medicaid-state-drug-utilization).

Lastly, the Medicaid Statistical Information System (MSIS) Datamart contains two public use data files: State Summary Datamart and the Drug Utilization Datamart. Both of these files can be used to produce tables covering a wide range of Medicaid program statistics on eligibility and claims data. These files contain data on Medicaid eligible, beneficiaries, and payment, maintenance assistance status, age group, gender, race/ethnicity, and service category and program type. For more information go to: [www.resdac.org/cms-data/files/msis-datamart](http://www.resdac.org/cms-data/files/msis-datamart).

### **Geographic Regions and Hospital Service Areas**

The next category includes two geographic public use files. The first file is the Hospital Service Area File. It contains summary data on hospital

discharges, length of stay, and total charges by CMS provider numbers and zip codes of the Medicare beneficiaries. Using these data hospital service areas can be determined for various services. More information on the file can be found at: [www.resdac.org/cms-data/files/hsaf](http://www.resdac.org/cms-data/files/hsaf).

The largest set of CMS geographic public use files is the Medicare Geographic Variation Files. They include 12 separate files – two files with state- and county-level data, four files with state-level data, and six files with hospital referral regions (HRRs). The files are divided into report and table formats for all Medicare beneficiaries, those under 65 years of age and those 65 years of age and older. These geographic files contain demographic, spending, utilization, and quality of care indicators for the Medicare fee-for-service population at the state, county, and hospital referral regions. The hospital referral regions were developed by the Dartmouth Atlas of Health Care Project and have been widely used by health services researchers to investigate regional differences in access, cost, quality, and the outcomes of care (Baker et al. 2014; Chen et al. 2014; Wennberg 2010). Detailed information on the files can be found at: [www.resdac.org/cms-data/files/medicare-geographic-variation](http://www.resdac.org/cms-data/files/medicare-geographic-variation).

### **Directories of Providers and Coding Systems**

The last category includes seven directories of providers and medical procedure coding systems public use data files. These files contain a listing of the unique CMS healthcare facility and healthcare professional provider identifiers and lists of CMS recognized medical procedure codes. The lists and procedure codes are primarily used for billing and payment purposes.

The public use Health Care Information System (HCIS) Data File contains information on each Medicare Part A and B institutional provider by type of facility and state. Specifically, it lists CMS provider identifiers, facility characteristics, total payment amounts, total number of Medicare beneficiaries served, and total utilization for hospitals, skilled nursing facilities, home health agencies, and hospices. For more information see: [www.resdac.org/cms-data/files/hcis](http://www.resdac.org/cms-data/files/hcis).

The Medicare Part B Summary Data Files consists of two separate public use files: Carrier File and National File. These files contain data summaries by Healthcare Common Procedure Coding System (HCPCS) code ranges. The HCPCS are medical codes used to report supplies, equipment, and devices provided to patients. The file includes allowed services, allowed charges, and payment amounts. More information on the files can be found at: [www.resdac.org/cms-data/files/part-b-summary-data](http://www.resdac.org/cms-data/files/part-b-summary-data).

The next public use file is the National Provider Identifier (NPI) Downloadable File. The NPI is a unique, ten-digit, identification number for each CMS-covered healthcare provider. By federal law, the NPI must be used in all administrative and financial healthcare transactions. The file contains NPI data on the name, gender, business address, and medical license number of provider. For more information see: [www.resdac.org/cms-data/files/nppes](http://www.resdac.org/cms-data/files/nppes).

The Physician Supplier Procedure Summary Master File contains data on all Medicare Part B carrier and durable medical equipment regional carrier (DMERC) claims that were processed by CMS. Carriers are private companies that have contracts with Medicare to process Part B claims. Durable medical equipment (DME) is equipment that can withstand repeated use and is appropriate for home use, for example wheelchairs, oxygen equipment, and hospital beds. The file includes data on each carrier; pricing locality; HCPCS procedure code; type and place of service; submitted, allowed, and denied services and charges; and payment amounts. More information can be found at: [www.resdac.org/cms-data/files/psps](http://www.resdac.org/cms-data/files/psps).

The Provider of Services (POS) File contains a record of each Medicare provider, including all institutional providers, ambulatory surgical centers, and clinical laboratories. The file, which is updated quarterly, includes CMS provider identification numbers and the characteristics of hospitals and other types of facilities, including the name, address, and type of Medicare services the facility provided. For further information see: [www.resdac.org/cms-data/files/pos](http://www.resdac.org/cms-data/files/pos).

The last two files in this category have been discontinued and replaced by the National

Provider Identifier (NPI) Downloadable File, which was previously discussed. These two files, which may be of interest to researchers investigating physicians in the mid-2000s, include the Unique Physician Identification Number (UPIN) Directory and the Unique Physician Identification Number (UPIN) Group File. The first file contains the name, specialty, license number, and zip code of physicians, limited licensed practitioners, and some nonphysician practitioners who were enrolled in the Medicare program. The second file provides data on group practices and the physicians who were members of them. Both files were discontinued in 2007 with the implementation of the NPI. Information on the two files can be obtained at: [www.resdac.org/cms-data/files/upin-directory](http://www.resdac.org/cms-data/files/upin-directory) and [www.resdac.org/cms-data/files/upin-group](http://www.resdac.org/cms-data/files/upin-group).

## Chronic Conditions Data Warehouse

Another important CMS data product is the Chronic Conditions Data Warehouse (CCW). Established in 2006, the CCW is a national Medicare and Medicaid research database containing claims and assessment data linked by beneficiary across the continuum of care. It also includes Medicare Part D prescription drug event data listing plan, pharmacy, prescriber characteristics, and a formulary file.

The CCW is designed to promote the use of current Medicare and Medicaid analytic easy-to-use data files by researchers and policy analysts, promote longitudinal research using data already linked by beneficiary across the continuum of care, identify areas to improve the quality of care provided to chronically ill beneficiaries, identify possible ways to reduce program spending, and provide thorough documentation so these data may be used accurately (General Dynamics Information Technology 2013; CCW website, [www.ccwdata.org/web/guest/about-ccw](http://www.ccwdata.org/web/guest/about-ccw)).

The CCW uses various computer algorithms to identify various conditions. The database includes 27 chronic disease conditions, 9 mental health and tobacco use conditions, and 15 conditions that are related to physical and intellectual disability and developmental disorders.

Specifically, the CCW's chronic disease conditions include acquired hypothyroidism, acute myocardial infarction, Alzheimer's disease, Alzheimer's or related dementia, anemia, asthma, atrial fibrillation, benign prostatic hyperplasia, cataract, chronic kidney disease (CKD), chronic obstructive pulmonary disease (COPD), heart failure, depression, diabetes, glaucoma, hip/pelvic fracture, hyperlipidemia, hypertension, ischemic heart disease, osteoporosis, rheumatoid/osteoarthritis, stroke/transient ischemic attack (TIA), breast cancer, colorectal cancer, lung cancer, prostate cancer, and endometrial cancer.

The CCW's mental health and tobacco conditions include conduct disorders and hyperkinetic syndrome, anxiety disorders, bipolar disorder, depressive disorders, personality disorders, post-traumatic stress disorder (PTSD), schizophrenia, schizophrenia and other psychotic disorders, and tobacco use disorder.

Lastly, the CCW's physical and mental disability conditions include autism spectrum disorder; cerebral palsy; cystic fibrosis and other metabolic developmental disorders; epilepsy; intellectual disabilities and related conditions; learning disabilities and other developmental delays; mobility impairments; multiple sclerosis and transverse myelitis; muscular dystrophy; sensory – deafness and hearing impairment; sensory – blindness and visual impairment; spina bifida and other congenital anomalies of the nervous system; spinal cord injury; traumatic brain injury and nonpsychotic mental disorders due to brain damage; and other developmental delays.

General information on the CCW can be obtained at [www.ccwdata.org/web/guest/home](http://www.ccwdata.org/web/guest/home). And a current detailed user guide (Buccaneer Computer Systems and Service 2015) can be found at: [www.ccwdata.org](http://www.ccwdata.org).

## Medicare Current Beneficiary Survey

A very widely used CMS data product is the Medicare Current Beneficiary Survey (MCBS). Since the survey's inception in 1991, the MCBS data files have been used to estimate the health status, healthcare use and expenditures, health

insurance coverage, satisfaction with the care they received, and socioeconomic and demographic characteristics of Medicare beneficiaries. It also has been used to study the occurrence and treatment of specific chronic conditions of the elderly such as depression, dementia, hip fractures, glaucoma, osteoporosis, and rheumatoid arthritis. A bibliography and copies of over 800 research articles published from 1992 to 2013, which used MCBS data, can be found at [www.cms.gov/Research-Statistics-Data-and-Systems/Research/MCBS/Bibliography.html](http://www.cms.gov/Research-Statistics-Data-and-Systems/Research/MCBS/Bibliography.html).

The MCBS is a continuous, in-person, longitudinal panel survey of a representative national sample of the Medicare population. Survey respondents are interviewed three times a year over a period of 4 years to form a continuous profile of their healthcare experience. Two types of interviews are conducted: a community interview done at the respondent's residence and a healthcare institutional interview of knowledgeable staff on behalf of the beneficiary. An important feature of the MCBS is that respondents are followed into and out of long-term care facilities during their panel participation. About 16,000 Medicare beneficiaries are interviewed every year (Adler 1994; Briesacher et al. 2012).

Two data products are derived each year from the MCBS: the Access to Care data file and the Cost and Use data file. The Access to Care file represents all persons enrolled in Medicare throughout the entire data collection year, which is referred to as the "always enrolled" beneficiary population. The file contains data on the beneficiaries' access to healthcare, satisfaction with care, and usual source of care. The Access to Care file is released within a year of the survey (Petroski et al. 2014).

The Cost and Use file represents all persons enrolled in Medicare at any point during the data collection year, which is referred to as the "ever-enrolled" beneficiary population. The file links Medicare claims data to survey-reported events and provides complete expenditure and source of payment data on all healthcare services, including those not covered by Medicare. The file contains data on the beneficiaries' use and cost of healthcare services, information supplementary

health insurance, living arrangements, income, health status, and physical functioning. The Cost and Use file is released within 2 years of the survey.

More information on the MCBS and its two files can be obtained at: [www.cms.gov/Research-Statistics-Data-and-Systems/Research/MCBS/index.html?redirect=/MCBS](http://www.cms.gov/Research-Statistics-Data-and-Systems/Research/MCBS/index.html?redirect=/MCBS). Additionally, an informative free webinar presentation, “Getting and Using the Medicare Current Beneficiary Survey (MCBS) for Health Services Research: Guidance from the Experts,” is available from Academy Health at: [www.academyhealth.org/Training/ResourceDetail.cfm?ItemNumber=11031](http://www.academyhealth.org/Training/ResourceDetail.cfm?ItemNumber=11031).

## Medicare Qualified Entity Program

The last data product to be discussed is the CMS’ Medicare Qualified Entity Program. This program, which was mandated by the Affordable Care Act of 2010, requires CMS to provide access to Medicare claims data by qualified entities (QEs) in order to produce public performance reports on physicians, hospitals, and other healthcare providers. The program enables the QEs to combine Medicare claims data with commercial insurance and Medicaid claims data. To become a QE, an organization must demonstrate existing expertise in performance measurement, the ability to combine Medicare data with other claims data, a process for allowing providers to review and correct their performance reports, and adherence to data privacy and security procedures (Hostetter and Klein 2013).

As of June 2014, CMS has certified 12 regional and one national QE: Oregon Health Care Quality Corporation (Q-Corp), Health Improvement Collaborative of Greater Cincinnati, Kansas City Quality Improvement Consortium, Maine Health Management Coalition Foundation, Health Insight (covering five counties in New Mexico), California Healthcare Performance Information System, Pittsburgh Regional Health Initiative, Minnesota Community Measurement, Wisconsin Health Information Organization, Center for Improving Value in Health Care (covering Colorado), Minnesota Department of Health, Division

of Health Policy, Midwest Health Initiative (covering the St. Louis area and 16 counties in Missouri), and the Health Care Cost Institute (covering all 50 states and the District of Columbia).

The QEs are beginning to release public reports using the combined Medicare and other payer data. The first report was published by the Oregon Health Care Quality Corporation, *Information for a Healthy Oregon: Statewide Report on Health Care Quality 2014* ([www.qcorp.org/reports/statewide-reports](http://www.qcorp.org/reports/statewide-reports)). It includes information on Oregon’s chronic disease care, preventive services, and ambulatory and hospital resource use.

More information on CMS’ Qualified Entity Program is available at: [www.resdac.org/cms-data/request/qualified-entity-program](http://www.resdac.org/cms-data/request/qualified-entity-program); [www.cms.gov/QEMedicareData](http://www.cms.gov/QEMedicareData); and [www.QEMedicareData.org](http://www.QEMedicareData.org).

---

## Conclusion

In the future, CMS will increasingly release more information and data products that will be useful to health services researchers, policymakers, educators, students, and the general public. CMS will continue to collect data on the Medicare, Medicaid, and Children’s Health Insurance Program (CHIP). At the same time, CMS will also expand its data collection efforts to measure its many new initiative programs, which are attempting to improve the quality of patient care, provide a greater emphasis on prevention and population health, and expand healthcare coverage. These initiatives will encourage all of the nation’s healthcare providers to use electronic health records, establish more Accountable Care Organizations (ACOs), increase value-based purchasing, better coordinate care for dual eligible beneficiaries, and reduce unnecessary hospital readmissions. As CMS moves from being a volume payer of healthcare services to a value-based payer, it will need much more data to identify the best ways to increase the quality of care while at the same time lower its costs (Burwell 2015; CMS Strategy 2013).

## References

- Adler GS. A profile of the medicare current beneficiary survey. *Health Care Financ Rev.* 1994;15(4):153–63. Available at: [www.cms.gov/Research-Statistics-Data-and-Systems/Research/HealthCareFinancingReview/Downloads/CMS1191330dl.pdf](http://www.cms.gov/Research-Statistics-Data-and-Systems/Research/HealthCareFinancingReview/Downloads/CMS1191330dl.pdf)
- Asper F. Introduction to Medicare cost reports. Slide presentation. Minneapolis: Research Data Assistance Center; 2013. Available at: [www.resdac.org/sites/resdac.org/files/IntroductiontoMedicareCostReports\(Slides\).pdf](http://www.resdac.org/sites/resdac.org/files/IntroductiontoMedicareCostReports(Slides).pdf)
- Baker LC, Kate Bundorf M, Kessler DP. Patients' preferences explain a small but significant share of regional variation in Medicare spending. *Health Aff.* 2014;33(6):957–63.
- Blumenthal D, Davis K, Guterman S. Medicare at 50 – moving forward. *N Engl J Med.* 2015;372(7):671–7. Available at: [www.nejm.org/doi/full/10.1056/NEJMp1414856](http://www.nejm.org/doi/full/10.1056/NEJMp1414856)
- Borck R, Laura R, Vivian B, Wagnerman K. The Medicaid analytic extract 2010 chartbook. Baltimore: Centers for Medicare and Medicaid Services; 2014. Available at: [www.mathematica-mpr.com/~media/publications/pdfs/health/maxchartbook\\_2010.pdf](http://www.mathematica-mpr.com/~media/publications/pdfs/health/maxchartbook_2010.pdf)
- Bowen SE. Evaluating outcomes of care and targeting quality improvement using Medicare Health Outcomes Survey data. *J Ambul Care Manage.* 2012;35(4):260–2.
- Brennan N, Oelschlaeger A, Cox C, Tavenner M. Leveraging the big-data revolution: CMS is expanding capabilities to spur health system transformation. *Health Aff.* 2014;33(7):1195–202.
- Briesacher BA, Tjia J, Doubeni CA, Chen Y, Rao SR. Methodological issues in using multiple years of the medicare current beneficiary survey. *Medicare Medicaid Res Rev.* 2012;2(1):E1–19. Available at: [www.cms.gov/mmrr/downloads/mmrr2012\\_002\\_01\\_a04.pdf](http://www.cms.gov/mmrr/downloads/mmrr2012_002_01_a04.pdf)
- Buccaneer Computer Systems and Service. Chronic conditions data warehouse: Medicare administrative data user guide. Version 3.1. 2015. Available at: [www.ccwdata.org](http://www.ccwdata.org)
- Bundy DG, Solomon BS, Kim JM, et al. Accuracy and usefulness of the HEDIS childhood immunization measures. *Pediatrics.* 2012;129(4):648–56. Available at: [www.ncbi.nlm.nih.gov/pmc/articles/PMC3313643/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3313643/)
- Burwell SM. Setting value-based payment goals – HHS efforts to improve U.S. health care. *N Engl J Med.* 2015;372(10):897–9. Available at: [www.nejm.org/doi/full/10.1056/NEJMp1500445](http://www.nejm.org/doi/full/10.1056/NEJMp1500445)
- Centers for Medicare and Medicaid Services (CMS). CMS announces data and information initiative. Fact Sheet. 2012. Available at: [www.cms.gov/Research-Statistics-Data-and-Systems/Research/ResearchGeninfo/Downloads/OIPDA\\_Fact\\_Sheet.pdf](http://www.cms.gov/Research-Statistics-Data-and-Systems/Research/ResearchGeninfo/Downloads/OIPDA_Fact_Sheet.pdf)
- Centers for Medicare and Medicaid Services (CMS). Centers for Medicare and Medicaid services: justification of estimates for appropriations committees. Baltimore: Centers for Medicare and Medicaid Services; 2015, p. 2. Available at: [www.cms.gov/About-CMS/Agency-Information/PerformanceBudget/Downloads/FY2015-CJ-Final.pdf](http://www.cms.gov/About-CMS/Agency-Information/PerformanceBudget/Downloads/FY2015-CJ-Final.pdf)
- Centers for Medicare and Medicaid Services (CMS). CMS strategy: the road forward: 2013–2017. Baltimore: Centers for Medicare and Medicaid Services; 2013. Available at: [www.cms.gov/About-CMS/Agency-Information/CMS-Strategy/Downloads/CMS-Strategy.pdf](http://www.cms.gov/About-CMS/Agency-Information/CMS-Strategy/Downloads/CMS-Strategy.pdf)
- Centers for Medicare and Medicaid Services (CMS). Medicare and you, 2015. Baltimore: Centers for Medicare and Medicaid Services; 2015. Available at: [www.medicare.gov/Pubs/pdf/10050.pdf](http://www.medicare.gov/Pubs/pdf/10050.pdf)
- Chen C, Petterson S, Phillips R, et al. Spending patterns in region of residency training and subsequent expenditures for care provided by practicing physicians for Medicare beneficiaries. *JAMA.* 2014;312(22):2385–92.
- Chronic Condition Data Warehouse. About chronic condition data warehouse. Available at: [www.ccwdata.org/web/guest/about-ccw](http://www.ccwdata.org/web/guest/about-ccw)
- Cohen AB, Colby DC, Wailoo K, Zelizer J, editors. Medicare and Medicaid at 50: America's entitlement programs in the age of affordable care. New York: Oxford University Press; 2015.
- Erdem E, Concannon TW. What do researchers say about proposed Medicare claims public use files? *J Comp Eff Res.* 2012;1(6):519–25.
- Erdem E, Korda HH, Haffer SC, Sennett C. Medicare claims data as public use files: a new tool for public health surveillance. *J Public Health Manag Pract.* 2014;20(4):445–52.
- Ewing MT, editor. State Children's Health Insurance Program (CHIP). New York: Nova; 2008.
- Feder J, Komisar HL. The importance of federal financing to the nation's long-term care safety net. 2012. Available at: [www.thescanfoundation.org](http://www.thescanfoundation.org)
- General Dynamics Information Technology. Centers for Medicare and Medicaid Services Chronic Condition Data Warehouse (CCW): national Medicare and Medicaid research database. Fairfax: General Dynamics Information Technology; 2013. Available at: [www.gdit.com/globalassets/health/6978\\_ccw.pdf](http://www.gdit.com/globalassets/health/6978_ccw.pdf)
- Haffer SC, Bowen SE. Measuring and improving health outcomes in Medicare: the Medicare HOS program. *Health Care Financ Rev.* 2004;25(4):1–3. Available at: [www.ncbi.nlm.nih.gov/pmc/articles/PMC4194894/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4194894/)
- Henry J. Kaiser Family Foundation. Medicaid moving forward. *Kaiser Commission on Medicaid and the Uninsured, Fact Sheet.* 2015. Available at: [www.kff.org](http://www.kff.org)
- Henry J. Kaiser Family Foundation. The medicare Part D prescription drug benefit. Fact Sheet. 2014. Available at: [www.kff.org](http://www.kff.org)
- Henry J. Kaiser Family Foundation. State demonstration proposals to align financing and/or administration for dual eligible beneficiaries. February 2015. Fact Sheet. 2015. Available at: [www.kff.org](http://www.kff.org)
- Holmes GM, Pink GH, Friedman SA. The financial performance of rural hospitals and implications for elimination of the critical access hospital program. *J Rural Health.* 2013;29(2):140–9.
- Hostetter M, Klein S. Medicare data helps fill in picture of health care performance. *Quality Matters: The*



- Commonwealth Fund Newsletter. 2013. Available at: [www.commonwealthfund.org/publications/newsletters/quality-matters/2013/april-may/in-focus](http://www.commonwealthfund.org/publications/newsletters/quality-matters/2013/april-may/in-focus)
- Kane NM, Magnus SA. The Medicare cost report and the limits of hospital accountability: improving financial accounting data. *J Health Polit Policy Law*. 2001;26(1):81–106.
- Lutfiyya MN, Gessert CE, Lipsky MS. Nursing home quality: a comparative analysis using CMS nursing home compare data to examine differences between rural and non-rural facilities. *J Am Med Dir Assoc*. 2013;14(8):593–8.
- National Conference of State Legislatures. Children's health: trends and options for covering kids. Washington, DC: National Conference of State Legislatures; 2014. Available at: [www.ncsl.org/documents/health/coveringkids914.pdf](http://www.ncsl.org/documents/health/coveringkids914.pdf)
- Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. The Medicare advantage program in 2014. ASPE Issue Brief. 2014. Available at: <http://aspe.hhs.gov>
- Orentlicher D. Medicaid at 50: no longer limited to the 'Deserving' poor? *Yale J Health Policy Law Ethics*. 2015;15(1):185–95.
- Palmsten K, Huybrechts KF, Kowal MK, et al. Validity of maternal and infant outcomes within nationwide Medicaid data. *Pharmacoepidemiol Drug Saf*. 2014;23(6):646–55.
- Petroski J, Ferraro D, Chu A. Ever enrolled Medicare population estimates from the MCBS access to care files. *Medicare Medicaid Res Rev*. 2014;4(2):E1–16. Available at: [www.cms.gov/mmrr/Downloads/MMRR2014\\_004\\_02\\_a05.pdf](http://www.cms.gov/mmrr/Downloads/MMRR2014_004_02_a05.pdf)
- Pugh MJV, Marcum ZA, Copeland LA, et al. The quality of quality measures: HEDIS quality measures for medication management in the elderly and outcomes associated with new exposure. *Drugs Aging*. 2013;30(8):645–54. Available at: [www.ncbi.nlm.nih.gov/pmc/articles/PMC3720786/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3720786/)
- Rust G, Zhang S, Reynolds J. Inhaled corticosteroid adherence and emergency department utilization among Medicaid-enrolled children with asthma. *J Asthma*. 2013;50(7):769–75. Available at: [www.ncbi.nlm.nih.gov/pmc/articles/PMC4017346/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4017346/)
- Saunders MR, Chin MH. Variation in dialysis quality measures by facility, neighborhood, and region. *Med Care*. 2013;51(5):413–7. Available at: [www.ncbi.nlm.nih.gov/pmc/articles/PMC3651911/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3651911/)
- Stein BD, Pangilnan M, Sorbero MJ, et al. Using claims data to generate clinical flags predicting short-term risk of continued psychiatric hospitalizations. *Psychiatr Serv*. 2014;65(11):1341–6.
- U.S. Government Accountability Office. Health care transparency: actions needed to improve cost and quality information for consumers. Washington, DC: U.S. Government Accountability Office; 2014. Available at: [www.gao.gov/products/GAO-15-11](http://www.gao.gov/products/GAO-15-11)
- Wennberg JE. Tracking medicine: a researcher's quest to understand health care. New York: Oxford University Press; 2010.
- Werner RM, Bradow ET. Relationship between Medicare's hospital compare performance measures and mortality rates. *JAMA*. 2006;296(22):2694–702.
- Williams A, Straker JK, Applebaum R. The nursing home five star rating: how does it compare to resident and family views of care? *Gerontologist*. 2014.
- Wright A, Febowitz J, Samal L, et al. The Medicare electronic health record incentive program: provider performance on core and menu measures. *Health Serv Res*. 2014;49(1 Pt 2):325–46. Available at: [www.ncbi.nlm.nih.gov/pmc/articles/PMC3925405/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3925405/)



# Health Services Data: Typology of Health Care Data

# 5

Ross M. Mullner

## Contents

<b>Introduction</b> .....	78
<b>Basic Units of Analysis</b> .....	80
Individuals .....	80
Households .....	81
Groups/Populations .....	81
Health Care Organizations .....	81
Health Care Programs .....	82
National Health Care Systems .....	83
<b>Collection Methods</b> .....	84
Literature Reviews .....	85
Observations .....	86
Focus Groups .....	86
Surveys .....	87
Medical Records, Administrative, and Billing Sources .....	88
Registries .....	88
Vital Records .....	89
<b>Data Sources and Holdings</b> .....	90
Government Organizations .....	90
Private Organizations .....	99
<b>Conclusion</b> .....	105
<b>References</b> .....	105

### Abstract

Health services researchers study access, cost, quality, and the outcome of health care. These researchers frequently use existing data collected by government agencies and private

organizations to monitor and evaluate current health care programs and systems and to predict the consequences of proposed new health policies. Primarily focusing on US data sources, this chapter outlines a practical typology, or classification framework, of health care data that is often used by these researchers when they are gathering data and conducting their studies. The typology addresses three important inextricably linked questions. First, what is the basic unit of

R. M. Mullner (✉)  
 Division of Health Policy and Administration, School of Public Health, University of Illinois, Chicago, IL, USA  
 e-mail: [rmullner@comcast.net](mailto:rmullner@comcast.net)

analysis for the study? These units include individuals, households, groups/populations, health care organizations, health care programs, and national health care systems. Second, how were these data collected? The methods used to collect data include literature reviews, observations, focus groups, surveys, medical records and administrative and billing sources, registries, and vital records. Third, which government agency or private organization collected and is currently holding these data? Government data collection and holding agencies include US health information clearinghouses and libraries, US registries, US government agencies and departments, health programs and systems of other (non-US) nations, and government sponsored international organizations. Private data collecting and holding organizations include health information clearinghouses and libraries; accreditation, evaluation, and regulatory organizations; associations and professional societies; foundations and trusts; health insurance and employee benefits organizations; registries; research and policy organizations; and survey research organizations. To illustrate each of the questions and classifications, many examples are provided and discussed. And many US and other public use data files are identified and described.

---

## Introduction

Health services research is a growing field of study that is becoming increasingly important to society. As medicine progresses and increasingly saves lives, becomes more technologically complex, and is ever more costly and demands a greater share of society's resources, a growing number of people are conducting health services research studies. These researchers include physicians, nurses, epidemiologist, demographers, health economists, medical sociologists, political scientists, public policymakers, hospital administrators, insurance executives, senior business managers, and consultants.

Health services research can be broadly defined as a multidisciplinary field of study that focuses on

assess, cost, quality, and the outcome of health care (Mullner 2009). Access to health care, which can be defined as encompassing everything that facilitates or impedes the use of health care services, is a basic requirement of any health care facility, program, or system. A number of factors influence an individual's access to health care including the environment, population characteristics, health behavior, and outcomes. Environmental factors include the health care system (e.g., whether it is acceptable to the individual or not) and external environmental factors (e.g., geographic distance, physical, and political barriers). Population characteristics include predisposing characteristics (e.g., age, and gender), enabling resources (e.g., income, and health insurance), and perceived need (e.g., health beliefs). Health behavior includes personal health practices and previous use of health services. Lastly, outcomes include perceived health status, evaluated health status, and consumer satisfaction with care (Andersen 1995).

Health services researchers studying access to health care investigate various topics such as identifying ethnic and racial disparities in medical care; determining the geographic locations of health professional shortage areas; studying the factors associated with the diffusion and use of new medical technology and facilities; measuring access to hospitals and other health care facilities; and identifying the availability of health insurance coverage, and determining its impact on the use of health care services (Agency for Healthcare Research and Quality 2014).

Cost of health care, which can be defined as the amount that has to be paid or spent to buy or obtain health care, can be differentiated and measured in many ways including average cost, fixed cost, incremental cost, marginal cost, total cost, and variable cost, as well as direct and indirect cost, avoided cost, cost of lost productivity, and the societal cost of illness (Culyer 2010; Feldstein 2011). It should be noted that health care cost frequently differs greatly from the price of health care, because the price is often not determined by cost, but rather it is greatly distorted by what health insurers are willing to pay (Painter and Chernew 2012).

Health services researchers studying the cost of health care investigate a large number of topics

such as conducting international comparisons of health care cost in various nations, determining the cost-benefit and cost-efficiency of medical procedures and drugs, investigating the impact of different methods of financing care, determining the impact of new payment reform models (i.e., pay-for-performance), identifying the impact of health care rationing, estimating the economic value of life, and identifying the economic and societal cost of particular medical conditions and diseases (Health Care Cost Institute 2014).

Quality of health care, which can be defined as getting the right care to the right patient at the right time – every time, is evaluated using three dimensions: structure, process, and outcome. Structure includes the characteristics of the care setting, such as type and size of the health facility, expertise of the medical staff, sophistication of the medical equipment, and the policies related to care delivery. Process consists of the methods of providing patients with consistent, appropriate, effective, safe, coordinated, timely, and patient-centered high quality care. Outcome evaluates the end result of care the patient received (Clancy 2009; Donabedian 1980).

Health services researchers studying the quality of health care investigate such topics as identifying the impact of accreditation and licensing of health care facilities and professionals; estimating the overuse, underuse, and misuse of health care services; determining the occurrence of preventable medical errors; identifying the frequency of health care-associated infections; studying patient safety problems; and developing and testing new medical quality indicators of care (Agency for Healthcare Research and Quality 2014; National Committee for Quality Assurance 2006).

Lastly, the outcome of health care reflects the interrelated issues of access, cost, and quality of care. Outcome of health care can be broadly defined and includes the occurrence and change in the number and rate of death, disease, disability, discomfort, and dissatisfaction with health care. Death or mortality also includes changes in longevity. Disease or morbidity addresses acute and chronic disease and complications with medical care. Disability deals with the change in physical functional status and psychosocial functioning.

Discomfort includes various levels of pain from “no pain” to “worst pain imaginable” and its duration. And dissatisfaction, which is the level of satisfaction, measures the specific and overall experience with care (Kane and Radosevich 2011).

Health services researchers studying the outcome of health care tend to investigate such topics as estimating the number of preventable deaths of enrollees in various health programs; determining the factors leading to the increase in longevity; identifying the health services provided to children and adolescents with chronic diseases and disabilities; developing and testing new pain scales; and analyzing and reporting the results of health satisfaction surveys (Halsey 2015; Perrin 2002; Williamson and Hoggart 2008).

Ideally, a health care facility, program, or system should provide the greatest access to health care, at the lowest possible cost, with the greatest level of quality, and achieve the best possible outcome of care. To work towards, this very difficult ideal, health services researchers frequently study the equity, efficiency, and effectiveness of health care. Equity can be broadly defined as fairness, efficiency as the ratio of inputs to outputs, and effectiveness as meeting stated objectives and goals, such as the US national health goals contained in *Healthy People, 2020* (Aday et al. 2004).

The overall aim of health services research is to influence health policy and to improve the practice of medicine and public health. Health services researchers do this by monitoring and evaluating current health care facilities, programs, and systems and by predicting the consequences of proposed future health care policies.

Health services researchers frequently conduct studies using existing data sources. They typically conduct secondary data analysis of large databases that were collected by various government agencies and private organizations. There are many advantages in using existing data: they are readily available, inexpensive, and save time in collection, and they may be used to conduct longitudinal and international comparisons (Huston and Naylor 1996).

Primarily focusing on US data sources, this chapter outlines a practical typology, or classification

framework, of health care data that is frequently used by these researchers. The typology addresses three important inextricably linked questions. First, what is the basic unit of analysis of the study? Second, how were these data collected? Third, which government agency or private organization collected and is currently holding these data?

## Basic Units of Analysis

After identifying a particular study area of interest, and a specific topic, a health services researcher must determine – What will be the basic unit of analysis of my proposed study? Table 1 shows a list of these units; it also presents some relevant questions that may be addressed for each unit. The basic units of analysis include individuals, households, groups/populations, health care organizations, health care programs, and national health care systems.

### Individuals

Many health services researchers conduct their studies focusing on individuals. Information on individuals may be obtained from many sources such as patient health care records, birth and death certificates, insurance claim forms, and various national health surveys. Data on them may include a very large number of potential variables including the person's age, sex, height, weight, race, ethnicity, place of birth, language most often spoken, marital status, highest level of education attained, main occupation, current work status, health insurance coverage, past medical history, current overall health status, physical activities, degree of mobility, disability status, individual risk factors (tobacco, alcohol use, and poor nutrition), environmental risk factors (air pollution, ground water contamination, and lack of sanitation), self care, the level of pain and discomfort experienced, cognition problems, interpersonal activities, sleep and energy level, inventory of medicines and drugs, health seeking behaviors, health screenings, reproductive and sexual health care, maternal health care, child health preventive care, and health goals. An example of a widely used

**Table 1** Basic units of analysis

<b>Individuals</b>
Identify the general demographic and social characteristics of individuals
Determine the overall health status of individuals
Measure the occurrence of specific diseases and medical conditions
<b>Households</b>
Identify the demographic and social characteristics of households
Measure the total household income and education levels
Determine the households overall use of health care services
<b>Groups/populations</b>
Identify the demographic, economic, and social characteristics of specific ethnic and minority groups
Determine the overall health status of high risk and vulnerable populations
Measure the gaps in health care among various groups
Identify health professional shortage areas
<b>Organizations</b>
Identify the total number health care organizations in a region
Access the operating characteristics of hospitals
Determine the number of long-term care facilities in an area
Measure the service areas and degree of competition between healthcare organizations
<b>Health care programs</b>
Identify the characteristics of Medicare beneficiaries
Access the number of type of providers of Medicaid services
Determine the unwarranted use of services
<b>National health care systems</b>
Compare the access, costs, quality, and outcomes of various national health care systems
Determine how each system rations care
Identify by country the highest and lowest levels of care

individual health questionnaire is the “World Health Survey, 2002,” which was implemented in 70 member states (countries) to gather data on a sample of 300,000 adults. Data from the surveys were used to strengthen each country's capacity to monitor critical health outcomes and systems. A copy of the long- and short-survey instruments can be found on WHO's websites, [www.who.int/healthinfo/survey/en/](http://www.who.int/healthinfo/survey/en/) (WHO 2002).

Another very important large-scale survey of individuals is the US Centers for Disease Control

and Prevention's (CDC) Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is a nationwide surveillance system that is conducted to monitor state-level prevalence of the major behavioral risks (e.g., exercise, alcohol consumption, tobacco use, immunizations, and various cancer screening) among adults who have conditions associated with premature morbidity and mortality. To collect data, the CDC works together with state health departments and conducts monthly telephone surveys. Currently, more than 500,000 interviews are conducted annually making the BRFSS the world's largest telephone survey. Data from the survey are published in various reports (Xu et al. 2014), and annual survey data for 1984–2013 can be downloaded. The BRFSS also offers statistical tools, Web Enabled Analysis Tool (WEAT), which let researchers conduct cross tabulations and logistic regression analysis, and an interactive mapping program to compare data across geographic areas. More information on the BRFSS can be obtained at: [www.cdc.gov/brfss/](http://www.cdc.gov/brfss/).

## Households

Health services researchers often study the health and health care seeking behavioral characteristics of households. In the USA, they frequently use data collected by the Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS), the nation's principal government health statistics agency. Many of NCHS' surveys collect data on the demographic, socioeconomic, and the health characteristics of households (NCHS "Summary" 2014b).

The oldest and arguably the most important National Center for Health Statistics' household survey is the National Health Interview Survey (NHIS). The NHIS, which is considered the principal source of information on the health of the US population, has been used to continuously monitor the nation's health since 1957. This large-scale household survey collects data on a statistically representative sample of the US civilian noninstitutional population. Each year interviewers visit 35,000–40,000 households across the nation and collect data on about 75,000–100,000 individuals.

The survey collects data on a broad range of topics including access to health care services, health insurance coverage, physical and mental health status, chronic medical conditions, health-related behaviors, functioning and activity limitations, immunizations, and injuries and poisonings. Current and past NHIS data public use files, questionnaires, documentation, and analytic reports are readily available and can be downloaded for free from NCHS' website [www.cdc.gov/nchs/nhis.htm](http://www.cdc.gov/nchs/nhis.htm) (NCHS 2010).

## Groups/Populations

When general sample surveys of individuals or households do not adequately yield reliable health care data on specific groups or populations, supplements may be added to existing surveys or new surveys may be developed to obtain data on those groups or populations. Some of these groups or populations may include racial and ethnic minority groups (American Indians and Alaska Natives, Asians, Native Hawaiian and Pacific Islanders, Blacks, and Hispanic/Latinos), high risk and vulnerable populations (infants, children under 5 years of age, pregnant women, and the elderly), and groups with a specific disease and medical condition (blind, hearing loss, and the severely disabled).

To obtain information on a group or population, the National Health Interview Survey (NHIS) often adds supplements to its standard survey and expands the number of households sampled. These supplements are sponsored by various government agencies and nonprofit organizations. In 2014, for example, the NHIS added 4,000 additional households to its survey to obtain more data on the health of Native Hawaiian and Pacific Islanders (NCHS 2014a).

## Health Care Organizations

Health services researchers frequently study health care organizations. They study many types of organizations such as medical group practices, outpatient surgery centers, home care

organizations, Health Maintenance Organizations (HMOs), and Accountable Care Organizations (ACOs). But they particularly study hospitals and nursing homes.

The hospital is arguably the single most important institution for the delivery of modern health care, while the nursing home is the major institution caring for the elderly. The most widely used data source on US hospitals is the American Hospital Association's (AHA) annual survey of hospitals. And the most important source on nursing homes data is the US Department of Health and Human Services, Centers for Medicare and Medicaid Services' (CMS) nursing home compare data program.

The American Hospital Association (AHA) conducts an annual survey of the nation's approximately 6,000 hospitals, which account for 920,000-staffed beds and 36 million admissions. The survey, which is the most comprehensive and authoritative source on US hospitals, collects almost 900 variables on each hospital. These data include the hospital's address, bed size, ownership (for-profit, not-for-profit, government), type of hospital (community, psychiatric, long-term care, federal, and units of institutions), membership in a multihospital system or network, teaching status, type of facilities and services offered, physician arrangements, information technology, total number of inpatients and outpatient visits, Medicare/Medicaid utilization, revenues and expenses, and number of hospital staff. Data from the survey are published in the annual *AHA Guide to the Health Care Field* and *AHA Hospital Statistics*, and the proprietary data can be purchased on CD (AHA Data Viewer 2015; AHA 2013).

The Centers for Medicare and Medicaid Services (CMS), which administers the nation's Medicare program and works in partnership with state governments to administer Medicaid programs, continuously gathers data on the country's nearly 16,000 certified-nursing homes. These nursing homes provide services to over 1.4 million residents, corresponding to nearly 3 % of the nation's over 65 population and 10 % of the over 85 population. Because CMS pays for nursing home services provided to Medicare beneficiaries and Medicaid recipients, it continuously

monitors and updates its files on them. CMS collects data on the address of each nursing home; the facilities' bed size, ownership type, and certification; number of nursing home residents; demographic and medical characteristics of the residents, including cognitive and functional impairments; and the number, type, and level of deficiencies these facilities experienced. The deficiencies include citations for substandard quality of care; abuse; improper restraint use; pressure sores; actual harm or worse; and the immediate jeopardy threat to the health or life of one or more nursing home residents. Data on individual nursing homes can be obtained at CMS's Medicare.gov Nursing Home Compare website [www.medicare.gov/nursinghomecompare](http://www.medicare.gov/nursinghomecompare), the entire database can be downloaded, and a summary nursing home data compendium is published annually, which is also available on the website (CMS 2014).

## Health Care Programs

Many health services researchers study large national health care programs. One of the most widely studied is the US Medicare program. This federal government administered national program provides health insurance for over 50 million people, including those 65 years of age or older, those with certain disabilities, and people of any age with End-Stage Renal Disease (ESRD) (permanent kidney failure requiring dialysis or a kidney transplant).

The Medicare program consists of four different parts: Part A (hospital insurance covering inpatient care, nursing home, hospice, and home health care), Part B (medical insurance covering physician services, outpatient and home health care, and durable medical equipment), Part C (Medicare Advantage, a managed care program covering Part A and B), and Part D (covering prescription drugs).

The program collects data on its various parts including claims for services provided to each beneficiary admitted to a certified hospital and nursing home. It codes the beneficiaries' address,

where they received care, their medical diagnoses, admission date, what services were provided, discharge date, discharge status, cost of each service, and the total cost of care. If the beneficiary dies after receiving care, it is coded up to 3 years after discharge. One widely used CMS database is the Medicare Provider Analysis and Review (MEDPAR) file, which can be obtained from CMS' website [www.cms.gov/Research-Statistics-Data-and-Systems/IdentifiableDataFiles/MedicareProviderAnalysisandReviewFile.html](http://www.cms.gov/Research-Statistics-Data-and-Systems/IdentifiableDataFiles/MedicareProviderAnalysisandReviewFile.html) (CMS 2014).

An exemplar of the innovative use of the MEDPAR database is the research conducted by the Dartmouth Atlas of Health Care Project. Health services researchers working on the project, which is housed at Dartmouth University's Institute for Health Policy and Clinical Practice, have studied a wide range of medical practice patterns at the national, regional, state, and local levels. For more than 20 years, these researchers have found and documented glaring unwarranted variations in surgeries, diagnostic testing, imaging exams, physician visits, referrals to specialists, hospitalizations, and stays in intensive care units. They have consistently found that more health care is not necessarily better care (Dartmouth Atlas of Health Care 2015).

Using Medicare data, the Dartmouth researchers have identified three broad categories of medical care: effective or necessary care, preference-sensitive care, and supply-sensitive care. Effective or necessary care includes services that are based on sound medical evidence, which work better than any alternative treatment (e.g., surgery for hip fractures and colon cancer). They estimate that this category of care accounts for no more than 15 % of total Medicare spending. Preference-sensitive or elective care includes interventions for which there are several options and where the outcomes vary depending on the option used (e.g., elective surgeries, mammography screening tests, and prostate specific antigen tests). This accounts for about 25 % of Medicare spending. Lastly, supply-sensitive care includes everyday medical care used to treat patients with acute and chronic diseases (e.g., physician visits, imaging exams, and admissions to hospitals). This

accounts for about 60 % of all Medicare spending.

To remedy the unwarranted variations in preference-sensitive care, the Dartmouth researchers argue for the greater use of evidence-based medicine to identify the best option, and they call for a fundamental reform of the physician-patient relationship, with greater shared decision-making and informed patient choice. To remedy the variations in supply-sensitive care, they argue that the common physician assumption that "more care is better" needs to change and there must be a new emphasis on improving the science of health care delivery (Wennberg 2010).

Published reports of the Dartmouth Atlas of Health Care Project as well as the data they used in many of their studies can be downloaded from their website [www.dartmouthatlas.org](http://www.dartmouthatlas.org).

## National Health Care Systems

Lastly, some health services researchers conduct cross-national studies of health care systems, such as comparing the US health system to that of Canada, the United Kingdom, and other industrialized nations. It is hoped that these multinational comparisons may help health policymakers learn from the experiences of other nations, lead to new insights and perspectives, held in evaluating existing policies, and identify possible new solutions to shared problems.

Three important sources of data on national health care systems are the World Health Organization (WHO), Organisation for Economic Co-operation and Development (OECD), and the Commonwealth Fund.

The World Health Organization (WHO), which is the directing and coordinating authority for health within the United Nations (UN), collects health-related data on its 194 member states (nations). These data on the states are published in its series *World Health Statistics*. Issued annually since 2005, *World Health Statistics* is the definitive source of information on the health of the world's people. The series is compiled using publications and databases produced and



maintained by the WHO's technical programs and regional offices, and from various databases of the UN and World Bank. Data in the publication provide a comprehensive summary of the current health status and health system of each member state. These data include nine areas: life expectancy and mortality, cause-specific morbidity and mortality, selected infectious diseases, health service coverage, risk factors, health systems, health expenditures, health inequities, and demographic and socioeconomic statistics. WHO's data in published form are available on its website [www.who.int](http://www.who.int) (WHO 2014).

The Organisation for Economic Co-operation and Development (OECD) is an international membership organization representing 34 industrialized nations that are committed to democracy and a free market economy. The OECD, working with its member nations, produces data and reports on a wide variety of economic and social topics, including health care. Each year it releases data comparing the health care systems of its member nations including: health care spending – average spending per capita, spending as a percentage of GDP, spending per hospital discharge, and pharmaceutical spending per capita; supply and utilization—number of practicing physicians per population, average number of physician visits per capita, Magnetic Resonance Imaging (MRI) machines per population, hospital discharges per population, and hip replacement inpatient cases per population; health promotion and disease prevention efforts – cervical cancer screening rates, flu immunization among adults 65 or older, and adults who report being daily smokers; quality and patient safety – mortality amenable to health care, breast cancer 5-year survival rate, and diabetes lower extremity amputation rates; prices – total hospital and physician prices for appendectomy and bypass surgery, diagnostic imaging prices, and long-term care and social supports – percent of population age 65 or older, beds in residential long-term care facilities per population age 65 or older, and health and social care spending as a percentage of GDP. OECD data and its reports, which are frequently used by health services researchers (Anderson 2014; Anderson and Squires 2010), can be downloaded from their website [www.oecd.org/statistics/](http://www.oecd.org/statistics/) (OECD 2013).

The Commonwealth Fund, a private, nonpartisan foundation headquartered in New York City that supports independent research on health care issues to improve health care practice and policy, conducts annual cross-national studies. Starting in 1998, its International Health Policy Center has conducted multinational surveys of patients and their physicians to identify their experiences with their health care systems. The surveys focus on various aspects of access, costs, and quality of health care.

One of the center's recent surveys was the "2014 Commonwealth Fund International Health Policy Survey of Older Adults," a telephone interview survey of more than 15,000 people age 65 or older in 11 industrialized countries (Australia, Canada, France, Germany, the Netherlands, New Zealand, Norway, Sweden, Switzerland, the United Kingdom, and the United States). The survey's major finding was that older adults in the US were sicker and more likely to have problems paying their medical bills and getting needed health care than those in the other 10 countries (Osborn et al. 2014).

The center has also conducted five surveys to monitor changes in multinational health care system performance, and the results have been published in a series of reports entitled *Mirror*; *Mirror on the Wall* (2004, 2006, 2007, 2010, 2014). Over the years, these reports have consistently found that among industrialized nations the US health care system has been the most expensive, but underperforms relative to other nations on most dimensions on access, efficiency, and equity (Davis et al. 2014).

---

## Collection Methods

The second question of this typology of health care data is – How were these data collected? This question is important, because the way the data were collected may limit the type of statistical methodology that can be used to analyze them, and it may greatly affect the reliability and validity of the results of the study. Each data collection method has advantages and disadvantages and the researcher should be well aware of

them. Table 2 shows the various data collection methods, and it also lists some relevant questions that may be addressed by each method. The methods include literature reviews; observations;

focus groups; surveys; medical records, administrative, and billing sources; registries; and vital records.

**Table 2** Data collection methods

<b>Literature reviews</b>
Identify what is known about a particular health care topic
Determine what are the gaps in knowledge on the topic
Conduct a meta-analysis to assess the clinical effectiveness of a health care intervention
Answer a research question
<b>Observations</b>
Observe patients taking their treatments
Measure the degree of hand hygiene adherence at a health care organization
Conduct a clinical observation, or shadowing, to determine how health care professionals actually provide patient care
<b>Focus groups</b>
Determine the perceptions, opinions, beliefs, and attitudes towards a health program
Identify specific problems with a health facility
Present options to a group and see which ones are viewed favorably
<b>Surveys</b>
Determine the past medical history of individuals
Identify the experiences of patients in receiving care
Measure the workload of physicians and other health care professionals
<b>Medical records, administrative, and billing sources</b>
Identify and implement best practices of care
Determine regional variations in the provision of health care
Measure the average costs of various health care services
<b>Registries</b>
Identify the occurrence of a disease within a population
Assess the natural history of a disease, its management, and its outcomes
Support health economic research
Collect postmarketing safety data on medical products and pharmaceuticals
<b>Vital records</b>
Determine trends in fetal and perinatal mortality
Identify the relationship between infant birth weight and health care problems
Determine trends in low-risk Cesarean delivery
Identify trends in drug-poisoning deaths involving opioid analgesics and heroin

## Literature Reviews

One of the easiest, fastest, and most economical ways to obtain data and information on a research topic or a specific research question is to conduct a literature review. A comprehensive literature review can help identify what is known and not known about a topic or question; what data sources are available; what variables were found to be important; what statistical methods were employed; what populations were studied; what sample sizes were used; and what are the gaps or possible errors in the studies.

A major resource in conducting literature reviews is the US National Library of Medicine's (NLM) PubMed search engine. PubMed accesses MEDLINE and other databases of citations and abstracts in the fields of medicine, nursing, public health, and health care systems. Currently, PubMed contains more than 24 million citations from over 5,600 worldwide journals and thousands of books and reports. PubMed is easy to use, it can be searched by entering Medical Subject Headings (MeSH) the NLM's controlled vocabulary, author names, title words or phrases, journal names, or any combination of these. It also links to many full-text articles and reports. The PubMed's website is: [www.ncbi.nlm.nih.gov/pubmed](http://www.ncbi.nlm.nih.gov/pubmed).

Another important source for conducting literature reviews is the Cochrane Collaboration. Consisting of a network of 14 centers around the world, the Cochrane Collaboration is a nonprofit international organization that promotes and disseminates systematic reviews of health care interventions, particularly clinical trials. Collaborators from over 120 countries conduct these systematic reviews. The Cochrane Library contains a number of useful databases including Cochrane Database of Systematic Reviews (CDSR); Cochrane Controlled Trials Register (CENTRAL); Database of Abstracts of Reviews of Effectiveness (DARE); Cochrane Methodology Register; Health Technology Assessment Database (HTA); and the

National Health Service Economic Evaluation Database (NHS EED). The Cochrane Collaboration's website is [www.cochrane.org](http://www.cochrane.org).

Many of the Cochrane Collaboration's systematic reviews include a meta-analysis of studies. Meta-analysis is a statistical technique that combines the findings from multiple research studies to develop a single conclusion that has greater statistical power. By pulling together a number of independent studies, researchers can make a more accurate estimate of the effect of a result (Borenstein et al. 2009; Higgins and Green 2008).

## Observations

Health services researchers sometimes conduct observational studies to obtain data. In these types of studies, individuals are observed or certain outcomes are measured, but no attempt is made to affect the outcome. They do not involve an experiment or intervention. Observational studies may be either cross-sectional or longitudinal. Cross-sectional studies are short quick snapshot studies, and they do not provide definitive information about a cause-and-effect relationship. However, longitudinal studies that are conducted over long periods of time with many observations can determine changes in individuals and populations. They can establish the sequence of events and suggest a cause-and-effect relationship.

Observational studies can vary greatly in size, scope, and complexity. Some observational studies are very small, inexpensive, quickly conducted, cross-sectional studies. An example of such a study would be a researcher investigating the waiting times of patients at a health care clinic. He or she might conduct the study by unobtrusively sitting in the waiting room for a few days observing and coding the demographic characteristics of each patient and the number of minutes they waited to be seen.

In contrast, other observational studies are very large, expensive, lengthy, longitudinal studies. One of the most famous longitudinal observational studies in modern medicine is the Framingham Heart Study. Begun in 1948 and continuing to

the present, this study has followed large cohorts of individuals from Framingham, Massachusetts, to determine their risk of developing cardiovascular disease. Today, much of what is now-common knowledge concerning the major risk factors of developing heart disease (hypertension, high "bad" cholesterol, diabetes, smoking, obesity, and a sedentary lifestyle) is based on the Framingham Study (Levy and Brink 2005).

## Focus Groups

Occasionally, health services researchers conduct focus groups to obtain data. Focus groups generally consist of five to ten participants who are asked their opinions about a topic in a group interview. Although the interviews are informal, open-ended, and relatively broad, a moderator asks the group a series of questions to help direct the discussion. Focus groups may be used to explore new research areas, topics that are difficult to observe, and very sensitive topics. They may also be used to gather preliminary data, aid in survey development and more formal structured interviews, and clarify complex research findings. As the focus group session is occurring, it is audio- and/or video-recorded. These recordings are then transcribed, reviewed, and studied.

Focus groups have advantages as well as disadvantages. They may generate new ideas and allow clarification of issues, and the group members may stimulate each other. However, members and the moderator can bias responses; some members may dominate the group; and the results of the focus group may be difficult to analyze or quantify (Krueger and Casey 2009).

Recently, the Robert Wood Johnson Foundation conducted a series of focus groups to gather information on what consumers think about the rising cost of health care in the USA. The foundation convened eight focus groups in four cities: Philadelphia; Charlotte, North Carolina; Chicago; and Denver. The participants included individuals with employer-sponsored insurance, those who purchased their insurance on the private market, those enrolled in Medicare, and those without any health insurance coverage. The major findings of

the focus groups were that the participants were very aware of their actual health care costs; they were aware of the rising costs of care, but did not understand why it was happening; the rising costs were affecting their daily lives and purchases; and they were increasingly angry about the increasing costs, but felt helpless in reversing the trends (Robert Wood Johnson Foundation 2013).

## Surveys

Some health services researchers rely heavily on surveys to gather data for their studies. They occasionally conduct their own health care surveys, but more often use data from surveys conducted by others. Using these data, they conduct health needs assessments, develop health profiles of groups/populations, monitor the health of cohorts and populations, and collect pre- and posttest health care measures.

Health care surveys are a very effective and efficient method of estimating the characteristics of large groups/populations using representative samples. Most health surveys are conducted with a large number of participants who are randomly selected to reduce the risk of selection bias. The surveys collect data in a structured, standardized manner from each respondent. Lastly, these data are typically summarized as counts or persons or events.

Health survey data are collected using two broad strategies, and the respondents are asked to reply to questions presented in questionnaires or read aloud by interviewers. These two strategies may be employed individually or in combination.

The most widely used type of survey is the self-administered mailed survey, whereby a questionnaire and an introductory cover letter are sent via standard mail to a sample of persons. The respondents are asked to complete the questionnaire and return it to the researcher using a preaddressed return envelope enclosed with the questionnaire. With the increasing use of home computers, self-administered surveys are also increasingly being sent to respondents via e-mail and the Internet.

Some health surveys are conducted by interviews, which may be completed over the telephone or face-to-face. Telephone interviews are more frequently used because of their versatility, data quality, and cost and time efficiency. In contrast, face-to-face interviews are generally considered to provide the very best data quality, but they are the most expensive and time-consuming surveys to complete (Aday and Cornelius 2006; Johnson 2014).

To collect longitudinal data to measure changes over time, health services researchers periodically send surveys to a panel of individuals or organizational respondents. An example of such a survey is the US Agency for Healthcare Research and Quality's (AHRQ) Medical Expenditure Panel Survey (MEPS). Begun in 1996, MEPS is a set of surveys of individuals and households, their medical providers, and employers across the nation. MEPS collects data to estimate the frequency and use of specific health services, the cost and payment for these services, and the health insurance coverage held by and available to US workers.

Specifically, MEPS consists of three components: household, insurance, and other. The household component collects panel data from a sample of families and individuals using several rounds of interviewing conducted over 2 years. Data from the interviews make it possible for researchers to identify how the changes in the respondent's health status, income, employment, health insurance, use of services, and payment of care are related. The insurance component gathers data by surveying employers about the health insurance coverage they offer their workers. The other component collects data on the hospitals, physicians, home health care providers, and pharmacies that provided care to respondents. It is used to supplement and/or replace information received from the respondents.

Data obtain from MEPS are published in various statistical briefs, which can be downloaded. Recent briefs have reported on the access to health care by adult men and women, ages 18–64 (Davis 2014); the number and characteristics of the long-term uninsured (Rhoades and Cohen 2014); and national health care expenses

by type of service and source of payment (Stagnitti and Carper 2014). MEPS household component public use data files and insurance component summary data tables are released on AHRQ's MEPS website on a regular annual schedule, [http://meps.ahrq.gov/mepsweb/about\\_meps/releaseschedule.jsp](http://meps.ahrq.gov/mepsweb/about_meps/releaseschedule.jsp).

### **Medical Records, Administrative, and Billing Sources**

A rich source of health care data can be obtained from medical records, administrative, and billing sources. The most widely used and easily accessible source of this type of data is the Medicare claims files. These data files have been widely used by health services researchers to identify: the factors that influence hospitalization; the geographic variations in the type of care patients receive, such as the previously discussed Dartmouth Atlas of Health Care Project; the cost-effectiveness of various clinical procedures; and the effect of health reform efforts such as the Affordable Care Act (ACA) on Medicare utilization rates.

CMS has numerous data files available to researchers. However, because of privacy concerns, some of the files are more restricted than others. CMS classifies its files into three categories: Research Identifiable Files (RIF), which are the most restricted files because they contain patient and condition identifiable data; Limited Data Sets (LDS), which are less restricted files because their patient-specific data are ranged or encrypted; and Public Use Files (PUF)/Non-identifiable Files, which are the least restricted files of all, are readily available, and can be easily downloaded.

CMS has released a number of public use data files. These "Basic Stand Alone (BSA) Medicare Claims Public Use Files (PUFs)" mainly consist of 5 % random samples of all Medicare beneficiaries from a reference year. Examples of these data files include: Hospital Inpatient Claims file, containing the variables: age, gender, base DRG, ICD-9 procedure code, length of stay, and the amount paid; Durable Medical Equipment (DME) Line Items

file, containing a list of equipment provided such as oxygen equipment, hospital beds, and wheelchairs; Prescription Drug Events file, containing the variables: age, gender, drug name, dose, cost, and payment by patient; Hospice Beneficiary file, containing the variables: age, gender, and length of stay; Carrier Line Items file, containing physician/supplier medical claims data, dates of service, and reimbursement amounts; Home Health Agency (HHA) Beneficiary file, containing demographic and claim-related variables; Outpatient Procedures file, containing demographic variables and procedures provided; Skilled Nursing Facility (SNF) Beneficiary file, containing demographic and nursing home claims; Chronic Conditions file, containing age, gender, various chronic conditions, and dual-eligibility status; Institutional Provider and Beneficiary Summary file, containing data on Medicare institutional claims paid during the calendar year and a summary of other measures; Prescription Drug Profiles file, containing demographic variables, plan-drug-and prescriber characteristics, and payment data; and the Geographic Variation Public Use file, containing demographic, spending, utilization, and quality indicators at the state, hospital referral region, and county level.

Further information about the data files can be obtained from the CMS-funded Research Data Assistance Center (ResDAC), which is located at the University of Minnesota, Minneapolis. Its website is [www.resdac.org](http://www.resdac.org).

### **Registries**

Health services researchers occasionally use data from registries to conduct their studies. Registries are tools that systematically collect a defined set of exposures, health conditions, and demographic data about individuals, with the data held in a central database for a specific purpose. They are used for a multitude of purposes including monitoring treatment benefits and risks, understanding the natural history of diseases, identifying unmet medical needs, and determining the quality of care. Registries can vary greatly in size, scope, and duration. Some registries collect data at a

single clinic for a few weeks, while others are international in scope and collect data for many decades. Registries may be sponsored by government agencies, nonprofit organizations, health care facilities, and/or private for-profit companies (Arts et al. 2002).

It is difficult to classify the various types of registries because of their great diversity and scope. Also, they may collect overlapping sets of data. However, they can be very roughly divided into product registries, disease or condition registries, and health services registries.

Product registries gather data on individuals who received a specific drug or medical device. To ensure safety, these registries have been established to monitor individuals who received such drugs as thalidomide, and those who were given medical devices such as implantable cardioverter defibrillators. Registries have also been established to monitor possible drug exposures during pregnancy and the neonatal consequences.

Disease or condition registries gather data on individuals with specific disorders. These registries may identify the natural history of a disease, evaluate possible treatments, and stimulate new research on the cause and outcome of the disorder. Diseases included in these registries can vary from rare diseases such as cystic fibrosis, to relatively common chronic diseases such as heart failure.

Health services registries tend to gather data on individual clinical encounters such as physician office visits, hospitalizations, clinical procedures, and total episodes of care. Some registries include all patients undergoing a procedure such as an appendectomy or those admitted to a hospital for a particular diagnosis such as community-acquired pneumonia. Many of these registries are used to evaluate the outcome of care and the associated quality of health care services (Gliklich and Dreyer 2010).

An example of a unique health services registry is the Health Resources and Services Administration's (HRSA) National Practitioner Data Bank (NPDB). The NPDB is a critical tool in the US' efforts to protect patients from incompetent, unprofessional, and often dangerous health care practitioners. Since 1986, the NPDB has collected

reports on medical malpractice payments, medical review actions, and sanctions by Board of Medical Examiners. It collects information from medical malpractice payments and adverse licensures, Drug Enforcement Administration (DEA) reports, and Medicare and Medicaid exclusion actions concerning physicians, dentists, and other licensed health care practitioners. The NPDB provides this information to health care providers, hospitals, and state and federal agencies to use when making important hiring or licensing decisions. This helps protect the public by preventing physicians and other practitioners from hiding their past when they move to a new state (Wakefield 2011).

The NPDB public use data file, which does not include any information that identifies individuals or reporting entities, is available for statistical analysis at [www.npdb.hrsa.gov/resources/publicData.jsp](http://www.npdb.hrsa.gov/resources/publicData.jsp).

## Vital Records

Vital records include birth certificates, marriage licenses and divorce decrees, and death certificates. In the USA, counties and state governments collect, manage, and disseminate vital records, not the federal government. Health services researchers frequently use data from birth and death certificates in their studies. They use these data to track health trends to determine changing public health and research priorities, identify racial and ethnic disparities, measure the impact of various diseases, ascertain the use of health care services, and to address quality of care issues (Children's Health Care Quality measures Core Set Technical Assistance and Analytic Support Program 2014; National Research Council 2009).

The US Standard Certificate of Live Birth contains a wealth of information on the newborn, as well as the mother and father. Data on the newborn include name, sex, time and place of birth, birth weight, Apgar scores, abnormal conditions, and congenital anomalies of the newborn. Data on the mother include name; address; education level; whether of Hispanic origin or not; race; date of first and last prenatal care visit; total number of prenatal visits; number of other pregnancy

outcomes; the degree of cigarette smoking before and during pregnancy; whether the mother was transferred for maternal medical or fetal indications for delivery; principal source of payment for the delivery; risk factors in the pregnancy such as diabetes, hypertension, and previous preterm birth; obstetric procedures used; onset of labor; characteristics of the labor and delivery; method of delivery; and maternal morbidity. Data on the father include: name, age, education level, whether of Hispanic origin or not; and race.

The US Standard Certificate of Death records the decedent's: legal name; age; sex; social security number; birthplace; residence; marital status at the time of death; place of death; place of disposition; date of death; cause of death including the immediate and underlying cause; manner of death; if the injury lead to death, the date and time of injury; and the location of injury.

There is also a separate certificate for fetal deaths. The US Standard Report of Fetal Death collects data on: the name of the fetus; sex; date and place where delivery occurred; initiating cause/condition; other significant causes or conditions; risk factors in the pregnancy; infections present and/or treated during the pregnancy; method of delivery; maternal morbidity; and congenital anomalies of the fetus.

Although birth, death, and fetal death certificates are confidential legal records, they can be obtained for research purposes from state public health departments. Summary data on births, deaths, fetal deaths, and linked birth/infant deaths can also be obtained from the National Center for Health Statistics (NCHS). Its data release and access policy for microdata and compressed vital statistics files can be found at [www.cdc.gov/nchs/nvss/dvs\\_data\\_release.htm](http://www.cdc.gov/nchs/nvss/dvs_data_release.htm).

---

## Data Sources and Holdings

The third question of this typology is – Which government agency or private organization collected and is currently holding these data? A large number of government and private organizations collect and disseminate health care data. Many

private organizations, sometimes with government support through contracts and grants, also collect health care data for research purposes, to monitor health policies, and to identify their member's views and opinions on various issues. Table 3 shows the classification of health care data collection organizations and holding sources, including a list of various representative organizations and their websites.

## Government Organizations

Federal, state, and local governments collect data on the health care programs they conduct and manage. These data are often readily available to researchers at little or no cost. From the perspective of health services research, government data collection and holding agencies can be broadly classified into the following categories: US health information clearinghouses and libraries; US registries; US government agencies and departments; health programs and systems of other (non-US) nations; and government sponsored international organizations.

### US Health Information Clearinghouses and Libraries

The federal government maintains many clearinghouses and libraries that are valuable resources for health services research. For example, the National Institutes of Health's (NIH) National Library of Medicine (NLM) is the world's largest biomedical library. The NLM maintains and makes available its vast print collection and produces and continuously updates its electronic information resources such as PubMed/MEDLINE. PubMed comprises more than 24 million citations from MEDLINE. The NLM also contains the National Information Center on Health Services Research and Health Care Technology (NICHSR). This center maintains databases and provides outreach and training, and information and publications on health services research. Its website is [www.nlm.nih.gov/nichsr/](http://www.nlm.nih.gov/nichsr/).

The US government's principal health statistical agency is the US Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS). Since 1960, the NCHS has conducted numerous national health

**Table 3** Data collection organizations and holding sources

<b>Government Organizations</b>
<b>US Health Information Clearinghouses and Libraries</b>
Area Health Resource Files (AHRF), <a href="http://www.ahrh.hrsa.gov">www.ahrh.hrsa.gov</a>
Congressional Research Service, <a href="http://www.loc.gov/crsinfo/">www.loc.gov/crsinfo/</a>
Data.gov, <a href="http://www.data.gov">www.data.gov</a>
HealthCare.Gov, <a href="http://www.healthcare.gov">www.healthcare.gov</a>
National Center for Health Statistics (NCHS), <a href="http://www.cdc.gov/nchs/">www.cdc.gov/nchs/</a>
National (Evidence-Based Clinical Practice) Guideline Clearinghouse, <a href="http://www.guideline.gov">www.guideline.gov</a>
National Health Information Center (NHIC), <a href="http://www.health.gov/nhic/">www.health.gov/nhic/</a>
National Information Center on Health Services Research and Health Care Technology (NICHSR), <a href="http://www.nlm.nih.gov/nichsr/">www.nlm.nih.gov/nichsr/</a>
National Institute on Deafness and Other Communication Disorders (NIDCD), <a href="http://www.nidcd.nih.gov/health/misc/pages/clearinghouse.aspx">www.nidcd.nih.gov/health/misc/pages/clearinghouse.aspx</a>
National Library of Medicine (NLM), <a href="http://www.nlm.nih.gov">www.nlm.nih.gov</a>
National Mental Health Information Center, <a href="http://www.samhsa.gov">www.samhsa.gov</a>
National Oral Health Clearinghouse, <a href="http://www.nider.nih.gov">www.nider.nih.gov</a>
<b>U.S. Registries</b>
FDA Adverse Event Reporting System (FAERS), <a href="http://www.fda.gov">www.fda.gov</a>
Global Rare Diseases Patient Registry Data Repository (GRDR), <a href="http://www.rarediseases.info.nih.gov">www.rarediseases.info.nih.gov</a>
National Practitioner Data Bank (NPDB), <a href="http://www.npdb.hrsa.gov">www.npdb.hrsa.gov</a>
National Registry of Evidence-Based Programs and Practices (NREPP), <a href="http://www.nrepp.samhsa.gov">www.nrepp.samhsa.gov</a>
National Vital Statistics System (NVSS), <a href="http://www.cdc.gov/nchs/nvss/">www.cdc.gov/nchs/nvss/</a>
NIH Genetic Testing Registry (GTR), <a href="http://www.ncbi.nlm.nih.gov/gtr/">www.ncbi.nlm.nih.gov/gtr/</a>
Surveillance, Epidemiology, and End Results (SEER) Cancer Registry, <a href="http://www.seer.cancer.gov">www.seer.cancer.gov</a>
<b>US Government Agencies and Departments</b>
Agency for Healthcare Research and Quality (AHRQ), <a href="http://www.ahrq.gov">www.ahrq.gov</a>
Centers for Disease Control and Prevention (CDC), <a href="http://www.cdc.gov">www.cdc.gov</a>
Centers for Medicare and Medicaid Services (CMS), <a href="http://www.cms.gov">www.cms.gov</a>
Congressional Budget Office (CBO), <a href="http://www.cbo.gov">www.cbo.gov</a>
Employee Benefits Security Administration (EBSA), <a href="http://www.dol.gov/ebsa/">www.dol.gov/ebsa/</a>
Federal Trade Commission (FTC), <a href="http://www.ftc.gov">www.ftc.gov</a>
Food and Drug Administration (FDA), <a href="http://www.fda.gov">www.fda.gov</a>
Government Accountability Office (GAO), <a href="http://www.gao.gov">www.gao.gov</a>
Health Resources and Services Administration (HRSA), <a href="http://www.hrsa.gov">www.hrsa.gov</a>
Internal Revenue Service (IRS), <a href="http://www.irs.gov">www.irs.gov</a>
Medicare Payment Advisory Commission (MedPAC), <a href="http://www.medpac.gov">www.medpac.gov</a>
National Institute on Aging (NIA), <a href="http://www.nia.nih.gov">www.nia.nih.gov</a>
National Institute on Drug Abuse (NIDA), <a href="http://www.drugabuse.gov">www.drugabuse.gov</a>
National Institute of Health (NIH), <a href="http://www.nih.gov">www.nih.gov</a>
Office of the National Coordinator for Health Information Technology (ONC), <a href="http://www.hhs.gov/healthit">www.hhs.gov/healthit</a>
Presidential Commission for the Study of Bioethical Issues, <a href="http://www.bioethics.gov">www.bioethics.gov</a>
Substance Abuse and Mental Health Services Administration (SAMHSA), <a href="http://www.samhsa.gov">www.samhsa.gov</a>
US Agency for International Development (USAID), Bureau for Global Health, <a href="http://www.usaid.gov">www.usaid.gov</a>
US Census Bureau, <a href="http://www.census.gov">www.census.gov</a>
US Department of Health and Human Services (HHS), <a href="http://www.hhs.gov">www.hhs.gov</a>
US Department of Justice, <a href="http://www.usdoj.gov">www.usdoj.gov</a>
US Department of Labor Statistics, <a href="http://www.bls.gov">www.bls.gov</a>
US Department of Veterans Affairs (VA), <a href="http://www.va.gov">www.va.gov</a>
US House of Representatives, <a href="http://www.house.gov">www.house.gov</a>

(continued)



**Table 3** (continued)

US Senate, <a href="http://www.senate.gov">www.senate.gov</a>
US Social Security Administration (SSA), <a href="http://www.ssa.gov">www.ssa.gov</a>
White House, <a href="http://www.whitehouse.gov">www.whitehouse.gov</a>
<b>Health Programs and Systems of Other (non-U.S.) Nations</b>
Australian Commission on Safety and Quality in Health Care, <a href="http://www.humanservices.gov.au">www.humanservices.gov.au</a>
Australian Government Department of Human Services, <a href="http://www.humanservices.gov.au">www.humanservices.gov.au</a>
Canadian Agency for Drugs and Technologies in Health, <a href="http://www.cadth.ca">www.cadth.ca</a>
Canadian Institute for Health Information, <a href="http://www.cihi.ca">www.cihi.ca</a>
Canadian Institutes of Health Research, <a href="http://www.cihr-irsc.gc.ca">www.cihr-irsc.gc.ca</a>
Health Canada, <a href="http://www.hc-sc.gc.ca">www.hc-sc.gc.ca</a>
United Kingdom's National Health Service (NHS), <a href="http://www.nhs.uk">www.nhs.uk</a>
United Kingdom's National Institute for Health and Care Excellence (NICE), <a href="http://www.nice.org.uk">www.nice.org.uk</a>
<b>Government Sponsored International Organizations</b>
European Commission, <a href="http://www.ec.europa.eu">www.ec.europa.eu</a>
European Observatory on Health Systems and Policies, <a href="http://www.euro.who.int/en/about-us/partners/observatory/about-us">www.euro.who.int/en/about-us/partners/observatory/about-us</a>
Organisation for Economic Co-operation and Development (OECD), <a href="http://www.oecd.org">www.oecd.org</a>
Pan American Health Organization (PAHO), <a href="http://www.paho.org">www.paho.org</a>
United Nation Children's Fund (UNICEF), <a href="http://www.unicef.org">www.unicef.org</a>
United Nations (UN), <a href="http://www.un.org">www.un.org</a>
World Bank, <a href="http://www.worldbank.org">www.worldbank.org</a>
World Health Organization (WHO), <a href="http://www.who.int">www.who.int</a>
<b>Private Organizations</b>
<b>Health information clearinghouses and libraries</b>
Centre for Evidence-Based Medicine (CEBM), <a href="http://www.cebm.net">www.cebm.net</a>
Cochrane Collaboration, <a href="http://www.cochrane.org">www.cochrane.org</a>
Cornell Disability Research Group, <a href="http://www.disabilitystatistics.org">www.disabilitystatistics.org</a>
Dartmouth Atlas of Health Care Project, <a href="http://www.dartmouthatlas.org">www.dartmouthatlas.org</a>
Data Resource Center for Child and Adolescent Health, <a href="http://www.childhealthdata.org">www.childhealthdata.org</a>
Health Care Cost Institute (HCCI), <a href="http://www.healthcostinstitute.org">www.healthcostinstitute.org</a>
Health Data Consortium, <a href="http://www.healthdataconsortium.org">www.healthdataconsortium.org</a>
IMS Health, <a href="http://www.imshealth.com">www.imshealth.com</a>
Inter-University Consortium of Political and Social Research (ICPSR), <a href="http://www.icpsr.umich.edu">www.icpsr.umich.edu</a>
National Association of Health Data Organization (NAHDO), <a href="http://www.nahdo.org">www.nahdo.org</a>
National Implementation Research Network (NIRN), <a href="http://www.preventionaction.org">www.preventionaction.org</a>
National Rehabilitation Information Center (NARIC), <a href="http://www.naric.com">www.naric.com</a>
National Rural Health Resource Center, <a href="http://www.ruralcenter.org">www.ruralcenter.org</a>
<b>Accreditation, evaluation, and regulatory organizations</b>
Accreditation Association for Ambulatory Health Care (AAAHC), <a href="http://www.aaahc.org">www.aaahc.org</a>
Accreditation Canada, <a href="http://www.accreditation.ca">www.accreditation.ca</a>
Accreditation Commission for Health Care (ACHC), <a href="http://www.achc.org">www.achc.org</a>
Association of American Medical Colleges (AAMC), <a href="http://www.aamc.org">www.aamc.org</a>
Board of Certification/Accreditation (BOC), <a href="http://www.bocusa.org">www.bocusa.org</a>
Center for Improvement in Healthcare Quality (CIHQ), <a href="http://www.cihq.org">www.cihq.org</a>
Community Health Accreditation Partner (CHAP), <a href="http://www.chapinc.org">www.chapinc.org</a>
Det Norske Veritas (DNV) Healthcare, <a href="http://www.dnvglhealthcare.com">www.dnvglhealthcare.com</a>
Health Grades, <a href="http://www.healthgrades.com">www.healthgrades.com</a>
Healthcare Facilities Accreditation Program (HFAP), <a href="http://www.hfap.org">www.hfap.org</a>
Healthcare Quality Association on Accreditation (HQAA), <a href="http://www.hqaa.org">www.hqaa.org</a>
Intersocietal Accreditation Commission (IAC), <a href="http://www.intersocietal.org">www.intersocietal.org</a>

(continued)

**Table 3** (continued)

---

Joint Commission, [www.jointcommission.org](http://www.jointcommission.org)

---

Leapfrog Group, [www.leapfroggroup.org](http://www.leapfroggroup.org)

---

Medical Travel Quality Alliance (MTQUA), [www.ntqua.org](http://www.ntqua.org)

---

National Business Group on Health (NBGH), [www.businessgrouphealth.org](http://www.businessgrouphealth.org)

---

National Committee for Quality Assurance (NCQA), [www.ncqa.org](http://www.ncqa.org)

---

National Quality Forum (NQF), [www.qualityforum.org](http://www.qualityforum.org)

---

URAC, [www.urac.org](http://www.urac.org)

---

**Associations and professional societies**

---

*Disease/condition associations*

---

ALS Association, [www.alsa.org](http://www.alsa.org)

---

American Association for Cancer Research (AACR), [www.aacr.org](http://www.aacr.org)

---

American Cancer Society (ACA), [www.cancer.org](http://www.cancer.org)

---

American Chronic Pain Association, [www.theacpa.org](http://www.theacpa.org)

---

American Diabetes Association (ADA), [www.diabetes.org](http://www.diabetes.org)

---

American Heart Association (AHA), [www.heart.org](http://www.heart.org)

---

American Stroke Association [www.strokeassociation.org](http://www.strokeassociation.org)

---

American Trauma Society (ATS), [www.amtrauma.org](http://www.amtrauma.org)

---

Canadian Mental Health Association (CMHA), [www.cmha.ca](http://www.cmha.ca)

---

CORD (Canadian Organization for Rare Disorders), [www.raredisorders.ca](http://www.raredisorders.ca)

---

EURORDIS (European Organisation for Rare Diseases), [www.eurordis.org](http://www.eurordis.org)

---

Mental Health America, [www.mentalhealthamerica.net](http://www.mentalhealthamerica.net)

---

National Alliance on Mental Illness (NAMI), [www.nami.org](http://www.nami.org)

---

National Health Council, [www.nationalhealthcouncil.org](http://www.nationalhealthcouncil.org)

---

National Organization for Rare Diseases (NORD), [www.rarediseases.org](http://www.rarediseases.org)

---

NORD (National Organization for Rare Disorders), [www.rarediseases.org](http://www.rarediseases.org)

---

Unite for Sight, [www.uniteforsight.org](http://www.uniteforsight.org)

---

*Demographic and population group associations*

---

AAPD (American Association of People with Disabilities), [www.aapd.com](http://www.aapd.com)

---

AARP, [www.aarp.org](http://www.aarp.org)

---

American Correctional Health Services Association (ACHSA), [www.achsa.org](http://www.achsa.org)

---

National Alliance for Hispanic Health, [www.hispanichealth.org](http://www.hispanichealth.org)

---

National Associations of Counties (NACO), [www.naco.org](http://www.naco.org)

---

National Coalition for the Homeless, [www.nationalhomeless.org](http://www.nationalhomeless.org)

---

National Medical Association (NMA), [www.nmanet.org](http://www.nmanet.org)

---

National Rural Health Association (NRHA), [www.ruralhealthweb.org](http://www.ruralhealthweb.org)

---

NCAI (National Congress of American Indians), [www.ncai.org](http://www.ncai.org)

---

Population Association of America, [www.populationassociation.org](http://www.populationassociation.org)

---

*Health care organizations and trade associations*

---

AAMI (Association for the Advancement of Medical Instrumentation), [www.aami.org](http://www.aami.org)

---

Advanced Medical Technology Association (AdvaMed), [www.advamed.org](http://www.advamed.org)

---

Ambulatory Surgery Center Association, [www.ascassociation.org](http://www.ascassociation.org)

---

American Association of Accountable Care Organizations (AAACO), [www.aaaco.org](http://www.aaaco.org)

---

American Association of Blood Banks (AABB), [www.aabb.org](http://www.aabb.org)

---

American Association of Eye and Ear Centers of Excellence (AAEECE), [www.aeece.org](http://www.aeece.org)

---

American Association of Homes and Services for the Aging (AAHSA), [www.aahsa.org](http://www.aahsa.org)

---

American Association of Preferred Provider Organizations (AAPPO), [www.aappo.org](http://www.aappo.org)

---

American Health Care Association (AHCA), [www.ahca.org](http://www.ahca.org)

---

American Health Information Management Association (AHIMA), [www.ahima.org](http://www.ahima.org)

---

American Hospital Association (AHA), [www.aha.org](http://www.aha.org)

---

(continued)

**Table 3** (continued)

---

Association for Behavioral Health and Wellness (ABHW), <a href="http://www.abhw.org">www.abhw.org</a>
Association of the British Pharmaceutical Industry (ABPI), <a href="http://www.abpi.org.uk">www.abpi.org.uk</a>
Association of Clinical Research Organization (ACRO), <a href="http://www.acrohealth.org">www.acrohealth.org</a>
Catholic Health Association of the United States (CHAUSA), <a href="http://www.chausa.org">www.chausa.org</a>
Children’s Hospital Association, <a href="http://www.childrenshospitals.net">www.childrenshospitals.net</a>
Federation of American Hospitals (FAH), <a href="http://www.fah.org">www.fah.org</a>
HealthCareCAN, <a href="http://www.healthcarecan.ca">www.healthcarecan.ca</a>
HOPE: European Hospital and Healthcare Federation, <a href="http://www.hope.be">www.hope.be</a>
International Hospital Federation (IHF), <a href="http://www.ihf-fih.org">www.ihf-fih.org</a>
Medical Device Manufacturers Association (MDMA), <a href="http://www.medicaldevices.org">www.medicaldevices.org</a>
National Association of ACOs (NAACOS), <a href="http://www.naacos.com">www.naacos.com</a>
National Association of Community Health Centers (NACHC), <a href="http://www.nachc.com">www.nachc.com</a>
National Association for Home Care and Hospice (NAHC), <a href="http://www.nahc.org">www.nahc.org</a>
PhRMA (Pharmaceutical Research and Manufacturers of America), <a href="http://www.phrma.org">www.phrma.org</a>
Trauma Center Association of America, <a href="http://www.traumacenters.org">www.traumacenters.org</a>
UHC (University Health System Consortium), <a href="http://www.uhc.edu">www.uhc.edu</a>
World Medical Association (WMA), <a href="http://www.wma.net">www.wma.net</a>
<i>Professional societies</i>
Academy Health, <a href="http://www.academyhealth.org">www.academyhealth.org</a>
American Academy of Family Physicians (AAFP), <a href="http://www.aafp.org">www.aafp.org</a>
American Academy of Pediatrics (AAP), <a href="http://www.aap.org">www.aap.org</a>
American Academy of Physician Assistants (AAPA), <a href="http://www.aapa.org">www.aapa.org</a>
American Board of Medical Specialties (ABMS), <a href="http://www.abms.org">www.abms.org</a>
American College of Emergency Physicians (ACEP), <a href="http://www.acep.org">www.acep.org</a>
American College of Healthcare Executives (ACHE), <a href="http://www.ache.org">www.ache.org</a>
American College of Surgeons (ACS), <a href="http://www.facs.org">www.facs.org</a>
American College of Radiology (ACR), <a href="http://www.acr.org">www.acr.org</a>
American College of Wound Healing and Tissue Repair (ACWHTR), <a href="https://acwound.org">https://acwound.org</a>
American Dental Association (ADA), <a href="http://www.ada.org">www.ada.org</a>
American Medical Association (AMA), <a href="http://www.ama-assn.org">www.ama-assn.org</a>
American Nurses Association (ANA), <a href="http://www.nursingworld.org">www.nursingworld.org</a>
American Osteopathic Association, <a href="http://www.osteopathic.org">www.osteopathic.org</a>
American Psychiatric Association (APA), <a href="http://www.psychiatry.org">www.psychiatry.org</a>
American Psychological Association (APA), <a href="http://www.apa.org">www.apa.org</a>
American Public Health Association (APHA), <a href="http://www.apha.org">www.apha.org</a>
American Society of Anesthesiologists, <a href="http://www.asahq.org">www.asahq.org</a>
American Society of Health Economists (ASHE), <a href="http://www.healtheconomics.us">www.healtheconomics.us</a>
American Society of Plastic Surgeons (ASPS), <a href="http://www.plasticsurgery.org">www.plasticsurgery.org</a>
Canadian Medical Association (CMA), <a href="http://www.cma.ca">www.cma.ca</a>
European Society for Health and Medical Sociology (ESHMS), <a href="http://www.eshms.eu">www.eshms.eu</a>
Health Services Research Association of Australia and New Zealand, <a href="http://www.hsraanz.org">www.hsraanz.org</a>
International Health Economics Association (iHEA), <a href="http://www.healtheconomics.org">www.healtheconomics.org</a>
National Association of Chronic Disease Directors, <a href="http://www.chronicdisease.org">www.chronicdisease.org</a>
National Association of Medicaid Directors (NAMD), <a href="http://www.medicaiddirectors.org">www.medicaiddirectors.org</a>
National Cancer Registrars Association (NCRA), <a href="http://www.ncra-usa.org">www.ncra-usa.org</a>
National Governors Association (NGA), <a href="http://www.nga.org">www.nga.org</a>
National League for Nursing (NLN), <a href="http://www.nln.org">www.nln.org</a>
Society of General Internal Medicine (SGIM), <a href="http://www.sgim.org">www.sgim.org</a>
Society for Medical Decision Making (SMDM), <a href="http://www.smdm.org">www.smdm.org</a>

---

(continued)

**Table 3** (continued)

<b>Foundations and trusts</b>
Canadian Foundation for Healthcare Improvement, <a href="http://www.cfhi-fcass.ca">www.cfhi-fcass.ca</a>
Commonwealth Fund, <a href="http://www.commonwealthfund.org">www.commonwealthfund.org</a>
Ford Foundation, <a href="http://www.fordfoundation.org">www.fordfoundation.org</a>
Gates (Bill and Melinda) Foundation, <a href="http://www.gatesfoundation.org">www.gatesfoundation.org</a>
Health Research and Educational Trust (HRET), <a href="http://www.hret.org">www.hret.org</a>
Kaiser (Henry J.) Family Foundation, <a href="http://www.kff.org">www.kff.org</a>
Kellogg (WK) Foundation, <a href="http://www.wkkf.org">www.wkkf.org</a>
Kresge Foundation, <a href="http://www.kresge.org">www.kresge.org</a>
MacArthur (John D. and Catherine T.) Foundation, <a href="http://www.macarthur.org">www.macarthur.org</a>
Milbank Memorial Fund, <a href="http://www.milbank.org">www.milbank.org</a>
National Patient Safety Foundation (NPSF), <a href="http://www.npsf.org">www.npsf.org</a>
New America Foundation, <a href="http://www.newamerica.net">www.newamerica.net</a>
NIHCM (National Institute for Health Care Management) Foundation, <a href="http://www.nihcm.org">www.nihcm.org</a>
Pew Charitable Trusts, <a href="http://www.pewtrusts.org">www.pewtrusts.org</a>
Physicians Foundation, <a href="http://www.physiciansfoundation.org">www.physiciansfoundation.org</a>
Pfizer Foundation, <a href="http://www.pfizer.com">www.pfizer.com</a>
Public Health Foundation, <a href="http://www.phf.org">www.phf.org</a>
Robert Wood Johnson Foundation (RWJ), <a href="http://www.rwjf.org">www.rwjf.org</a>
Wellcome Trust, <a href="http://www.wellcome.ac.uk">www.wellcome.ac.uk</a>
<b>Health insurance and employee benefits organizations</b>
American Academy of Insurance Medicine (AAIM), <a href="http://www.aaimedicine.org">www.aaimedicine.org</a>
America's Health Insurance Plans (AHIP), <a href="http://www.ahip.org">www.ahip.org</a>
American Insurance Association (AIA), <a href="http://www.aiadc.org">www.aiadc.org</a>
Association for Community Affiliated Plans (ACAP), <a href="http://www.communityplans.net">www.communityplans.net</a>
Blue Cross and Blue Shield Association (BCBS), <a href="http://www.bcbs.com">www.bcbs.com</a>
Canadian Life and Health Insurance Association (CLHIA), <a href="http://www.clhia.ca">www.clhia.ca</a>
Employee Benefit Research Institute (EBRI), <a href="http://www.ebri.org">www.ebri.org</a>
Healthcare Financial Management Association (HFMA), <a href="http://www.hfma.org">www.hfma.org</a>
Insurance – Canada, <a href="http://www.insurance-canada.ca">www.insurance-canada.ca</a>
Medicaid Health Plans of America (MHPA), <a href="http://www.mhpa.org">www.mhpa.org</a>
National Academy of Social Insurance (NASI), <a href="http://www.nasi.org">www.nasi.org</a>
National Association of Health Underwriters (NAHU), <a href="http://www.nahu.org">www.nahu.org</a>
National Association of Insurance Commissioners (NAIC), <a href="http://www.naic.org">www.naic.org</a>
Physicians for a National Health Program (PNHP), <a href="http://www.pnhp.org">www.pnhp.org</a>
<b>Registries</b>
Alzheimer's Prevention Registry, <a href="http://www.endalznow.org">www.endalznow.org</a>
American Burn Association, National Burn Repository, <a href="http://www.ameriburn.org">www.ameriburn.org</a>
Australian Orthopaedic Association National Joint Replacement Registry (AOANJRR), <a href="http://www.aoa.org.au">www.aoa.org.au</a>
British Society for Rheumatology Rheumatoid Arthritis Register (BSRBR-RA), <a href="http://www.inflammation-repair.manchester.ac.uk">www.inflammation-repair.manchester.ac.uk</a>
Congenital Muscle Disease International Registry (CMDIR), <a href="http://www.cmdir.org">www.cmdir.org</a>
Cystic Fibrosis Foundation (CFF) Patient Registry, <a href="http://www.cff.org">www.cff.org</a>
DANBIO Registry of Biologics Used in Rheumatoid Arthritis Patients, <a href="http://www.danbio-online.dk">www.danbio-online.dk</a>
Danish Hip Arthroplasty Register (DHR), <a href="http://www.kea.au.dk">www.kea.au.dk</a>
EPIRARE (European Platform for Rare Disease Registries), <a href="http://www.epirare.eu">www.epirare.eu</a>
International Society of Heart and Lung Transplantation (ISHLT), <a href="http://www.isHLT.org">www.isHLT.org</a>
Kaiser Permanente Autoimmune Disease Registry, <a href="http://www.kaiserpermanente.org">www.kaiserpermanente.org</a>
NAACCR (North American Association of Central Cancer Registries), <a href="http://www.naacr.org">www.naacr.org</a>
National Cancer Data Base, <a href="http://www.facs.org/quality/programs/cancer/ncdb/">www.facs.org/quality/programs/cancer/ncdb/</a>

(continued)

**Table 3** (continued)

---

National Cardiovascular Data Registry, <a href="http://www.cardiosource.org">www.cardiosource.org</a>
National Marrow Donor Program's Be the Match Registry, <a href="http://www.bethematch.org">www.bethematch.org</a>
National Trauma Data Bank, <a href="http://www.ntdsdictionary.org">www.ntdsdictionary.org</a>
Register of Information and Knowledge about Swedish Heart Intensive-care Admissions (RIKS-HIA), <a href="http://www.ucr.uu.se">www.ucr.uu.se</a>
Scientific Registry of Transplant Recipients (SRTR), <a href="http://www.srtr.org">www.srtr.org</a>
Swedish Childhood Cancer Registry, <a href="http://www.cceg.ki.se">www.cceg.ki.se</a>
Swedish Hip Arthroplasty Register (SHAR), <a href="http://www.shpr.se">www.shpr.se</a>
Swedish National Cataract Register (NCR), <a href="http://www.kataraktreg.se">www.kataraktreg.se</a>
Swedish Rheumatology Quality Register (SRQ), <a href="http://www.srq.nu/en/">www.srq.nu/en/</a>
United Kingdom Cataract National Data Set for Adults, <a href="http://www.rcophth.ac.uk">www.rcophth.ac.uk</a>
United Kingdom Myocardial Ischaemic National Audit Project (MINAP), <a href="http://www.hqip.org.uk">www.hqip.org.uk</a>
United Network for Organ Sharing (UNOS), <a href="http://www.unos.org">www.unos.org</a>
<b>Research and policy organizations</b>
Abt Associates, <a href="http://www.abtassociates.com">www.abtassociates.com</a>
American Enterprise Institute for Public Policy (AEI), <a href="http://www.aei.org">www.aei.org</a>
American Health Policy Institute, <a href="http://www.americanhealthpolicy.org">www.americanhealthpolicy.org</a>
American Research Institute for Policy Development (ARIPD), <a href="http://www.aripd.org">www.aripd.org</a>
Battelle Memorial Institute, <a href="http://www.battelle.org">www.battelle.org</a>
Brookings Institution, <a href="http://www.brookings.edu">www.brookings.edu</a>
Canadian Association for Health Services and Policy Research (CAHSPR), <a href="http://www.cahspr.ca">www.cahspr.ca</a>
Cato Institute, <a href="http://www.cato.org">www.cato.org</a>
Deloitte Center for Health Solutions, <a href="http://www2.deloitte.com">www2.deloitte.com</a>
ECRI Institute, <a href="http://www.ecri.org">www.ecri.org</a>
Families USA, <a href="http://www.familiesusa.org">www.familiesusa.org</a>
Galen Institute, <a href="http://www.galen.org">www.galen.org</a>
George Washington University Center for Health Policy Research, <a href="http://www.publichealth.gwu.edu">www.publichealth.gwu.edu</a>
Institute for Clinical Evaluative Sciences (Ontario, Canada), <a href="http://www.chspr.ubc.ca">www.chspr.ubc.ca</a>
Institute for e-Health Policy, <a href="http://www.e-healthpolicy.org">www.e-healthpolicy.org</a>
Institute for the Future (ITF), <a href="http://www.iftf.org">www.iftf.org</a>
Institute for Healthcare Improvement (IHI), <a href="http://www.ihl.org">www.ihl.org</a>
Institute of Medicine (IOM), <a href="http://www.iom.edu">www.iom.edu</a>
International Health Economics Association (iHEA), <a href="http://www.healtheconomics.org">www.healtheconomics.org</a>
Lewin Group, <a href="http://www.lewin.com">www.lewin.com</a>
Manitoba Centre for Health Policy, <a href="http://www.umanitoba.ca/centres/mchp/">www.umanitoba.ca/centres/mchp/</a>
Mathematica Policy Research, <a href="http://www.mathematica-mpr.com">www.mathematica-mpr.com</a>
McMaster University Centre for Health Economics and Policy Analysis (CHEPA), <a href="http://www.chepa.org">www.chepa.org</a>
National Bureau of Economic Research (NBER), <a href="http://www.nber.org">www.nber.org</a>
National Center for Policy Analysis (NCPA), <a href="http://www.ncpa.org">www.ncpa.org</a>
National Center for Public Policy Research, <a href="http://www.nationalcenter.org">www.nationalcenter.org</a>
National Coalition on Health Care (NCHC), <a href="http://www.nchc.org">www.nchc.org</a>
National Health Policy Forum (NHPF), <a href="http://www.nhpf.org">www.nhpf.org</a>
National Health Policy Group (NHPG), <a href="http://www.nhpg.org">www.nhpg.org</a>
National Institute for Health Care Reform (NIHCR), <a href="http://www.nihcr.org">www.nihcr.org</a>
Nuffield Trust, <a href="http://www.nuffieldtrust.org.uk">www.nuffieldtrust.org.uk</a>
Patient-Centered Research Institute (PCORI), <a href="http://www.pcori.org">www.pcori.org</a>
RAND Corporation, <a href="http://www.rand.org">www.rand.org</a>
RTI International, <a href="http://www.rti.org">www.rti.org</a>
Stanford University Center for Health Policy/Center for Primary Care and Outcomes Research, <a href="http://www.stanford.edu">www.stanford.edu</a>
Transamerica Center for Health Studies, <a href="http://www.transamericacenterforhealthstudies.org">www.transamericacenterforhealthstudies.org</a>

---

(continued)

**Table 3** (continued)

University of British Columbia, Centre for Health Services and Policy Research, <a href="http://www.chspr.ubc.ca">www.chspr.ubc.ca</a>
University of California Los Angeles Center for Health Policy Research, <a href="http://www.healthpolicy.ucla.edu">www.healthpolicy.ucla.edu</a>
University of Illinois at Chicago Institute for Health Research Policy, <a href="http://www.ihrp.uic.edu">www.ihrp.uic.edu</a>
University of Nebraska Center for Health Policy Analysis and Rural Health Research, <a href="http://www.unmc.edu">www.unmc.edu</a>
Urban Institute, <a href="http://www.urban.org">www.urban.org</a>
Westat, <a href="http://www.westat.com">www.westat.com</a>
<b>Survey research organizations</b>
AAPOR (American Association for Public Opinion Research), <a href="http://www.aapor.org">www.aapor.org</a>
AASRO (Association of Academic Survey Research Organizations), <a href="http://www.aasro.org">www.aasro.org</a>
American Statistical Association, Survey Research Methods Section, <a href="http://www.amstat.org/sections/srms">www.amstat.org/sections/srms</a>
CASRO (Council of American Survey Research Organizations), <a href="http://www.casro.org">www.casro.org</a>
ESRA (European Survey Research Association), <a href="http://www.europeansurveyresearch.org">www.europeansurveyresearch.org</a>
Gallup, Inc., <a href="http://www.gallup.com">www.gallup.com</a>
GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany, <a href="http://www.gesis.org/en/institute/">www.gesis.org/en/institute/</a>
Harris Interactive, <a href="http://www.harrisinteractive.com">www.harrisinteractive.com</a>
Institute for Social Research, York University, <a href="http://www.isr.yorku.ca">www.isr.yorku.ca</a>
NORC at the University of Chicago, <a href="http://www.norc.org">www.norc.org</a>
ORC (Opinion Research Corporation) International, <a href="http://www.orcinternational.com">www.orcinternational.com</a>
Population Research Laboratory, University of Alberta, <a href="http://www.ualberta.ca/PRL/">www.ualberta.ca/PRL/</a>
Rasmussen Reports, <a href="http://www.rasmussenreports.com">www.rasmussenreports.com</a>
Roper Center for Public Opinion Research, University of Connecticut, <a href="http://www.ropercenter.uconn.edu">www.ropercenter.uconn.edu</a>
Survey Health Care, <a href="http://www.surveyhealthcare.com">www.surveyhealthcare.com</a>
Survey Research Laboratory, University of Illinois at Chicago, <a href="http://www.srl.uic.edu">www.srl.uic.edu</a>
University of Virginia Center for Survey Research, <a href="http://www.virginia.edu/surveys/moreinfo.htm">www.virginia.edu/surveys/moreinfo.htm</a>

surveys, and it has also worked closely with state governments to gather vital records. The NCHS currently has four major data collection programs: National Vital Statistics System (NVSS); National Health Interview Survey (NHIS); National Health and Nutrition Examination Survey (NHANES); and the National Health Care Surveys. A summary of NCHS' surveys and data collection systems can be found at [www.cdc.gov/nchs/data/factsheets/factsheet\\_summary2.pdf](http://www.cdc.gov/nchs/data/factsheets/factsheet_summary2.pdf).

An important source of health care reports and issue briefs is the US Library of Congress' (LOC) Congressional Research Service (CRS). The CRS, which is the research arm of the US Congress, conducts its analysis at the requests of Congressional committees and individual members of both the House and Senate. CRS reports are unbiased, timely, and comprehensive. Recent reports have addressed the various aspects of the Affordable Care Act (ACA), the most extensive reform of the American health care system in the last 50 years. CRS' reports are posted on its website, [www.loc.gov/crsinfo/](http://www.loc.gov/crsinfo/).

### US Registries

The federal government conducts, and/or sponsors, a number of health care registries. For example, the National Cancer Institute's (NCI) Surveillance, Epidemiology, and End Results (SEER) Program is the nation's premier source of cancer statistics. The SEER Program is a coordinated system of population-based cancer registries located across the United States. It currently collects cancer incidence and survival data from 20 geographic areas of the nation, which together represent about 28 % of the US population. The SEER Program provides essential data to track the nation's medical progress against cancer. It also enables researchers to study access, quality, and outcomes of health care, geographic patterns of cancer care, and health disparities. SEER's data are widely available through factsheets, reports, databases, analytical software, websites, and linkages to other national data sources (National Cancer Institute 2010). More information can be obtained at [www.seer.cancer.gov](http://www.seer.cancer.gov).

The US Food and Drug Administration (FDA) also has many registries that may be useful to health services researchers. There are registries that identify all US registered drugs and medical devices and their manufacturers; record the occurrence of adverse drug events and medication errors; and list drugs in short supply and the reasons for the drug shortages. The FDA also has a registry of all new and generic drug approvals. More information on these and other FDA registries can be found at [www.fda.gov](http://www.fda.gov).

### US Government Agencies and Departments

All of the 15 executive departments of the US federal government are involved in some way with collecting health care data. For example, the US Department of Labor (DOL) collects data on the nation's health care workers and makes projections of future needs; the US Department of Defense (DOD) records the health care services they provide to military personnel and their families at bases throughout the world; and the US Department of Homeland Security (DHS) works and collects data on the nation's hospitals and public health departments to prepare for natural disasters and possible terrorist attacks.

The US Department of Health and Human Services (HHS) is by far the largest and arguably the most important collector of health care data of all. The HHS, with a budget of \$1 trillion in fiscal year (FY) 2015, has many staff offices and operating divisions, which implement national health care policy, manage health care programs, deliver health care services, conduct medical research, and collect health care data. The major data collection systems sponsored by HHS, by agency and division, are listed in Table 4 (Office of the Assistant Secretary for Planning and Evaluation *n.d.*):

### Health Programs and Systems of Other (Non-U.S.) Nations

Some nations provide data and information about their health care programs and systems, which may be useful for health services research. This information can be quickly obtained via the Internet. Some of the best and most accessible English-

**Table 4** List of major U.S. Department of Health and Human Services data sources by division/agency

<b>Agency for Healthcare Research and Quality (AHRQ)</b>
Healthcare Cost and Utilization Project (HCUP)
Medical Expenditure Panel Survey (MEPS)
<b>Centers for Disease Control and Prevention (CDC)</b>
Behavioral Risk Factor Surveillance System (BRFSS)
National Ambulatory Medical Care Survey (NAMCS)
National Health Interview Survey (NHIS)
National Health and Nutrition Examination Survey (NHANES)
National Home and Hospice Care Survey (NHHCS)
National Hospital Ambulatory Medical Care Survey (NHAMCS)
National Hospital Care Survey (NHCS)
National Hospital Discharge Survey (NHDS)
National Immunization Survey (NIS)
National Nursing Home Survey (NNHS)
National Survey of Children's Health (NSCH)
National Survey of Family Growth (NSFG)
National Survey of Residential Care Facilities (NSRCF)
National Vital Statistics System (NVSS)
State and Local Area Integrated Telephone Survey (SLAITS)
Youth Risk Behavior Surveillance System (YRBSS)
<b>Centers for Medicare and Medicaid Services (CMS)</b>
CMS Administrative Datasets
Home Health Outcome and Assessment Information Set (OASIS)
Medicare Current Beneficiary Survey (MCBS)
<b>Health Resources and Services Administration (HRSA)</b>
Area Health Resource File (AHRF)
<b>National Institutes of Health (NIH)</b>
Health and Retirement Study (HRS)
National Children's Study (NCS)
<b>Substance Abuse and Mental Health Services Administration (SAMHSA)</b>
National Survey on Drug Use and Health (NSDUH)

language websites are: United Kingdom's National Health Service (NHS) at [www.nhs.uk](http://www.nhs.uk), including its the National Institute for Health and Care Excellence (NICE) [www.nice.org.uk](http://www.nice.org.uk); Health Canada [www.hc-sc.gc.ca](http://www.hc-sc.gc.ca), and the Canadian Institute of Health Research [www.cihr-irsc.gc.ca](http://www.cihr-irsc.gc.ca); and Australian Government Department of Human Services, [www.humanservices.gov.au](http://www.humanservices.gov.au).

## Government Sponsored International Organizations

Nearly all of the world's nations are members of international organizations that address health care policy issues. For example, the World Health Organization (WHO), which is the directing and coordinating authority for health within the United Nations (UN), represents 194 member states (nations). Established in 1946 and headquartered in Genève, Switzerland, the WHO provides leadership on health matters worldwide, and it sets norms and standards on health issues. Much of the WHO's work is concentrated on supporting research and providing technical advice to the health departments and ministries of governments. The WHO compiles health statistics on its members and publishes numerous reports in print and on the Internet (Lee 2008). Its website can be accessed at [www.who.int](http://www.who.int).

Another important international health organization is the Pan American Health Organization (PAHO). Founded in 1902 and headquartered in Washington, D.C., PAHO is the oldest international public health agency. PAHO provides technical cooperation and mobilizes partnerships to improve the health and quality of life in the nations of the Americas. It represents 50 nations and territories in North and South America and also serves as the regional office for the Americas of the World Health Organization. PAHO collects health statistics on its members, publishes reports, and posts them on the Internet at [www.paho.org](http://www.paho.org).

Other government sponsored international organizations that collect health statistics include: European Commission (EC), [www.ec.europa.eu](http://www.ec.europa.eu); Organisation for Economic Cooperation and Development (OECD), [www.oecd.org](http://www.oecd.org); United Nations (UN), [www.un.org](http://www.un.org); and the World Bank, [www.worldbank.org](http://www.worldbank.org).

## Private Organizations

Private data collecting and holding organizations include health information clearinghouses and libraries; accreditation, evaluation, and regulatory organizations; associations and professional societies; foundations and trusts; health insurance and

employee benefits organizations; registries; research and policy organizations; and survey research organizations.

## Health Information Clearinghouses and Libraries

There are a number of private health information clearinghouses and libraries. Examples include the Cochrane Collaboration and the Dartmouth Atlas of Health Care Project, which have previously been discussed. Two other examples are IMS Health and the Inter-University Consortium of Political and Social Research (ICPSR).

IMS Health is a large global information and technology services for-profit corporation, whose stock is traded on the New York Stock Exchange. Established in 1954, and headquartered in Danbury, Connecticut, IMS Health operates in more than 100 countries. It maintains several very large databases (more than 10 petabytes) on various diseases, treatments, costs, and outcomes of care. Using annual data from 100,000 suppliers, which include physicians who report on the number and type of drug prescriptions they write, and more than 55 billion health care transactions, the company serves over 5,000 clients globally. IMS Health customers include health care manufacturers, medical providers, government agencies, policymakers, and researchers. More information on the company can be obtained at [www.imshealth.com](http://www.imshealth.com).

The Inter-University Consortium of Political and Social Research (ICPSR) is a unit of the Institute for Social Research at the University of Michigan with offices in Ann Arbor. Established in 1962, the ICPSR acquires, preserves, and distributes original social science research data to an international consortium of to more than 700 university and research institution members. ICPSR maintains a very large data archive of more than 500,000 research files. These data span many academic disciplines including economics, sociology, political science, demography, gerontology, and public health. It has several special topic collections that relate to health services research: health and medical care archive; minority data resource center; national archive of computerized data on aging; substance abuse and mental health data



archive; and the terrorism and preparedness data resource center. Faculty, staff, and students of member institutions have full access to ICPSR's data archives and to all of its services. Data files are available in SAS, SPSS, Strata, and R format. ICPSR's website is [www.icpsr.umich.edu](http://www.icpsr.umich.edu).

### **Accreditation, Evaluation, and Regulatory Organizations**

To ensure that patients receive safe high quality care, health care professionals, laboratories, programs, and health care facilities are accredited and regulated.

One of the most important accrediting organizations is the Joint Commission. Founded in 1951, the Joint Commission, an independent, not-for-profit organization, is the largest and oldest accrediting health care organization in the USA. It accredits and certifies more than 20,500 health care organizations and programs in the nation including: all types of hospitals; home care organizations, medical equipment services, pharmacy, and hospice services; nursing homes and rehabilitation centers; behavioral health care and addiction services; ambulatory care organizations, group practices and office-based surgery practices; and independent and freestanding clinical laboratories.

To receive Joint Commission accreditation, hospitals, for example, must meet certain evidence-based process standards that are closely linked to positive patient outcomes. These process or accountability measures include heart attack care, pneumonia care, surgical care, children's asthma care, inpatient psychiatric services, venous thromboembolism care, stroke care, immunization, and perinatal care (Chassin et al 2010). The Joint Commission grants accreditation based on periodic reviews by its survey teams who conduct unannounced onsite visits, and quarterly self-assessment reports submitted by the hospitals. The quality and safety results for specific hospitals are available at [www.qualitycheck.org](http://www.qualitycheck.org).

A more recently established popular health care evaluation organization is the Healthgrades Operating Company, which is simply known as Healthgrades. Founded in 1998 in Denver,

Colorado, Healthgrades has amassed data on over three million US health care providers. Healthgrades provides online data to consumers on physicians, hospitals, and dentists. For example, the website identifies the name of physicians in a city or zip code, the conditions they treat, the procedures they perform, the physician's qualifications and patient feedback, and other criteria. In terms of qualifications, the site identifies whether the physician is board certified and has sanctions or malpractice claims against them, the report of eight measures of care, and their patient's willingness to recommend the physician to their family and friends. Today, nearly one million people a day use the Healthgrades website. It should be noted that some physicians have criticized Healthgrades for having erroneous data, and not screening for false reviews. Healthgrades website is [www.healthgrades.com](http://www.healthgrades.com).

### **Associations and Professional Societies**

The largest category of private health care organizations is associations and professional societies. This category, which includes hundreds of organizations (Swartout 2014), can be roughly subdivided into disease/condition associations, demographic and population group associations, health care organizations and trade associations, and professional societies.

There are associations for nearly every disease and medical condition. These associations help individuals and their families suffering from the disease, and they advocate on their behalf, educate the general public, and work to prevent and end the disease. An example of this type of association is the American Cancer Society (ACS). Founded in 1913, the American Cancer Society is one of the largest voluntary health organizations in the USA. With its headquarters in Atlanta, Georgia, the ACS also has over 350 local offices nationwide. The ACS works to prevent cancer and detect it as early as possible. The society offers free information, programs, and services, and it provides community referrals to patients, survivors, and caregivers. It funds research to identify the causes of cancer, to determine the best way to prevent cancer, and to discover new ways to cure the disease. It also works with lawmakers to

promote policies, laws, and regulations to prevent cancer. The ACS has a National Cancer Information Center, which is open 24 h a day, every day of the year, to answer questions from individuals. And it also offers advice online. The ACS website is [www.cancer.org](http://www.cancer.org).

Some associations represent specific demographic and population groups. For example, the National Rural Health Association (NRHA) works on behalf of the rural population of the USA. Nearly 25 % of the nation's population lives in rural areas and many of them, who tend to be poorer, have higher suicide rates and experience higher death and serious injury accidents than their urban counterparts, also face physician shortages and have to travel long distances to health facilities. The NRHA works to improve the health and well-being of rural Americans. It provides leadership on health issues through advocacy, communications, education, and research. Founded in 1980, with headquarters in Leawood, Kansas, the NRHA has more than 21,000 individual and organizational members, all sharing a common interest in rural health. Its website is [www.ruralhealthweb.org](http://www.ruralhealthweb.org).

Other associations represent health care organizations and trade associations. The Pharmaceutical Research and Manufacturers of America (PhRMA) is an example of a large influential trade association. Founded in 1958, and headquartered in Washington, D.C., PhRMA represents the nation's largest biopharmaceutical research and biotechnology companies, such as Amgen, Bayer, Eli Lilly, Merck, and Pfizer. Since 2000, PhRMA member companies have invested more than \$550 billion in drug development, including an estimated \$51.1 billion in 2013. PhRMA is an advocate for public policies to encourage the discovery of new medicines. To accomplish this PhRMA is dedicated to achieve: broad patient access to medicines through a free market, without price controls; strong intellectual property incentives; and effective regulation and a free flow of information to patients. PhRMA publishes policy papers, profiles and reports, fact sheets, newsletters, and speeches (Pharmaceutical Research and Manufacturers of America 2014). These publications are available at its website, [www.phrma.org](http://www.phrma.org).

Another type of health care association is professional societies. These societies advocate and lobby for their members, provide continuing education, and attempt to advance the field. They typically publish newsletters, factsheets, journals, and hold local meetings and an annual convention for their members.

For example, one of the oldest professional medical societies is the American Medical Association (AMA). Founded in 1847, and incorporated in 1897, the AMA is the largest association of physicians and medical students in the USA. Starting as a small association, the AMA would become the single most influential organization on the practice of medicine in the nation. The AMA gained national prominence by publishing its flagship *Journal of the American Medical Association* and by reorganizing into local and state-level constituent societies, a national House of Delegates, a Board of Trustees, and national officers. With these changes, the membership of the AMA grew from around 8,000 in 1900 to approximately 220,000 today. During the 1960s, the membership market share of the AMA reached its zenith, representing about 70 % of the nation's physicians, but today it only represents about 25 %. Its membership, and to some degree its influence, has declined because of the profusion of competing national specialty medical societies, and the decline of solo practices and the rise of salaried physicians who work for various organizations (American Medical Association 1997).

Today, the stated mission of the AMA is to promote the art and science of medicine and the betterment of public health. Headquartered in Chicago and with an office in Washington, D.C., the AMA advocates for its members by developing health care policies. The top items on the AMA's current policy agenda include modification of the Affordable Care Act (ACA), the improvement of diabetes care delivery, changes in drug reporting, and increasing Medicaid payments making them comparable to those paid by Medicare. The AMA also produces a number of important products and services. The association is one of the largest publishers of medical information in the world. For example, its weekly *Journal of the American Medical Association*

(JAMA) is published in 10 languages and print editions are circulated in over 100 countries. The AMA also publishes a number of other specialty journals. Another of its publications is the *Current Procedural Terminology* (CPT), a guidebook for physicians' offices on how to classify and code medical procedures and services for reimbursement from Medicare, and other insurance companies. An important AMA resource that supports membership services, marketing activities, and research is the Physician Master file, a large database that contains biographic, medical education and training, contact, and practice information on more than 1.4 million physicians, residents, and medical students in the USA. The file, which is updated continuously, also contains information on medical schools, graduate medical education programs, teaching institutions, and medical group practices. More information on the AMA can be found on its website at [www.ama-assn.org](http://www.ama-assn.org).

### Foundations and Trusts

Foundations and trusts are an important source of health care data and information. They publish health care reports, policy briefs, and newsletters. They also often support and fund projects on various health services research topics.

One of the largest health care foundations in the USA is the Robert Wood Johnson Foundation (RWJF). Located in Princeton, New Jersey, RWJF was founded in 1936, and it became a national philanthropy in 1972. Its mission is to improve the health and health care of all Americans. Over the past 40 years, RWJF has become the nation's largest philanthropy devoted solely to the public's health. It currently provides grant funds primarily to public agencies, universities, and public charities in six broad areas of focus: child and family well-being; childhood obesity, health insurance coverage, healthy communities, health leadership and the workplace, and health system improvement. RWJF attempts to fund innovative projects that will have a measurable impact and that can create meaningful, transformation change, such as service demonstrations, gathering and monitoring of health statistics, public education, training and fellowship programs, policy analysis, health services research, technical assistance,

communications activities, and evaluations. RWJF funds both projects it proposes that are issued through call for proposals (CFP) as well as unsolicited proposals. Each year, RWJF makes hundreds of awards, with funds ranging from \$3,000 to \$23 million. However, most awards range from \$100,000 to \$300,000 for a period of 1–3 years. A list of 1,225 awards RWJF has given from 1972 to 2015 totaling \$789,305, 241 can be found at [www.rwjf.org](http://www.rwjf.org).

### Health Insurance and Employee Benefits Organizations

Many health services researchers study the function of health insurance, the various types of insurance plans, and the impact of insurance on the use of health care services. In the USA, most people obtain their health insurance coverage through their workplace, with health insurance being one of the most important employment-based benefits.

An example of one organization that studies health insurance and other benefits is the Employee Benefit Research Institute (EBRI). Founded in 1978, and located in Washington, D.C., the EBRI is a nonprofit, nonpartisan organization that conducts research relating to employee benefit plans, compiles and disseminates information on employee benefits, and sponsors educational activities such as lectures, roundtables, forums, and study groups on employee benefit plans. The EBRI publishes a number of special reports, books, and monthly issue briefs. It also conducts annual health and retirement benefit surveys. In terms of health benefits, the EBRI has four research centers: Center for Research on Health Benefits Innovation, which focuses on helping employers measure the impact of new benefit plan designs in terms of cost, quality, and access to health care; Center for Studying Health Coverage and Public Policy, which monitors the trends in the availability of health coverage and the impact of public policy on employment-based health benefits; Center for Research on Health Care in Retirement, which studies the trends in retiree health benefits and its impact upon them; and the Center for Survey Research, which conducts the Health Confidence Survey and the Consumer Engagement in Health Care Survey. More information can be found at [www.ebri.org](http://www.ebri.org).

## Registries

This category is very similar to government registries, which has been discussed previously. However, it differs mainly in terms of sponsorship and funding source. Most of the private organization registries are funded, constructed, and maintained by professional medical societies.

For example, the National Trauma Data Bank (NTDB), which is managed by the American College of Surgeons' (ACS) Committee on Trauma, is the largest aggregation of trauma patient data in the USA. The NTDB obtains its data from trauma registries maintained by hundreds of hospitals across the nation. Currently, the NTDB contains more than five million trauma patient records. Since 2003, the NTDB has published an annual report summarizing these data. The 2013 report contains data based on 833,311 trauma patient records submitted by 805 hospitals. The report contains a wealth of summary information on the patient's age; gender; primary payment source; alcohol and drug use; mechanism of injury; injury severity score; pre-hospital time; hospital geographic location, bed size, and trauma level; patient transfer information; number of ICU and ventilator days; hospital complications; length of hospital stay; place of discharge; and the number of deaths, including those who were dead on arrival (DOA), and specific and overall mortality rates (American College of Surgeons 2013). Data contained in the NTDB are available to qualified researchers in two forms: a dataset containing all records sent to the NTDB for each admission year and national estimates for adult patients seen in Level I and II trauma centers. More information is available at the NTDB website at [www.ntdb.org](http://www.ntdb.org).

## Research and Policy Organizations

There are many private research and policy organizations in the USA. Most of them conduct contract or grant funded research studies for the federal government. Depending upon the scope of work, these contracts and grants can amount to hundreds of thousands to hundreds of millions of dollars. Some of the largest contract research and policy organizations that often receive these funds include Abt Associates, Brookings Institution,

Mathematica Policy Research, RAND Corporation, RTI International, and the Urban Institute.

The RAND Corporation is the largest, and one of the most prestigious, research and policy organizations in the USA. Incorporated in 1948, the RAND (a contraction of "research and development") Corporation is an independent, nonprofit organization that conducts research and analysis for many US government departments, foreign governments, international organizations, professional associations, and other organizations. Headquartered in Santa Monica, California, the RAND Corporation has a professional staff of 1,700 people. It annually receives over \$250 million in contracts and grants and works on about 500 projects at any given time.

One of the RAND Corporation's largest research divisions is RAND Health. With a staff of 280 health care experts, about 70 % of RAND Health's research is supported by contracts and grants from the US federal government, with the remainder coming from professional associations, universities, state and local governments, and foundations.

Over the years, RAND Health has conducted hundreds of health care research studies, including the very famous Health Insurance Experiment (HIE). The RAND HIE was one of the largest and most comprehensive social science experiments ever conducted in the USA. Funded by the federal government, HIE addressed two key questions: How much more medical care will people use if it is provided free of charge? What are the consequences for their health? To answer these questions in 1971 the HIE randomly assigned several 1,000 households in different geographic regions in the USA to health insurance with varying levels of co-insurance, and then followed them for 5-years to evaluate the effect on their medical utilization and health. The HIE, which took 15-year to complete at a cost of about \$200 million, remains the largest health policy study ever conducted in US history. Its rich findings are still being discussed today (Aron-Dine et al. 2013; Newhouse 1993). More information about the HIE can be found on the project's home page at: [www.rand.org/health/projects/hie.html](http://www.rand.org/health/projects/hie.html).

Currently, RAND Health is working on several major projects including developing global

HIV/AIDS prevention strategies using antiretroviral drugs in South Africa, India, and the USA; measuring the total costs of dementia in the USA; determining the impact of lowering the costs of healthy foods in supermarkets in the diet patterns of households in South African; identifying the effect of the Affordable Care Act (ACA) on hospital emergency department use by young adults who remained on their parent's health insurance; and developing new models of patient-centered medical homes and nurse-managed health centers to help alleviate the growing shortage of primary care physicians in the USA (RAND Corporation 2013).

The RAND Corporation publishes all of its reports on its website. Further, RAND Health makes all of its surveys publicly available without charge. Examples of available surveys by topic are shown in Table 5. More information can be found at [www.rand.org/health/surveys\\_tools.html](http://www.rand.org/health/surveys_tools.html).

### Survey Research Organizations

Academic and commercial survey research organizations frequently collect health care data. They often conduct health care surveys for various government agencies, commercial companies, and research and public policies organizations. Sometimes they also add health care questions to the general population surveys they conduct to determine changing attitudes, beliefs, and public opinions. Data from these surveys are often archived by the survey organizations and eventually are made available to researchers. Many of these organizations also provide lists of the survey questions they have used. This can be a valuable resource for researchers, because it is difficult to design nonbiased questions, and they can judge the validity and reliability of the questions already used. Researchers may include these questions in the surveys they are designing.

An example of one of the oldest independent academic-based survey research organizations is NORC at the University of Chicago. Founded in 1941, NORC, which originally stood for National Opinion Research Center, is headquartered in downtown Chicago with additional offices on the University of Chicago's campus and in Washington, D.C. During the past 70 years, NORC has conducted many landmark national

**Table 5** RAND Corporation health surveys by topic

<b>Aging and health</b>
Assessing Care of Vulnerable Elders (ACOVE)
Vulnerable Elders Survey (VES-13)
<b>Diversity and health</b>
Homelessness survey
<b>Health economics</b>
Hospital competition measures
Managed health care survey
<b>HIV, STDs, and sexual behavior</b>
HIV Cost and Services Utilization Study (HCSUS)
HIV Identification, Prevention, and Treatment Services Surveys
HIV Patient-Assessed Report of Status and Experience (HIV-PARSE)
<b>Maternal, child, and adolescent health</b>
Pediatric Asthma Symptom Scale
Pediatric Quality of Life Inventory (PedsQL Measurement Model)
<b>Mental health</b>
Mental health inventory
Depression screener
Improving Care for Depression in Primary Care (Partners in Care)
<b>Military health policy</b>
Chronic Illness Care Evaluation Instruments (ICICE website)
Dialysis Patient Satisfaction Survey (DPSS)
Patient Satisfaction Questionnaires (PSQ-III and PSQ-18)
Patient Satisfaction Survey for the Unified Medical Group Association
<b>Quality of life</b>
Epilepsy Surgery Inventory Survey (ESI-55)
Kidney Disease Quality of Life Instrument (KDQOL)
Medical Outcomes Study (MOS)
Measures of quality of life
Measures of patient adherence
Mental health inventory
Sexual problems measures
Sleep scale
Social support survey
National Eye Institute Refractive Error Quality of Life Instrument
Pediatric Quality of Life Inventory (PedsQL Measurement Model)
Quality of Life in Epilepsy Inventory (QOLIE-89 and QOLIE-31)
RAND Negative Impact of Asthma on Quality of Life
Visual Function Questionnaire (VFQ-25)
<b>Research methods</b>
Socially Desirable Response Set Five-Item Survey (SDRS-5)
The Homelessness Survey

large-scale health surveys including: National Ambulatory Medical Care Survey, the first-ever survey of medical care delivered to patients by office-based physicians; National Children's Study, the largest study of children's health and development tracking 100,000 children before birth through age 21; and the National Social Life, Health and Aging Project, a longitudinal study of the health of older Americans.

One of NORC's flagship surveys and longest-running projects is its General Social Survey (GSS). Begun in 1972, and continuing today, this annual survey is the most widely regarded single best source of data on societal trends. Hundreds of researchers, policymakers, and students have used the survey's data to study a wealth of topics. The GSS contains a standard set of demographic, behavioral, and attitudinal questions, plus various topics of special interest. For more than 40 years the GSS has been tracking the opinions of Americans. Over the years, many health care questions have been included in the survey asking about choice of physicians, difficulty receiving care, health insurance coverage, coverage changes, use of Medicare/Medicaid, incentives for physicians, opinions on HMOs, and whether they sought medical care for mental health problems. Data from the GSS and its various questionnaires and codebooks can be downloaded. A cross-tabulation program is also available (NORC at the University of Chicago 2011). More information on NORC and its surveys, including the GSS, can be obtained at [www.norc.org](http://www.norc.org).

---

## Conclusion

This chapter has presented a practical typology of health care data, and it has identified and described many important data sources and public use files. Although much health care data are currently available, in the future much more data will be needed. The demand for more accessible, transparent, and comprehensive health care data will be driven by advances in medical science, rising public expectations, the continuing growth of the Internet and social media, and the ever increasing cost of health care.

In the future, patients, health care providers, policymakers, and health services researchers

will all increasingly demand having more health care data. Patients will need these data to help them make better evidence-based informed decisions. They need to know: Who are the best physicians for the care I need? What innovative treatments are available? What are the benefits and risks of the treatments? Which hospitals are the best providers of the treatments? Where can I get a second or even a third medical opinion? How much will the treatments cost? And which treatments are covered by my current health insurance policy?

Health care providers will need more data to better monitor the care they provide. They will need to hold down their costs, provide high quality services, and justify what they charge to health care insurers. They also will have to increasingly deal with patients demanding more data on the cost and quality of the care they received. Already many hospitals and clinics, insurers, and employers enable patients to access their electronic medical and billing records online.

Policymakers will need more data to develop new more effective policies to help bend the cost curve. They will use these data to construct and test new medical care reimbursement models, which will hopefully lower costs and at the same time increase the quality of care. They will also develop policies to encourage more disease prevention and wellness programs.

Health services researchers will demand more data to better evaluate existing health care programs. They will increasingly conduct research to compare the relationship between the cost and quality of health care to determine its value to patients and society. Over time, using these data sources, health services researchers will forge an important new evidence-based science of health care delivery – a new science that will continue to build on the crucial concepts of access, cost, quality, and the outcome of health care.

---

## References

- Aaron HJ, Schwartz WB, Cox MA. Can we say no?: the challenge of rationing health care. Washington, DC: Brookings Institution Press; 2005.

- Aday LA, Cornelius LJ. Designing and conducting health surveys: a comprehensive guide. 3rd ed. San Francisco: Jossey-Bass; 2006.
- Aday LA, Begley CE, Lairson DR, Balkrishnan R. Evaluating the healthcare system: effectiveness, efficiency, and equity. 3rd ed. Chicago: Health Administration Press; 2004.
- Agency for Healthcare Research and Quality (AHRQ). 2013 National healthcare disparities report. Rockville: Agency for Healthcare Research and Quality; 2014a. Available at: [www.ahrq.gov/research/findings/nhqrd/index.html](http://www.ahrq.gov/research/findings/nhqrd/index.html)
- Agency for Healthcare Research and Quality (AHRQ). 2013 National healthcare quality report. Rockville: Agency for Healthcare Research and Quality; 2014b. Available at: [www.ahrq.gov/research/findings/nhqrd/index.html](http://www.ahrq.gov/research/findings/nhqrd/index.html)
- American College of Surgeons. National Trauma Data Bank 2013: annual report. Chicago: American College of Surgeons; 2013. Available at: [www.ntdb.org](http://www.ntdb.org)
- American Hospital Association (AHA). AHA guide to the health care field, 2014. Chicago: Health Forum; 2013a.
- American Hospital Association (AHA). AHA hospital statistics, 2014. Chicago: Health Forum; 2013b.
- American Hospital Association (AHA). AHA Data Viewer website. 2015. [www.ahadataviewer.com](http://www.ahadataviewer.com)
- American Medical Association. Caring for the country: a history and celebration of the first 150 years of the American Medical Association. Chicago: American Medical Association; 1997.
- Andersen RM. Revisiting the behavioral model and access to medical care: does it matter? *J Health Soc Behav.* 1995; 36:1–10. Available at: [www.mph.ufl.edu/files/2012/01/session6april2RevisitingBehavioralModel.pdf](http://www.mph.ufl.edu/files/2012/01/session6april2RevisitingBehavioralModel.pdf)
- Anderson C. Multinational comparisons of health system data, 2014. New York: Commonwealth Fund; 2014. Available at: [www.commonwealthfund.org](http://www.commonwealthfund.org)
- Anderson GF, Squires DA. Measuring the U.S. health care system: a cross-national comparison. *Issues in International Health Policy*, Commonwealth Fund, Pub. 1412, Vol. 90, June 2010. Available at: [www.commonwealthfund.org](http://www.commonwealthfund.org)
- Aron-Dine A, Einav L, Finkelstein A. The RAND Health Insurance Experiment, three decades later. *J Econ Perspect.* 2013;27(1):197–222. Available at: <http://economics.mit.edu/files/8400>
- Arts DGT, de Keizer NF, Scheffer G-J. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc.* 2002;9(6):600. Available at: [www.ncbi.nlm.nih.gov/pmc/articles/PMC349377](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC349377)
- Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ.* 1996;312(7040):1215–18. Available at: [www.bmj.com/content/312/7040/1215](http://www.bmj.com/content/312/7040/1215)
- Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. Chichester: Wiley; 2009.
- Centers for Medicare and Medicaid Services (CMS). Medicare and you, 2015. Baltimore: Centers for Medicare and Medicaid Services; 2014a. Available at: [www.cms.gov](http://www.cms.gov)
- Centers for Medicare and Medicaid Services (CMS). Nursing home data compendium 2013 edition. Baltimore: Centers for Medicare and Medicaid Services; 2014b. Available at: [www.cms.gov](http://www.cms.gov)
- Chassin MR, Loeb JM, Schmalz SP, Wachter RW. Accountability measures – using measurement to promote quality improvement. *N Engl J Med.* 2010;363(7):683–88. Available at: [www.nejm.org/doi/full/10.1056/NEJMs1002320](http://www.nejm.org/doi/full/10.1056/NEJMs1002320)
- Children's Health Care Quality Measures Core Set Technical Assistance and Analytic Program. Strategies for using vital records to measure quality of care in Medicaid and CHIP programs. Medicaid/CHIP Health Care Quality Measures: Technical Assistance Brief 4: Jan 2014, 1–11. Available at: [www.medicaid.gov/Medicaid-CHIP-Program-information/By-Topics/Quality-of-Care/Downloads/Using-Vital-Records.pdf](http://www.medicaid.gov/Medicaid-CHIP-Program-information/By-Topics/Quality-of-Care/Downloads/Using-Vital-Records.pdf)
- Clancy CM. What is health care quality and who decides? Statement before the Committee on Finance, Subcommittee on Health Care, U.S. Senate, 18 Mar 2009. Available at: [www.hhs.gov/asl/testify/2009/03/t20090318b.html](http://www.hhs.gov/asl/testify/2009/03/t20090318b.html)
- Culyer AJ. The dictionary of health economics, second edition. Northampton: Edward Elgar; 2010.
- Dartmouth Atlas of Health Care Project. 2015. [www.dartmouthatlas.org](http://www.dartmouthatlas.org)
- Davis KE. Access to health care of adult men and women, ages 18–64, 2012. Medical Expenditure Panel Survey (MEPS) Statistical Brief #461. Rockville: U.S. Agency for Healthcare Research and Quality (AHRQ); 2014. Available at: [www.meps.ahrq.gov/mepsweb/data\\_files/publications/st461/stat461.pdf](http://www.meps.ahrq.gov/mepsweb/data_files/publications/st461/stat461.pdf)
- Davis K, Stremikis K, Squires D, Schoen C. Mirror, mirror on the wall: how the performance of the U.S. health care system compares internationally. Pub. No. 1755. New York: Commonwealth Fund; 2014. Available at: [www.commonwealthfund.org](http://www.commonwealthfund.org)
- Donabedian A. The definition of quality and approaches to its assessment. Vol. 1. Explorations in quality assessment and monitoring. Ann Arbor: Health Administration Press; 1980.
- Feldstein PJ. Health care economics. 7th ed. New York: Thomson Deimar Learning; 2011.
- Gliklich RE, Dreyer NA, editors. Registries for evaluating patient outcomes: a user's guide. 2nd ed. AHRQ Publication No. 10-EHC049. Rockville: U.S. Agency for Healthcare Research and Quality; 2010. p. 15–16. Available at: [www.effectivehealthcare.ahrq.gov/ehc/products/74/531/Registries2nd ed Final to Eisenberg 9-15-10.pdf](http://www.effectivehealthcare.ahrq.gov/ehc/products/74/531/Registries2nd%20ed%20Final%20to%20Eisenberg%209-15-10.pdf)
- Halsey MF, Albanese SA, Thacker M, The Project of the POSNA Practice Management Committee. Patient satisfaction surveys: an evaluation of POSNA members' knowledge and experience. *J Pediatr Orthop.* 2015; 35(1):104–7.

- Health Care Cost Institute (HCCI). 2013 health care cost and utilization report. Washington, DC: Health Care Cost Institute; 2014. Available at: [www.healthcostinstitute.org](http://www.healthcostinstitute.org)
- Healthy People 2020. [www.healthypeople.gov](http://www.healthypeople.gov)
- Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. Chichester: Wiley-Blackwell; 2008.
- Huston P, Naylor CD. Health services research: reporting on studies using secondary data sources. *Can Med Assoc J.* 1996;155(12):1697–1702. Available at: [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)
- Johnson TP, editor. Handbook of health survey methods. New York: Wiley; 2014.
- Kane RL, Radosevich DM. Conducting health outcomes research. Sudbury: Jones and Bartlett Learning; 2011.
- Krueger RA, Casey MA. Focus groups: a practical guide for applied research. 4th ed. Thousand Oaks: Sage; 2009.
- Larsson S, Lawyer P, Garellick G, Lindahl B, Lundstrom M. Use of 13 disease registries in 5 countries demonstrates the potential to use outcome data to improve health care's value. *Health Aff.* 2012;31(1):220–7.
- Lee K. Global institutions: the World Health Organization (WHO). New York: Routledge; 2008.
- Levy D, Brink S. A change of heart: how the people of Framingham, Massachusetts, helped unravel the mysteries of cardiovascular disease. New York: Knopf; 2005.
- Mullner RM, editor. Encyclopedia of health services research. 2 Vol. Thousand Oaks: Sage; 2009, xxix.
- National Cancer Institute. SEER as a research resource. NIH Publication No. 10-7519. Bethesda: SEER Program, National Cancer Institute; 2010. Available at: [www.seer.cancer.gov/about/factsheets/SEER\\_Research\\_Brochure.pdf](http://www.seer.cancer.gov/about/factsheets/SEER_Research_Brochure.pdf)
- National Center for Health Statistics (NCHS). National health survey: the principal source of information on the health of the U.S. population. Hyattsville: National Center for Health Statistics; 2010. Available at: [www.cdc.gov/nchs/data/nhis/brochure2010January.pdf](http://www.cdc.gov/nchs/data/nhis/brochure2010January.pdf)
- National Center for Health Statistics (NCHS). Health, United States, 2013: with special feature on prescription drugs. Hyattsville: National Center for Health Statistics; 2014a. Available at: [www.cdc.gov/nchs/data/hus/hus13.pdf](http://www.cdc.gov/nchs/data/hus/hus13.pdf)
- National Center for Health Statistics (NCHS). Summary of current surveys and data collection systems. National Center for Health Statistics; 2014b. Available at: [www.cdc.gov/nchs/data/factsheets/factsheet\\_summary1.pdf](http://www.cdc.gov/nchs/data/factsheets/factsheet_summary1.pdf)
- National Committee for Quality Assurance (NCQA). The essential guide to health care quality. Washington, DC: National Committee for Quality Assurance; 2006. Available at: [www.ncqa.org](http://www.ncqa.org)
- National Research Council, Committee on National Statistics. Vital statistics: summary of a workshop. Washington, DC: National Academies Press; 2009. Available at: [www.ncbi.nlm.nih.gov/books/NBK219877/](http://www.ncbi.nlm.nih.gov/books/NBK219877/)
- Newhouse JP, The Insurance Experiment Group. Free for all?: lessons from the RAND health experiment. Cambridge, MA: Harvard University Press; 1993.
- NORC at the University of Chicago. Social science research in action. Chicago: NORC at the University of Chicago; 2011. Available at: [www.norc.org/PDFs/Brochures-Collateral/NORC\\_Book\\_Social\\_Science\\_Research\\_in\\_Action.pdf](http://www.norc.org/PDFs/Brochures-Collateral/NORC_Book_Social_Science_Research_in_Action.pdf)
- Office of the Assistant Secretary for Planning and Evaluation (ASPE). U.S. Department of Health and Human Services (HHS). Guide to HHS surveys and data resources. Washington, DC: U.S. Department of Health and Human Services; n.d. Available at: [www.aspe.hhs.gov/sp/surveys/index.cfm](http://www.aspe.hhs.gov/sp/surveys/index.cfm)
- Organisation for Economic Co-operation and Development (OECD). Health at a glance 2013: OECD indicators. Paris: Organisation for Economic Co-operation and Development; 2013. Available at: [https://doi.org/10.1787/health\\_glance-2013-en](https://doi.org/10.1787/health_glance-2013-en)
- Osborn R, Moulds D, Squires D, et al. International survey of older adults finds shortcomings in access, coordination, and patient-centered care. *Health Aff.* 2014;33(12):2247–55.
- Painter MJ, Chernew ME. Counting change: measuring health care prices, costs, and spending. Princeton: Robert Wood Johnson Foundation; 2012. Available at: [www.rwjf.org](http://www.rwjf.org)
- Perrin JM. Health services research for children with disabilities. *Milbank Q.* 2002;80(2):303–24. Available at: [www.ncbi.nlm.nih.gov/pmc/articles/PMC2690116/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2690116/)
- Pharmaceutical Research and Manufacturers of America. 2014 Biopharmaceutical Research Industry Profile. Washington, DC: Pharmaceutical Research and Manufacturers of America; 2014. Available at: [www.phrma.org/sites/default/files/pdf/2014\\_PhRMA\\_PROFILE.pdf](http://www.phrma.org/sites/default/files/pdf/2014_PhRMA_PROFILE.pdf)
- RAND Corporation. RAND Corporation: annual report 2013. Santa Monica: RAND Corporation; 2013. Available at: [www.rand.org/pubs/corporate\\_pubs/CPI-2013.html](http://www.rand.org/pubs/corporate_pubs/CPI-2013.html)
- Rhoades JA, Cohen SB. The long-term uninsured in America, 2009–12 (selected intervals): estimates for the U.S. civilian noninstitutionalized population under age 65. Medical Expenditure Panel Survey (MEPS) Statistical Brief #464. Rockville: U.S. Agency for Healthcare Research and Quality (AHRQ); 2014. Available at: [www.meps.ahrq.gov/mepsweb/data\\_files/publications/st464/stat464.pdf](http://www.meps.ahrq.gov/mepsweb/data_files/publications/st464/stat464.pdf)
- Robert Wood Johnson Foundation. Consumer attitudes on health care costs: insights from focus groups in four U.S. cities. Princeton: Robert Wood Johnson Foundation; 2013. Available at: [www.rwjf.org/content/dam/farm/reports/issue\\_briefs/2013/rwjf403428](http://www.rwjf.org/content/dam/farm/reports/issue_briefs/2013/rwjf403428)
- Stagnitti MN, Carper K. National health care expenses in the U.S. civilian noninstitutionalized population, distributions by types of service and source of payment, 2012. Medical Expenditure Panel Survey (MEPS) Statistical Brief #456. Rockville: U.S. Agency for Healthcare Research and Quality (AHRQ); 2014. Available at: [www.meps.ahrq.gov/mepsweb/data\\_files/publications/st456/stat456.pdf](http://www.meps.ahrq.gov/mepsweb/data_files/publications/st456/stat456.pdf)



- Swartout KA. Encyclopedia of associations: national organizations of the U.S. Farmington Hills: Gale Cengage Learning; 2014. Available at: [www.gale.cengage.com](http://www.gale.cengage.com)
- Wakefield MK. Statement by HRSA administrator Mary K. Wakefield, Ph.D., R.N. on the National Practitioner Data Bank Public Use File, 9 Nov 2011. Available at: [www.npdb.hrsa.gov/resources/publicDataStatement.jsp](http://www.npdb.hrsa.gov/resources/publicDataStatement.jsp)
- Wennberg JE. Tracking medicine: a researcher's quest to understand health care. New York: Oxford University Press; 2010. p. 1–13.
- Williamson A, Hoggart B. Pain: a review of three commonly used pain rating scales. *J Clin Nurs*. 2008;14(7):798–804. Available at: [www.onlinelibrary.wiley.com/doi/10.1111/j.1365-2702.2005.01121.x/pdf](http://www.onlinelibrary.wiley.com/doi/10.1111/j.1365-2702.2005.01121.x/pdf)
- Workman TA. Engaging patients in information sharing and data collection: the role of patient-powered registries and research networks. AHRQ Community Forum White Paper, AHRQ Publication No. 13-EHC124-EF. Rockville: U.S. Agency for Healthcare Research and Quality; 2013. Available at: [www.effectivehealthcare.ahrq.gov/ehc/assets/File/Patient-Powered-Registries-white-paper-130911.pdf](http://www.effectivehealthcare.ahrq.gov/ehc/assets/File/Patient-Powered-Registries-white-paper-130911.pdf)
- World Health Organization (WHO). World health survey, 2002, B – individual questionnaire. 2002. Available at: [www.who.int/healthinfo/survey/en/](http://www.who.int/healthinfo/survey/en/)
- World Health Organization (WHO). World health statistics, 2014. Geneva: WHO Press; 2014. Available at: [www.who.int](http://www.who.int)
- Xu F, Mawokomatanda T, Flegel D, et al. Surveillance for certain health behaviors among states and selected local areas – behavioral risk factor surveillance system, United States, 2011. *Morb Mortal Wkly Rep (MMWR)*. 2014;63(9):1–149. Available at: [www.cdc.gov/mmwr/pdf/ss/ss6309.pdf](http://www.cdc.gov/mmwr/pdf/ss/ss6309.pdf)



# Health Services Information: Application of Donabedian's Framework to Improve the Quality of Clinical Care

# 6

A. Laurie W. Shroyer, Brendan M. Carr, and Frederick L. Grover

## Contents

<b>Introduction</b> .....	110
<b>National Committee for Quality Assurance</b> .....	111
<b>Dr. Ernest Amory Codman's Data-Driven Approach to Defining and Measuring Quality of Care</b> .....	112
<b>Dr. Avedis Donabedian's Process-Structure-Outcome Model for Quality of Care</b> .....	113
<b>Processes of Care</b> .....	113
<b>Structures of Care</b> .....	114
<b>Outcomes of Care</b> .....	115
<b>Process-Structure-Outcomes in Cardiac Surgery</b> .....	115
<b>Risk Adjustment</b> .....	117
<b>Uncertainty</b> .....	119
<b>Implementation of VA National Quality Improvement Programs</b> .....	120
<b>The Processes, Structures, and Outcomes of Cardiac Surgery Study</b> .....	121

---

A. L. W. Shroyer (✉)  
Department of Surgery, School of Medicine, Stony Brook  
University, Stony Brook, NY, USA  
e-mail: [annielaurie.shroyer@stonybrookmedicine.edu](mailto:annielaurie.shroyer@stonybrookmedicine.edu)

B. M. Carr  
Department of Emergency Medicine, Mayo Clinic,  
Rochester, MN, USA

F. L. Grover  
Department of Surgery, School of Medicine at the  
Anschutz Medical Campus, University of Colorado,  
Aurora, CO, USA

<b>Hypotheses of the PSOCS Study</b> .....	122
<b>Methods of the PSOCS Study</b> .....	123
<b>Findings of the PSOCS Study</b> .....	124
<b>The CICS-P-X Program</b> .....	124
<b>Measuring Processes of Care</b> .....	125
<b>Monitoring Trends Over Time</b> .....	127
<b>Implementation of National Quality Improvement Programs</b> .....	129
<b>Uncovering Quality Trends</b> .....	130
<b>The Michigan Society of Thoracic and Cardiovascular Surgeons Quality Collaborative</b> .....	131
<b>The American College of Surgeons' Private Sector Initiative</b> .....	132
<b>Implementation Challenges: Dilemmas Faced by Quality Measurement Projects</b> .....	134
<b>Summary</b> .....	135
<b>References</b> .....	136

### Abstract

This chapter provides a summary of the well-established conceptual models related to measuring and improving the overall quality of medical care as described by the Institute of Medicine and founded upon Donabedian's historical triad for quality measures. The subcomponents required for quality measurement are first identified, including (1) patient risk factors, (2) processes of care, (3) structures of care, (4) clinical outcomes, and (5) resource utilization or costs of care. The key challenges associated with applying this quality of care conceptual model to designing and implementing new research projects are then discussed, including the following cutting-edge measurement-related topics: (1) dealing with missing data (e.g., clinical substitution versus statistical imputation), (2) differentiating planned versus unplanned processes of care (e.g., distinguishing between interventions used as a matter of routine and those interventions that were initiated in response to observed changes in the patient's status), (3) evaluating the differential impact of sequential versus nonsequential timing of

events (e.g., cascading of outcomes), and (4) assessing the relative impact of medical versus nonmedical care influences upon the quality of patient medical care rendered. Historical projects designed to define, measure, and evaluate the quality of cardiac surgical care in the Department of Veterans Affairs Continuous Improvement in Cardiac Surgery Program and Society of Thoracic Surgeons National Adult Cardiac Database are presented to illustrate how these quality of care concepts can be applied. The challenges in using clinical databases to evaluate quality of care are then summarized. Finally, several innovative approaches are described toward advancing the future practice of quality measurement research.

### Introduction

Within the healthcare field, a diversity of approaches has been used to define, to measure, and to improve the quality of medical care services, attempting to optimize the opportunities to improve clinical care outcomes while

evaluating whether the actual outcomes incurred achieve the original expectations. As perhaps one of the earliest documented descriptions related to defining or measuring the quality of medical care, King Hammurabi's Code (1,700 BC) provided insights as to what were considered unacceptable care outcomes as compared to the expectations, providing clear instructions as to the direct consequences to clinicians for the delivery of substandard care:

If a physician performed a major operation on a nobleman with a bronze lancet and caused the nobleman's death, or he opened the eye-socket of a nobleman and destroyed the nobleman's eye, they shall cut off his hand. (Magno 1975)

To optimize quality of medical care, there exist at many facilities patient safety initiatives focused on engaging healthcare professionals, organizations, and patients toward the attainment of a healthcare system that reduces errors with a focus to consistently improve the care provided (based on previously identified challenges occurring) and to create an institutional culture focused upon assuring *patient safety as a top priority*.

Institutional patient safety cultures can foster and support the design and implementation of ideal clinical practices. This would be exemplified by an institutional culture that focuses on reducing the risk of adverse events occurring. Even with application of the best evidence available, unforeseen adverse consequences of the medical care provided unfortunately still do occur. As an example, the perioperative administration of prophylactic antibiotic therapy for patients undergoing surgery is commonly cited as a patient safety practice employed to prevent surgery-related infections in the postoperative period (van Kasteren et al. 2007). In spite of this important intervention, however, postoperative infections still remain an outstanding challenge faced by many healthcare institutions, with multiple approaches implemented to keep postoperative infection rates low (e.g., conscience handwashing techniques used routinely, combined with sterile techniques for wound dressing changes).

## National Committee for Quality Assurance

Toward the goal of providing quality rankings, the National Committee for Quality Assurance (NCQA) provides an infrastructure support of a broad array of programs and services focused on measuring, analyzing, and continually improving the healthcare provided by US-based health plans. The National Committee for Quality Assurance has defined quality metrics that can be used to identify opportunities for quality improvement. The routine reporting of quality metrics has been useful to inform decisions at the clinical program, facility, health plan, and policy levels. By providing publicly available statistical reports evaluating health plan performance, important quality improvements have been documented and translated into reduced adverse event rates impacting patient care. For example, the use of beta-blockers for the subgroup of patients with a prior acute myocardial infarction (aka AMI or "heart attack") has been documented in the peer-reviewed literature to reduce the chance of a repeat AMI by 40% (National Committee for Quality Assurance 2014a). Thus, beta-blocker use has been cited as an NCQA successful metric used to facilitate positive trends documented for quality of care outcomes.

Moreover, as part of the National Committee for Quality Assurance, the Healthcare Effectiveness Data and Information Set (HEDIS) was developed, and, as of 2014, the vast majority of US-based health plans submitted HEDIS metrics (which consisted of 81 measures for quality of care across five different care domains). The National Committee for Quality Assurance HEDIS requires that plans report the continued post-AMI use rates for beta-blocker medications for their eligible population. That is, health plans must calculate the proportion of their eligible enrollees (aged 18 years or older) who received persistent beta-blocker treatment for 6 months after discharge following their AMI hospitalization over the past year period. Although this specific HEDIS metric is most relevant to a smaller sized subgroup of ischemic heart disease patients, the National Committee for Quality Assurance

PERSISTENCE OF BETA-BLOCKER TREATMENT RATE					
YEAR	COMMERCIAL		MEDICAID	MEDICARE	
	HMO	PPO	HMO	HMO	PPO
2013	83.9	81.4	84.2	90.0	89.4
2012	83.9	79.5	82.0	88.9	88.5
2011	81.3	77.0	80.5	87.3	86.2
2010	75.5	71.3	76.3	83.1	82.5
2009	74.4	69.6	76.6	82.6	78.9
2008	75.0	68.8	73.6	79.7	76.7
2007	71.9	62.9	62.0	75.5	70.4
2006	72.5	65.5	68.1	69.6	70.9
2005	70.2	64.3	69.8	65.4	58.5

**Fig. 1** National Committee for Quality Assurance Healthcare Effectiveness Data and Information Set 2014 report on persistence of beta-blocker treatment

trends over time documented in facility performance are impressive. Specifically, the reported improvements for this HEDIS metric over time for the National Committee for Quality Assurance health plans (including commercial, Medicaid, and Medicare populations) for both health maintenance organizations (HMO) and preferred provider organizations (PPO) were reported, as shown in Fig. 1 (National Committee for Quality Assurance 2014b).

Importantly, the National Committee for Quality Assurance seal of approval is given to health plans that meet their published requirements, including adherence to more than 60 preestablished standards, as well as reported performance for more than 40 quality metrics. The National Committee for Quality Assurance provides the *Quality Rating System (QRS) Measure Technical Specifications*, a technical manual that details each quality metric's specification and provides guidelines for data capture. Thus, the National Committee for Quality Assurance Quality Rating System fulfills the reporting requirements set forth as part of the recent

Affordable Care Act, leading the way to providing publicly available information to aid consumers in selection of qualified health plans (QHPs), as well as routinely monitoring qualified health plan quality.

### Dr. Ernest Amory Codman's Data-Driven Approach to Defining and Measuring Quality of Care

As one of the “founding fathers” of the historical healthcare quality movement, Dr. Ernest Amory Codman (1869–1940) conceived of the “end result” hospital, where the long-term outcomes following surgery would be documented and evaluated for each patient to identify opportunities for future medical care improvements (Codman 2009). As part of his original hand-tallied quality-of-care report card, Dr. Codman tabulated the findings for over 600 abdominal surgical cases over a decade, classifying his findings by individual surgeon operators and by diagnosis and treatment approaches used. As a tribute to his outcomes-measurement legacy, the Joint Commission created the “Ernest Amory Codman Award” in 1996, designed to enhance knowledge and encourage the use of performance measurement to improve healthcare quality and safety (The Joint Commission 2014).

In his multiple roles as a clinician, a leader, and an advocate for quality management, Dr. Codman emphasized the need for monitoring and improving surgical quality of care. His original concerns related to data-driven improvement of quality-of-care decisions remain critically relevant even today (Nielsen 2014). The distinguishing factor between the historical and current healthcare debates, however, relates to the plethora of current data available versus the scarcity of data that had been gathered historically. In spite of the overabundance of data captured to meet government, insurer, and accreditation requirements, an outstanding challenge remains to identify *relevant, meaningful information* that can be used to improve the quality of care and to further advance the field of quality measurement.

## Dr. Avedis Donabedian's Process-Structure-Outcome Model for Quality of Care

As a more contemporary leader shaping today's quality of healthcare paradigm, Dr. Avedis Donabedian (1919–2000) established a conceptual model for quantifying healthcare quality improvements. Specifically, Dr. Donabedian noted that “. . .quality may be judged based on improvements in patient status obtained, as compared to those changes reasonably anticipated based on the patient's severity of illness, presence of comorbidity, and the medical services received” (Donabedian 1986). As part of his approach to defining and measuring quality of care, Dr. Donabedian described two related processes of care domains for assessing the quality of medical care: *interpersonal excellence* and *technical excellence* (Donabedian 1980).

Excellence in the *interpersonal quality domain* is associated with a *patient-centered focus to how the care is provided*, and the degree of excellence achieved is based on the degree to which the care meets the patient's unique needs (including information, physical, and emotional needs) in a manner consistent with the patient's expectations and preferences. As part of the interpersonal care domain, incorporating the patient's decisions directly in the decision-making process is recognized as an important component of excellent quality care. Further, this implies that patient satisfaction with the care provided may not mirror quality if the patient reports high satisfaction while the degree to which their needs are met is less than desired – such that the quality of care rendered would be classified as “low.” Thus, there may not always be a concordance between patient satisfaction and patient-centered quality of care assessments.

The second domain, the technical quality of care, is related to the *degree of alignment between the care that was provided and the care that might have been provided based upon the current professional care standards*, as well as the degree of improvements in patient outcomes that occurred (as compared with the changes in outcomes that may have been otherwise anticipated). To

systematically evaluate quality, Dr. Donabedian put forward an approach that is now commonly used based on his “*Process-Structure-Outcomes*” framework that incorporated the domains of (1) processes of care, (2) structures of care, and (3) clinical outcomes (Donabedian 1988). In 1997, Dr. Donabedian was recognized by the Joint Commission as the first recipient in the individual category to receive the Ernest Amory Codman Award (The Joint Commission 2014).

---

### Processes of Care

Dr. Donabedian's approach to assessing quality focused first on evaluating the processes of care. *Processes of care* may be defined as the “set of procedures and/or skills with which health care technology of proven or accepted efficacy is delivered to individual patients” (Shroyer et al. 1995). This includes the processes associated with care provider actions, as well as the patient's activities related to seeking and obtaining care. Thus, processes of care assessments verify that patients received what is known to be the appropriate care, by the standards of evidence-based medicine. Processes of care may include communication processes, such as a post-discharge telephone follow-up call, as well as social or emotional support-related activities. In evaluating processes of care, it is important to recognize that the patient's characteristics and historical medical care received need to be factored into this assessment. Specific patient subpopulations may be targeted for specific processes of care, whereas these same processes of care may be contraindicated for other patient subgroups. Finally, the patient-based or family-based actions taken – that is, the patient's or their family's actions to seek care, to adhere to the treatment plan, or to select to not participate in care offered (e.g., by requesting “do-not-resuscitate” orders) – are important considerations to complement the provider-based processes of care evaluated (National Quality Forum 2009).

For example, the use of the left internal mammary artery (LIMA) as a conduit for coronary artery bypass grafting (CABG) surgical

procedures has been documented to improve long-term survival in comparison to the use of other conduits, e.g., saphenous vein grafts (SVGs) (Goldman et al. 2004). The earlier cardiac surgery clinical guidelines identified that the use of LIMA for CABG surgery should be considered where longer-term survival may be an important consideration. More recent published literature has extended the LIMA benefits documented to include the elderly population. As with any surgical procedure, there are risks and benefits associated with every procedure, including LIMA use. The use of a LIMA graft generally takes more time; therefore, a LIMA graft may be contraindicated for emergent/urgent patients where surgical cross-clamp time may be critically important. Hence, the CABG LIMA use rates may be used as a quality of care metric for elective patients, but may not be a meaningful measure of quality of care for the emergent/urgent patient subgroups (Karthik and Fabri 2006).

---

## Structures of Care

*Structures of care*, as another important metric to assess quality, were defined by Dr. Donabedian as being related to the “overall context or environment in which care is rendered to a group of patients,” including the characteristics of healthcare team members (e.g., credentials and experience) and healthcare facilities (e.g., the type and age of equipment) (Shroyer et al. 1995). Representing an important arm of Donabedian’s triad, structures of care include the manner in which healthcare facilities are organized and operated, the approaches used for care delivery, and the policies and procedures related to care including quality oversight processes.

For example, structures of surgical care may involve the physicians’ provider-specific characteristics, e.g., international medical graduate (IMG) or board certification status. Though not definitive, studies have shown no difference in mortality outcomes among hospitalized patients treated by graduates of US medical schools versus IMGs. There may, however, be a correlation between board certification and better clinical

outcomes (Sharp et al. 2002; Norcini et al. 2010).

Facility characteristics, such as a hospital’s affiliation (academic versus community) or location (urban versus rural versus frontier hospitals), have also been studied as structural characteristics that have been documented to impact the quality of care provided. Academic affiliation, for instance, has not been shown to be a predictor of better outcomes (Papanikolaou et al. 2006). Location may be important, however, as rural hospitals have been shown to have worse performance on quality of care indicators than urban hospitals, in spite of studies showing their outcomes to not be inferior to those at urban hospitals (Nawal Lutfiyya et al. 2007; Dowsey et al. 2014; Tran et al. 2014). The identification of disparities such as this not only demonstrates the important role of structures of care in affecting the quality of care but may serve as an impetus to identify changes that can be made in the structures themselves to improve patient care.

Importantly, the entire process associated with accreditation, including the Joint Commission, is intended to coordinate a quality oversight mechanism, which, in theory, should validate the importance of structural measures for care. For example, the field of cardiac surgery has established minimal acceptable standards for nurse staffing ratios to be coordinated in critical care units for immediate post-CABG patient care. In order to be deemed of “acceptable” quality, standards for the number and type of nurse staffing must be met to assure that a high quality of care may be provided. For example, a study by VillaNueva and colleagues looked at risk-adjusted outcomes of cardiac surgery patients in relation to (1) “the demographics, education, experience, and employment of operation room (OR) and surgical intensive care unit (ICU) nurses involved in their care” and (2) “the staffing and vacancy ratios of OR and surgical ICU nurses involved in their care.” Significant variations were observed in processes of care between participating cardiac surgery centers, but there was insufficient data to draw conclusions on their effect on patient outcomes (VillaNueva et al. 1995). For this study, therefore, the theoretical link between structures

of care and outcomes of care could not be confirmed directly. Within Donabedian's quality triad, there is a fundamental assumption underlying the assessment of structural quality elements; that is, the healthcare setting in which the care is rendered is a very important factor influencing the quality of medical care provided. In spite of the data-driven evidence being sparse, this assumption extends to the current Joint Commission accreditation assessments focused on evaluating the adequacy of healthcare facility basic structure of care.

## Outcomes of Care

Finally, *outcomes of care*, the third piece of the triad, were defined by Dr. Donabedian as the measurable end points of the healthcare process (Malenka and O'Connor 1998).

Ideally, a broad range of clinically relevant outcomes should be assessed including (but not limited to) traditional measures of mortality and morbidity, health-related quality of life, condition-specific or disease-specific metrics of symptom status or functionality, general health status, and general overall functionality, or patient satisfaction. The outcomes measured should be related to the full range of care end points salient to the patients impacted by the treatment received. Prioritized in importance based upon the nature of the question raised, outcomes may reflect a patient's status at a single point in time (e.g., 30-day operative mortality) or changes over points in time (e.g., pre-CABG angina frequency compared to post-CABG 6-month follow-up angina frequency). For quality assessment purposes, moreover, outcomes may be subclassified

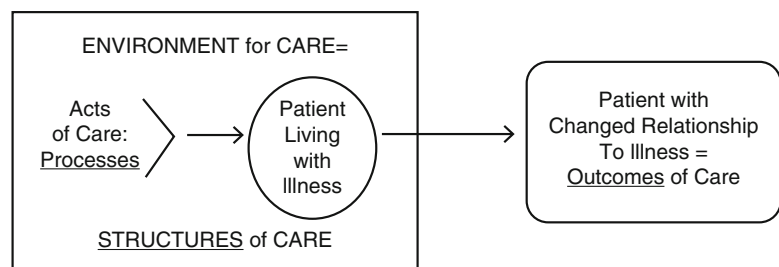
in many ways including (1) planned (intended) or not planned (unintended) (Mavroudis et al. 2014), (2) preventable versus not preventable (Lee et al. 2003), (3) major versus minor in importance, and (4) related or not related to the medical care rendered (Shann et al. 2008).

Figure 2 illustrates the hypothetical interactions between processes, structures, and changes in the patient's outcomes of care, where patients present to the healthcare system with an illness, in context of their other patient risk characteristics. The medical care interventions received represent processes of care, as well as the actions taken by patients themselves to address their illness state. These actions are coordinated within a healthcare environment, representing the structures to care. Pending the passage of time, the patient emerges from their episode of care with a changed relationship to their illness, which is the outcome of care measured. This "Process-Structure-Outcome" paradigm can be extended from a single episode of care to the full series of care encounters, in order to assess and to improve the quality of patient care received.

## Process-Structure-Outcomes in Cardiac Surgery

Supporting these different outcome-based classification systems, multiple examples have been reported within the field of cardiothoracic surgery. Delayed sternal closure, for example, may be planned or unplanned (1). In pediatric cardiac surgery in particular, the surgeon may plan to leave the sternum open at the end of the procedure because this may allow for better heart function in certain patients. In other cases, however, the

**Fig. 2** Theoretical "process-structure-outcome" framework





surgeon may have initially planned to close the patient's chest, but found himself or herself unable to as a result of bleeding, myocardial edema, or arrhythmia (Yasa et al. 2010; Ozker et al. 2012). It is important to distinguish between the two when investigating the incidence of delayed sternal closure as a surgical complication, because planned delays in closure could inflate the apparent incidence of surgical complications. On the other hand, reintubation is rarely planned, but may be preventable (2) if it is brought about by unplanned extubation or as a complication of a neuromuscular blocking agent rather than a non-iatrogenic respiratory problem (Lee et al. 2003). Major and minor outcomes (3) are easily envisioned based on the degree to which they impact the patient (e.g., death or nonfatal myocardial infarction versus new-onset atrial fibrillation after CABG surgery, respectively). Finally, outcomes may be unrelated to the medical care rendered (4) when they're accepted as a normal consequence of a procedure in a certain fraction of patients. Microembolic events, for example, are known to be an unpreventable consequence of the use of extracorporeal circulation (i.e., cardiopulmonary bypass during cardiac surgery), while an embolic stroke involving a territory of brain circulation is not (Shann et al. 2008).

To support clinical decision-making, the outcomes identified for the medical care rendered should focus on the most clinically relevant end points or changes and may be judged in comparison with the best possible outcomes anticipated with the use of good processes and structures of care. Outcomes are often reported as rates, for example, the rate of a serious adverse event following a surgical procedure. For coronary artery bypass graft (CABG-only) procedures, for example, the national rate reported for a 30-day operative mortality by the Society of Thoracic Surgery (STS) for the period from 1996 to 2009 was 2.24% (Puskas et al. 2012). For an outcome to be useful, it must be compared across different populations that have the potential to achieve this desired end point and also compared to reference standards determined from the expected ideal outcome rate to be achieved. For example, the rates for Department of Veterans Affairs (VA) CABG

30-day operative mortality may be compared to non-VA/STS hospital rates (Public Law 99-166 1985), or these rates can be compared across time, by examining the metric for different periods. To be most useful as quality assessment metrics, it may be important to make comparisons of different outcome rates across key patient subgroups that did or did not receive specific treatments (e.g., rates of mediastinitis during the 30-day perioperative period for post-CABG patients treated versus not treated with a prophylactic antibiotic therapy). Moreover, goals for specific procedure-based outcomes can be proactively established, such as the STS national objective to achieve a 1% 30-day operative mortality rate for lower-risk CABG-only patients in the future (Mack 2012).

As part of a National Institutes of Health (NIH) initiative in 2004, a new repository entitled the Patient Reported Outcomes Measurement Information System (PROMIS<sup>®</sup>) system of measures was established. The PROMIS<sup>®</sup> metrics included patient self-reported mental, physical, and social health status as assessments of the patient's perception of their overall well-being. The PROMIS<sup>®</sup> surveys identified how patients reacted and described how patients felt during specific times during care received for a preestablished set of conditions (National Institutes of Health 2014). To evaluate treatment effectiveness, PROMIS<sup>®</sup> assessments can be used as primary or secondary end points in clinical studies.

*Intermediate outcomes*, as observations in the pathway that directly lead to the final longer-term outcomes, have also been commonly measured. Specifically, intermediate outcomes may be commonly associated with processes of care, as key steps in the journey to obtaining a desired longer-term health states. For example, the current ischemic heart disease guidelines promulgated by the American Heart Association would recommend that CABG patients be discharged from the hospital receiving lipid-lowering medications. At discharge, the use of lipid-lowering medications can be documented, as well as the patient's current total cholesterol level (as well as high-density lipoprotein and low-density lipoprotein subcomponents). As an important marker related to post-

CABG patient's long-term survival, therefore, both lipid-lowering medication use (as a process of care measure) and patient cholesterol measures following the CABG hospital discharge over time (as intermediate outcomes) may be assessed as part of a quality assurance program (Hiratzka et al. 2007).

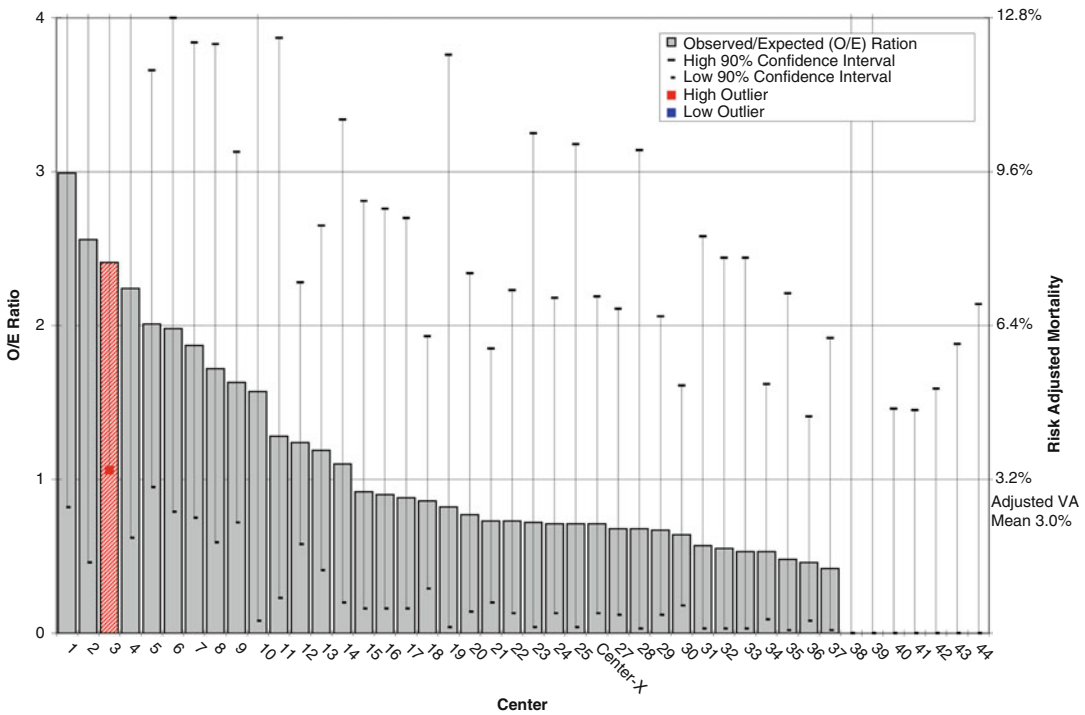
---

## Risk Adjustment

Although outcomes are considered by many to be the ultimate measure of quality of care, they are, to a large degree, influenced by the patient's pretreatment condition and their unique characteristics (e.g., *risk factors* that influence outcomes). The goal of risk adjustment (a statistical analysis isolating the relationship between the outcomes of interest and the treatment effects of interest) is to control for the effects of other patient-relevant factors, although a patient's pretreatment status may not be easy to measure. Specifically, patient risk factors may be defined as those characteristics that "***place patients at an enhanced risk that an unfavorable event may occur***" (Blumberg 1986). Generally, risk factors may be classified as modifiable (e.g., related to lifestyle or health behavior choices) or non-modifiable (e.g., related to the patients' demographic characteristics, socioeconomic status, or their genetic propensity to incur disease-related adverse conditions). In evaluating a patient's risk profile, it is of paramount importance to identify the patient's severity of disease and comorbidities (e.g., other diseases that may impact a patient's likelihood of experiencing an adverse event related to the primary disease being considered). In the realm of cardiac surgery risk adjustment of outcomes, patients' demographic factors (e.g., age or gender) and socioeconomic status (e.g., highest educational level attained) along with the severity of their coronary disease and complexities of their comorbidities have been demonstrated to be related to risk-adjusted outcomes. Moreover, patient-based choices related to healthy behaviors (e.g., body mass index) and lifestyle (e.g., smoking status) may also influence their probability for having a major adverse event (Nashef et al. 2012).

Importantly, patient risk factors may predispose patients to appropriately receive different types of treatments or be excluded from consideration for a specific treatment or set of treatments. Based on these same risk characteristics, therefore, patients may be pre-selected by providers to be eligible to receive care or differential types of treatments. The patient characteristics that relate to the *propensity* of a provider (or set of providers) to select them for treatment may be considered in a slightly different modeling approach, related to a propensity analysis (Blackstone 2001, 2002). Based on a patient population's likelihood to receive a specific treatment (which may also be related to their risk characteristics) using a propensity analysis, a risk-adjusted analysis can be performed to identify the quality of care rendered to the patient.

The risk-adjustment process, using a statistical modeling approach, calculates an "expected" risk ("E") for each patient uniquely. Based on aggregating patient data, the sum of the "expected" risks for an adverse outcome may be compared to the sum of the "observed" adverse outcomes to identify a patient subpopulation **O/E ratio**. Any specific patient subpopulation or provider-based O/E ratio that is statistically different from the value of 1.0 (i.e., where the ratio of the "expected" event rate to the "observed" event rate falls outside of the preestablished confidence interval) may be classified as a "high" outlier – that is, the O/E ratio is statistically higher than the value of 1.0. Similarly, a "low" outlier can be identified based on an O/E ratio that is statistically significantly lower than the value of 1.0. In general, the risk-adjusted outcome "high" outliers are identified for more intensive quality reviews (e.g., expanded chart reviews or site visits) for potential quality challenges by oversight groups. In contrast, the "low" outliers may serve as potential opportunities to identify differential processes or structures of care that may be exemplary to serve as a "benchmark" for others, as a template to consider for quality improvement. In general, quality assurance processes may tend to use more generous confidence intervals (e.g., 90% confidence intervals) in order to be sensitive – that is, to screen in additional patients or provider



**Fig. 3** Example: observed/expected ratio comparison for Department of Veterans Affairs coronary artery bypass graft 30-day operative mortality

subgroups for closer quality assurance (QA) review activities (Shroyer et al. 2008).

As illustrated in Fig. 3 (which is an example report), VA medical center #3 would likely be identified in preliminary reviews as a “high-outlier” facility and may subsequently be screened for potential quality of care concerns, given that the observed rate for 30-day operative mortality is statistically significantly higher than the rate that would have been expected based on evaluating the patient risk characteristics for a CABG procedure. As documented, many quality assurance reports commonly will use a liberal p-value threshold (such as  $p < 0.10$ ) to attempt to screen in more facilities for an in-depth quality review, casting a broader net for the next step in the review process. As O/E ratios (in and of themselves) are not definitive measures of quality of care, VA medical center #3 potentially might be selected for a detailed chart review and possibly a site visit (pending the results of the chart review) to explore for possible quality of care challenges. In contrast,

center X had no statistically significant difference identified between their O and E rates, indicating no need for further quality investigation related to this specific end point. Finally, there are no “low-outlier” facilities in this example, as the confidence intervals for the O/E ratios for facilities #38–44 encompass the value of 1.0. If, however, there were “low outliers” identified, then these may be facilities to explore further with both in-depth chart reviews and/or site visits to identify “benchmark” care activities that may be useful to share and disseminate to other VA medical centers as “best practices.”

Though important, outcomes do have inherent limitations when used as quality of care metrics. Outcomes only indirectly provide information that a potential challenge may exist related to quality of care, but generally outcomes do not identify the specific actions needed to improve the quality of patient care. Moreover, outcomes do not usually provide an adequate level of information to guide the required changes as “action items” that can be

taken by providers directly. Hence, the importance of Donabedian's triad assessment for quality of care, as a complement of outcomes with processes and structures, is required.

---

## Uncertainty

For many quality-of-care endeavors, there is no adequate understanding of the relative impact of the patient risk factors upon adverse outcomes, nor adequate understanding of what might be the natural course of events had the patients not received any treatment or an alternative course of treatment. With a variety of care alternatives often available, the best approach to address a patient's unique risk factor profile is not always clear. For example, in treating patients with ischemic heart disease, there is strong evidence suggesting CABG to be the best care strategy for patients with two- or three-vessel disease. However, the situations where medical management should be used to optimize long-term survival versus manage angina symptoms versus a revascularization may not be completely clear, particularly for high-risk patients subgroups (e.g., patients with two prior heart surgical procedures, as well as current severe angina symptoms). Given that clinical guidelines may provide evidence-based care strategies for some but not all patient subpopulations (particularly the highest-risk patient subgroups), compliance with state-of-the-art evidence provides an important indicator of quality of care – that is, a process-based assessment to augment the risk-adjusted outcomes assessments that may be coordinated. Unfortunately, there is not always adequate evidence basis to coordinate guidelines: a recent evaluation identified that for the current ACC/AHA guidelines promulgated from 1994 to 2008, only 11% of the guidelines were based on rigorous scientific, high-quality data-driven evidence (based on a review of 53 guidelines on 22 topics, with a total of 7,196 recommendations evaluated) (Tricoci et al. 2009).

To improve quality of care, it is important not only to identify and to monitor outcomes but also to subject these risk-adjusted outcome reports to

critical review by the academic, industry, patient, and public targeted audiences. Over the past two decades, there has been an increasing emphasis placed on improving the public transparency as well as sharing reports of risk-adjusted provider-specific and facility-specific outcomes. As a case in point, the Society of Thoracic Surgeons has partnered with Consumer Reports to provide online provider-specific outcome reports, with risk-adjusted outcomes (The Society of Thoracic Surgeons 2012). Given that the availability of risk-adjusted outcomes information is increasing, it will be very interesting to observe the changes in both referral patterns and patient-provider choices that may occur over time in cardiac surgery utilization rates, revealing to what degree changes in patient patterns in obtaining care may be or may not be related to the use of risk-adjusted outcomes-based reports.

Emphasizing the clinician's role in quality improvement, Dr. Donabedian noted that "*An ideal physician is defined as one who selects and implements the strategy of care that maximizes health status improvement without wasted resources*" (Donabedian et al. 1982). Toward this goal, new quality of care metrics may be added to evaluate "timeliness" of care rendered. For example, Dr. Boris Sobolev and his Canadian-based research team have forged the way to identify patterns in surgery wait times, evaluating the impact of the timeliness of care rendered for patients upon both their short-term and longer-term outcomes (Sobolev and Fradet 2008). Dr. Sobolev has also done similar research in other surgical fields (e.g., general surgery and orthopedics) that has demonstrated that longer wait times do appear to have detrimental effects on patient outcomes across a variety of surgical fields and procedures (Sobolev et al. 2003; Garbuz et al. 2006). Moreover, the referral patterns related to the risk-adjusted outcomes may be stratified based on wait time delays, taking into consideration the patient's disease-related care processes – not just focusing on a patient's single cardiac surgical care encounter. Although early in the evolutionary process, the current focus of quality of care, which uses the patient encounter as the primary unit of analysis, is beginning to

transition to a disease management focus (e.g., evaluating the care provided related to the patient's ischemic heart disease) and toward a patient-based holistic health perspective (Fihn et al. 2012).

---

## Implementation of VA National Quality Improvement Programs

In 1972, the Department of Veterans Affairs (VA) established the Cardiac Surgery Consultants Board (CSCB) to provide quality assurance oversight for all VA-based cardiac surgery programs. Initially, the Cardiac Surgery Consultants Board review focus was placed on evaluating descriptive reports of observed mortality cases, as well as monitoring rates for both mortality and major morbidity outcomes. Chart audits and site visits were performed by the Cardiac Surgery Consultants Board to assure that minimum standards for quality of cardiac surgery were met by means of a peer-review process (Veterans Health Administration 2008).

In 1985, the Health Care Financing Administration (HCFA) release of hospital report cards raised the public's awareness of the wide variations experienced by hospitals for their surgical outcomes reported. Additionally, the Administration Health Care Amendments Act was passed, requiring that the VA establish a new quality assurance program which would identify significant deviations in risk-adjusted and unadjusted mortality and morbidity rates for surgical procedures when compared with prevailing national rates (Public Law 99-166 1985). Accordingly, the VA had also to determine if any discrepancies that were identified were related to differences in the quality of the VA-based healthcare services (Grover et al. 1990).

To address these legislative requirements, Drs. Hammermeister and Grover implemented in 1987 a new program entitled the "Continuous Improvement in Cardiac Surgery Program" (CICSP), gathering data related to each cardiac surgical patient's unique set of risk factors, surgical procedural details, and 30-day operative death outcomes. In December 1987, the first risk-adjusted reports for

30-day operative mortality were produced; these were further refined in June 1990. With the VA CICSP fully implemented, the first risk-adjusted outcomes reports (focused on mortality and major perioperative complications) were produced comparing the performance across of all VA-based cardiac surgery programs.

Before the end of 1990, the CICSP data form (originally comprised of 54 elements on a single sheet of paper) with associated definitions for risk, procedure-related, and outcome variables was mandated nationally by the VA as a new quality assurance requirement for all cardiac surgery programs. Based on the CICSP endeavor, a new noncardiac surgical quality improvement program, entitled the National Surgical Quality Improvement Program (NSQIP), was initiated in 1991 by Drs. Shukri Khuri and Jennifer Daley (Khuri et al. 1998). Expanding the focus to include a diversity of general surgical procedures, the VA NSQIP initiative partnered with the CICSP to obtain funding for local nurse or data coordinators to prospectively gather the patient preoperative risk characteristics, the detailed surgical processes of care, and the mortality and perioperative morbidity-related outcomes to be able to coordinate risk-adjusted mortality reports. Similar to the CICSP oversight coordinated by the Cardiac Surgery Consultants Board, the NSQIP established an Executive Committee (EC) with key analytical support coordinated by Dr. William Henderson. Working in concert, the VA Central Office of Surgical Services (under the leadership and guidance of Drs. Gerald McDonald and Ralph DePalma) synchronized the CICSP and NSQIP efforts to provide data-driven reports routinely to both the national oversight committees (Cardiac Surgery Consultants Board and NSQIP Executive Committee) as well as to share these reports with local and regional surgical program leaders (including Cardiothoracic Division Chiefs, Chiefs of Surgical Services, Medical Center leaders, and VA Regional Office leaders). As a primary focus, both CICSP and NSQIP chose to make their top priority the provision of good information to drive good local and regional decisions – to support internal VA-based self-assessment and self-improvement initiatives.

With directives and continuous improvement communications coordinated by Drs. McDonald and DePalma, they were able to successfully provide the right information at the right time to the right individuals, as key decision-makers, to empower them to take the right actions to improve the safety and the quality of patient care.

As the first national comprehensive surgical quality improvement endeavor, the efforts of these key VA leaders, including Drs. Hammermeister, Grover, Shroyer, Khuri, Daley, and Henderson, radically shifted the quality-of-care paradigm from crisis identification, focused on uncovering problem facilities or providers, where urgent action was needed to address deficiencies in care. The new goal was to improve the quality of care for all facilities and focused on evaluating metrics comprehensively over time (Itani 2009a, Rodkey and Itani 2009). These data-driven quality improvement programs have made major impacts. The NSQIP program has identified risk factors for morbidity and mortality across a wide range of surgical subspecialties, including general surgery, orthopedics, neurosurgery, and many others (Itani 2009b). These risk factors have set the stage for continuous improvement in the field of surgery by providing tools with which to better evaluate the role of surgery in individual patients' care and better identify patients for prophylactic measures or closer monitoring in the intra- and postoperative periods. Having established the initial CICSP and NSQIP's legacy, these VA programs provided an impetus, serving as models for others (such as the Northern New England Cardiovascular Consortium) to follow and to expand upon – with innovative enhancements (Malenka and O'Connor 1998).

---

### **The Processes, Structures, and Outcomes of Cardiac Surgery Study**

During the early CICSP implementation period (1987–1991), however, it is important to realize that both Drs. Hammermeister and Grover recognized that there were inherent limitations in

focusing on risk-adjusted outcome metrics as the ultimate quality of care metrics. Mortality, in and of itself, was a relatively rare event (under 3% mortality rate for CABG procedures). Given that the chart reviews and site visits performed by the VA Cardiac Surgery Consultants Board members often provided meaningful insights into the challenges that occurred with processes and structures of care, they initiated a new VA Health Services Research and Development Study entitled *Processes, Structures, and Outcomes of Cardiac Surgery* (PSOCS) to identify the important components of the cardiac surgical care rendered to veterans that may benefit by closer quality monitoring and reporting (Shroyer et al. 1995).

Funded in late 1991, the PSOCS study was initiated in May 1992 at 14 VA Medical Centers with active cardiac surgery programs (out of the 44 total VA cardiac surgery programs). The PSOCS study was a prospective cohort study, with funded research nurses and data support personnel. They gathered an extensive set of detailed data related to processes of care (including preoperative, intraoperative, postoperative, and post-discharge), structures of care related to the entire care provider team (e.g., team member's educational background, specialty training, years of experience, and level of certification), and the environment in which the care was rendered. The environment was comprehensively assessed, including data about the key features of the operating room, recovery room intensive care units, telemetry monitoring, staffing levels, and the quality and scope of oversight mechanisms. Additionally, the care provider interactions and communications were assessed via surveys. Finally, the nature and scope for surgical resident training were assessed, including the degree of supervision provided to the residents engaged in cardiac surgical patient care.

To complement the traditional mortality and morbidity outcome metrics routinely monitored by CICSP, a very broad array of outcomes was incorporated into the PSOCS study assessments. Focusing on the primary end points of death and major perioperative complications, outcome assessments were made at both 30 days following surgery or at the completion of the inpatient

hospitalization (whichever came sooner) and at 6 months post-CABG procedure. Comparing to baseline assessments, both a generic health-related quality-of-life instrument (i.e., the Veterans' version of the Short-Form 36 health-related quality of life survey, the VR-36) and a disease-specific survey (i.e., the Seattle Angina Questionnaire) were used to assess the PSOCS veteran self-perceptions of changes in physical, emotional, and social functionality related to changes in health status. Additionally, patient satisfaction with care was assessed to identify the concordance of patient self-reported outcomes with clinical outcomes of care, as well as to identify the factors that may influence a patient's CABG surgical care-related experiences.

---

## Hypotheses of the PSOCS Study

As the overarching research question, the PSOCS study identified the specific processes and structures of care that could be revised in the future to improve the quality of cardiac surgery patient care. Importantly, the PSOCS study established a vision that was based on a clinically relevant, conceptual framework of the wide diversity of processes and structures of care that may be related to patient risk-adjusted outcomes. Specifically, the PSOCS study evaluated comprehensively the literature for all factors known in surgery to be directly or indirectly related to changes in patient outcomes, coordinating these findings into a conceptual model that measured the variables identified. There were six specific process and three specific structure hypotheses, with corresponding sets of sub-hypotheses, that were related to the dimensions (and correspondingly the subdimensions) of the PSOCS conceptual model, tying each variable for which data was gathered into an organized hierarchical relationship of sets of variables, which could be analyzed in concert to address the specific research questions raised. For example, one PSOCS hypothesis focused on the intraoperative processes of care performed that may influence the short-term and intermediate-term patient outcomes. For intraoperative processes of care, there were ten

different sub-hypotheses evaluating a variety of the different intraoperative care dimensions, including operation duration, hemodynamic and physiologic monitoring techniques, management of hemodynamic function, anesthesia techniques used, blood management approaches, myocardial preservation technique, the use of the cardiopulmonary bypass machine, the surgeon-specific operative techniques used, the completeness of the documentation for intraoperative care provided, and the use of early extubation approaches. Given that a research nurse was located in the operating room for the duration of the procedure to independently record the care provided, the medical chart's completeness and quality of the documentation (e.g., the completeness of the surgeon's dictated operative note) could be assessed (O'Brien et al. 2004).

Each PSOCS hypothesis (or sub-hypothesis) was action driven; that is, the goal was to identify the specific actions that care providers or healthcare administrators or healthcare policy-makers would be able to take to improve the quality of future cardiac surgery patients' care. The PSOCS research questions raised were based on the following assumptions:

1. A significant proportion of post-CABG patients' risk-adjusted healthcare outcomes could be explained by processes and/or structures of care that could be improved.
2. The processes of cardiac surgical care that were most likely to impact risk-adjusted outcomes included the completeness and quality of the preoperative care processes, the intraoperative care processes, and the post-CABG processes of care, as well as the continuity of follow-up care in the post-discharge period.
3. The structures of cardiac surgical care that were most likely to impact risk-adjusted patient outcomes included the degree of supervision by senior physicians, the degree and effectiveness of communications both among care provider team members as well as between team members and the patient and family, and the nature and scope of the quality-related oversight coordinated as part

of the medical staff organization and regulatory activities that were performed as part of the hospital's quality integrating system.

4. The structures of care that may impact outcomes also included the number, education, experience, and specialty training of the physician provider team members (e.g., the surgeon, cardiologist, and anesthesiologist). Fundamentally, the provider team member characteristics, mix of providers providing care, and staffing levels, along with hospital and physician experience, were important structures that were hypothesized to impact patient outcomes, after holding patient-specific baseline risk factors constant (Shroyer et al. 1995).

Building on Dr. Donabedian's paradigm for quality of care, the PSOCS study assumed that good processes and good structures of care were very likely to lead to improved patient outcomes. Uncovering problems with specific processes of care or structure-related weaknesses in the provider-based characteristics, the clinical care team mix, or facility-based characteristics, could indicate targets for scrutiny, where different actions could be taken to improve care.

---

## Methods of the PSOCS Study

Given that PSOCS outcomes included assessments at 6 months post-discharge, a series of "interval events" was monitored, including both health-related and non-health-related life events during this post-discharge time period. The sequence and timing of post-discharge events were gathered to evaluate the potential for interactions between post-discharge healthcare and non-healthcare events upon risk-adjusted 6-month patient outcomes of care.

Importantly, a comprehensive array of patient-specific risk factors was gathered. Risk factors were classified in four dimensions assessed at baseline, including severity of cardiac disease, comorbidities (i.e., noncardiac diseases), demographic and socioeconomic factors, and health status evaluations performed by both the care

provider team and by the patients themselves (for both cardiac disease-specific and general health status domains). The risk factors were also analyzed to evaluate to what degree modifiable risk factors (e.g., patient's alcohol use, smoking, and exercise habits) had a differential impact as compared to the non-modifiable risk factors (e.g., the patient's age, gender, or race/ethnicity). Finally, a series of control variables was used (e.g., provider identifier, facility identifier, date/time sequencing variables) to coordinate the complex analyses required.

In total, there were 1,453 variables gathered for each PSOCS patient, including 249 outcome-related dependent variables (which were ultimately used to calculate three short-term and five intermediate 6-month outcomes) along with 1,102 independent variables (209 patient risk variables, 509 process-of-care variables, and 303 structure-of-care variables) and 23 interval events with 153 "control" variables used for analytical purposes. Across the 14 participating medical centers, the PSOCS study enrolled 3,988 patients during the period from 1992 to 1996, with follow-ups coordinated through early 1997 (O'Brien et al. 2004).

Due to the large number of variables, an initial task was data reduction, addressing the missing data and evaluating patterns of data completeness across surgeons and VA medical centers. Because intraoperative complications directly impacted outcomes, these were addressed analytically. As a first step, statistical risk models were built to predict the 30-day operative and 6-month outcomes. Within domains and coordinated in a nested analysis across sub-domains, the impact of processes of care upon risk-adjusted outcomes was evaluated. Specifically, processes of care related to operative duration (i.e., increased operative time), the use of inotropic agents, the use of transesophageal echocardiographic (TEE) monitoring and systemic temperature monitoring, and the use of hemoconcentration/ultrafiltration systems were powerful predictors of adverse composite outcomes. Since some of these processes of care may be initiated in response to adverse intermediate outcomes (e.g., intraoperative complications), a more complex analytical approach was



used to evaluate for the main effects (rather than interaction-related effects) for processes of care. Following these adjustments, the use of intraoperative transesophageal echocardiography and the use of hemoconcentration/ultrafiltration remained significantly associated with increased risk for an adverse outcome (O'Brien et al. 2004), which was likely driven by patient complexity.

---

## Findings of the PSOCS Study

An important finding of this study, unanticipated in the original PSOCS design, was that, retrospectively, it is extremely difficult to differentiate planned versus unplanned processes of care. Intermediate outcomes, such as intraoperative complications, may cause providers to initiate new processes, previously unplanned, to address unforeseen challenges. Thus, differentiating between a planned process of care (i.e., a process of care that would be generally initiated for all patients) versus an unplanned process of care (i.e., a process of care that was initiated in response to an unforeseen challenge) is a critically important distinction for meaningful quality assessments. Quite simply, capturing the *unplanned processes of care* may be – in and of itself – an important indicator as a quality metric. With this important concept documented by PSOCS, it became clear that the use of state-of-the-art techniques and equipment for monitoring may provide for the early identification of potential adverse events.

To facilitate future quality-related research, the PSOCS study successfully built upon the historical literature basis, denoting that inotropic use, transesophageal echocardiography use, and the use of hemoconcentration/ultrafiltration appear to potentially impact post-CABG risk-adjusted outcomes. The PSOCS found that there was a consistent relationship documented between key times (i.e., cardiopulmonary bypass time or operative time) and risk-adjusted adverse outcomes, for which there is an association with the surgeon-specific and/or facility-specific practices. Not surprisingly, therefore, the PSOCS study identified that processes (e.g., operative times) were

intertwined with structures of care (e.g., surgeon-specific years of experience). Moreover, the PSOCS study challenged the ability of research to isolate process-specific or structure-specific impacts on adverse risk-adjusted outcomes, as well as identified the need to differentiate unplanned versus planned processes of care, an important advancement forging forward the frontier of quality assessment. Finally, the PSOCS study documented that the statistical risk modeling approaches used may need to evolve, to be process- or structure specific, in order to identify the unique risk factors that emerged (e.g., a new intraoperative complication) directing the change from planned to unplanned approaches (O'Brien et al. 2004).

---

## The CICSP-X Program

Having recently completed the PSOCS study's data capture and preliminary analyses, the VA CICSP was dramatically expanded (entitled CICSP-X [as an expansion of CICSP], under the leadership of Dr. Shroyer) in 1997 as a clinical national quality improvement database to identify the interrelationships of risk factors with processes and structures of care, as well as to include a broader set of clinical outcomes (Shroyer et al. 2008). The CICSP-X program established the feasibility of coordinated multidimensional quality database reports to address a more comprehensive set of quality of care metrics, with a comprehensive “dashboard” of summary metrics reported for different quality of care dimensions, including a series of preestablished outcome metrics, as well as processes and structures of care measures.

In 1997, Department of Defense (DoD) and VA guidelines for Ischemic Heart Disease (IHD) became an impetus for additional changes to the VA Criteria and Standards, where new post-CABG hospital medication-use requirements were established (Veterans Health Administration and Department of Defense 1997). As a key processes of care measure, the CABG-only patients use of key evidence-based medical therapies was required for (1) lipid-lowering agents,

(2) beta-blockers for patients with a prior myocardial infarction, and (3) angiotensin-converting enzyme (ACE) or angiotensin II receptor blocker (ARB) medications for patients with a baseline low ejection fraction ( $\leq 40\%$ ). For CABG-only patients in high-risk subgroups, monitoring extended to additional guidelines, measuring compliance with standards including the use of diabetic agents for diabetic patients and antihypertensive medications for those with hypertension.

Due to the VA's extensive Pharmacy Benefits Management (PBM) program (and outstanding leadership of the Pharmacy Benefits Management enterprise), the rates of guideline-based medication use could be identified for a CABG-only patient based on their preoperative risk profile. Although limited to identification of medications filled via the VA pharmacy (medications filled at non-VA pharmacies could not be easily ascertained), the compliance rates for all of the guideline-required medications (using an "all-or-none" evaluation) were routinely coordinated to assess overall cardiac surgery program performance. By improving compliance with Department of Defense/VA guidelines, the goal was to improve long-term survival post-CABG surgery, as well as to optimize veterans' long-term health status and quality of life (Veterans Health Administration CARE-GUIDE Working Group et al. 1996).

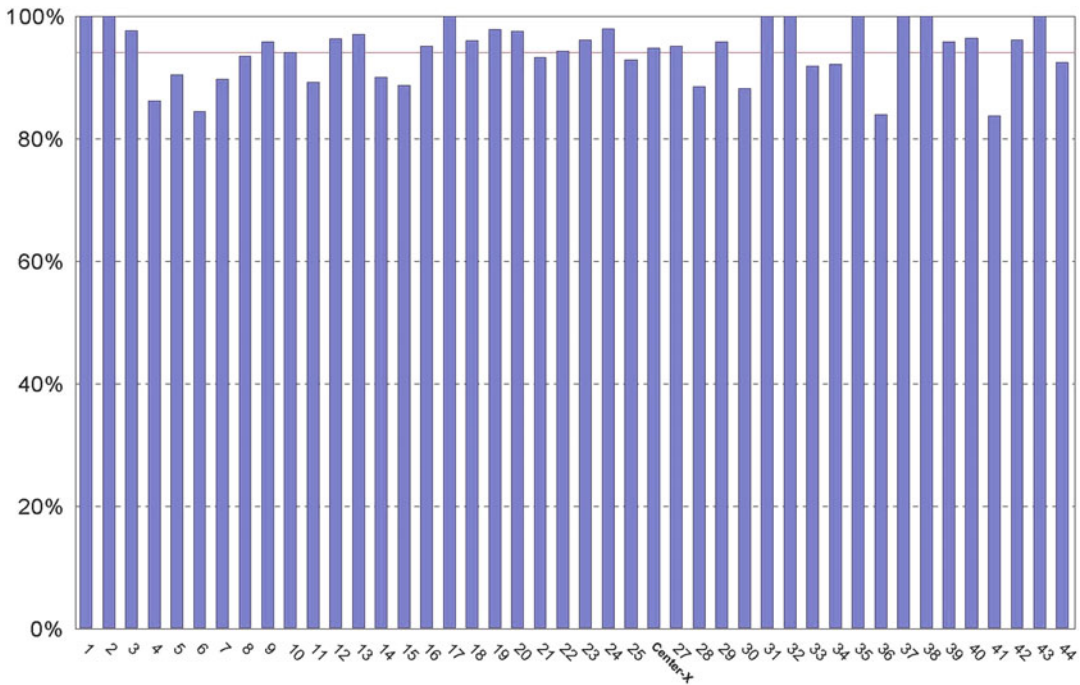
---

## Measuring Processes of Care

During the late 1990s, a wide variety of national watchdog agencies arose with the goal of providing quality of care oversight such as the Leapfrog initiative (Milstein et al. 2000). The National Quality Forum was developed (Miller and Leatherman 1999) and published a set of performance indicators that were intended to serve as internal quality improvement metrics (National Quality Forum 2004). At that time, the National Quality Forum metrics represented the best data-driven evidence (or in the cases where evidence is lacking, the best clinical consensus) about the optimal approaches to provide cardiac surgical

care to patients. As reference, these National Quality Forum quality metrics specified what would be anticipated "best practices" as well as established goals for surgeons to strive for in coordinating the care for their patients. For example, the use of internal mammary artery (IMA) conduits for a CABG graft placed to the left anterior descending artery (LAD) artery was generally preferred based on improved long-term survival rates, as well as reduced rates for repeat revascularization procedures. Since it may take slightly longer to take down the internal mammary artery, compared to harvesting a saphenous vein graft (SVG) conduit, this approach may not be advantageous for emergent patients. Similarly, elderly patients may not live long enough to document the internal mammary artery survival benefit (Ferguson et al. 2002). Based on the National Quality Forum standards combined with literature-based evidence and the feasibility of data to be captured, the CSCB identified as "best practice" the use of an internal mammary artery graft for CABG-only procedures, particularly emphasizing that this practice should be used for the subgroup of non-emergent, patients (e.g., elective and urgent cases). Starting in 2008, therefore, the VA Criteria and Standards for Cardiac Surgery Programs specified that a CSCB review would be performed for cardiac surgery programs that performed less than 80% of their CABG-only procedures using internal mammary artery grafts during a 6-month reporting period. Figure 4, a sample report, illustrates the variability in internal mammary artery graft use across VA medical centers. Within this 6-month reporting period, center "X" had a CABG-only procedure internal mammary artery graft use rate of  $>80\%$ . Hence, no quality reviews of center "X" would normally be required for this preestablished internal mammary artery graft use quality threshold.

In addition to assessing that the right processes of care were provided to the right patient, the VA CICS-P-X reports were expanded to also evaluate cardiac surgical resource utilization, toward the goal of improving the efficiency of the VA care provided (Shroyer et al. 2008). The resource utilization metrics included evaluating the rates of the same-day surgery, the preoperative length of




**Fig. 4** Example: rate of internal mammary artery graft use at Veterans Affairs Medical Centers

stay, the operating room times, the postoperative length of stay, and the total length of stay for the veterans served. Because some patients underwent preoperative cardiac catheterizations during the CABG hospitalization and others did not, these two groups were considered separately, since this difference could impact both the rates for same-day surgery and the total length of stay.

As an example of important resource use metrics routinely evaluated by CICSP historically, the proportion of patients with same-day surgery, the preoperative length of stay (both for patients with and without a cardiac catheterization procedure during the CABG hospitalization), the postoperative length of stay, and the total length of stay were monitored. For example, Fig. 5 (which is a sample report) illustrates the types of resource consumption profiles provided by center. Within this example 6-month reporting period, center “X” might have had several areas that were flagged for potential efficiency reviews to examine practices of discharge-related processes and structures of

care (e.g., early discharge planning and social work support systems).

Recent studies have attempted to further characterize the importance and utility of these types of resource utilization metrics. For example, the Virginia Cardiac Surgery Quality Initiative (VCSQI) database of over 42,000 patients undergoing CABG was recently analyzed to investigate the relationship between quality (as determined by various risk-adjusted measures of morbidity and mortality) and resource utilization (i.e., costs and length of stay) at individual hospitals. The VCSQI research team documented strong correlation between risk-adjusted morbidity and mortality with length of stay but not directly with costs. This appears to support the importance of these types of process of care and outcome measures in assessing the value of services rendered at cardiac surgical centers. Further, it was shown that both preoperative and postoperative factors (e.g., comorbidities and complications, respectively) influence both length of stay and costs, reinforcing the importance of healthcare quality

CICSP Cardiac Surgery Dashboard For All Centers Resource Use Measures						
Six-Month Report Period	Center	Percent Same Day Surgery (no cath) Figure R3	Pre-Op Length Of Stay without Cath (median days) Figure R6	Pre-Op Length Of Stay without Cath (median days) Figure R9	Total Post-Op Length of Stay (median days) Figure R10	Total Length of Stay (median days) Figure R11
FY07-2	1					
FY07-2	2					
FY07-2	3					
FY07-2	4					
FY07-2	5					
FY07-2	6					
FY07-2	7					
FY07-2	8					
FY07-2	9					
FY07-2	10					
FY07-2	Center-X					
FY07-2	12					
FY07-2	13					
FY07-2	14					
FY07-2	15					
FY07-2	16					
FY07-2	17					
FY07-2	18					
FY07-2	19					
FY07-2	20					
FY07-2	21					
		0% or Lower quartile	Upper quartile (longer)	> 1 day	Upper quartile (longer)	Upper quartile (longer)
		Mid range	Mid range	1 day	Mid range	Mid range
		Upper quartile	Lower quartile	Same day	Lower quartile	Lower quartile
 Centers in the Upper / Lower quartiles and Mid range are not outliers						

**Fig. 5** Example: Veterans Affairs coronary artery bypass grafting procedural resource consumption dashboard report

initiatives in containing the costs associated with healthcare and increasing the value of the care rendered (Osnabrugge et al. 2014a, b).

### Monitoring Trends Over Time

Across all processes of care, structures of care, resource use, and risk-adjusted outcomes, reports for the most recent 6-month period, trends over time

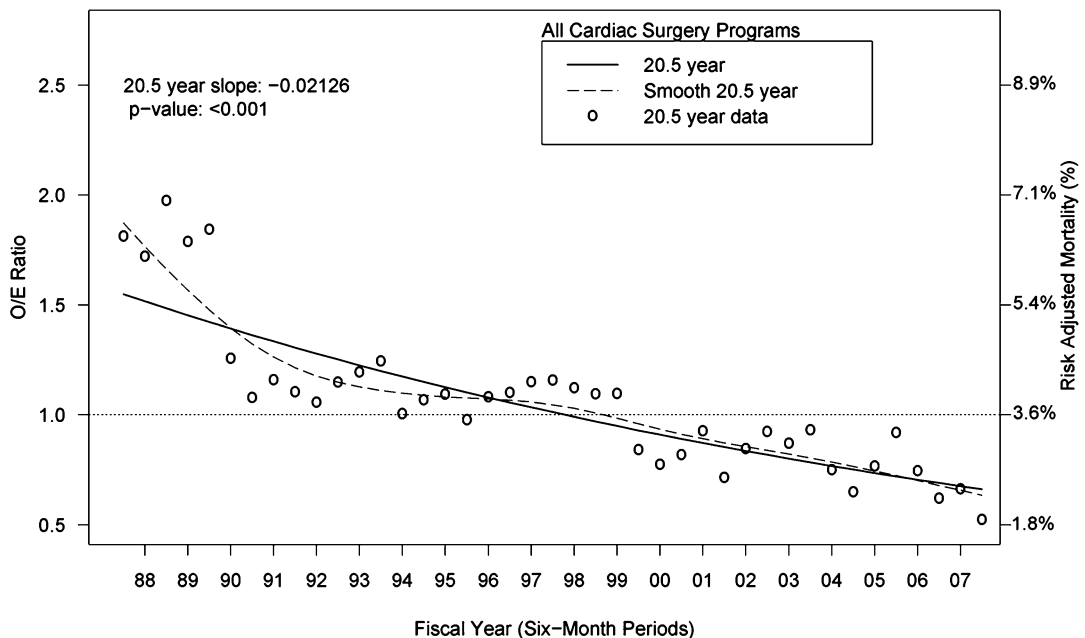
for the most recent 3-year period, and trends over time for the entire period monitored (from 1991 to the current reporting period) were coordinated. These “Time Series Monitors of Outcome” (TSMO) metrics were evaluated to identify if a cardiac surgery program might be a “high outlier,” “not an outlier,” or “low outlier” based on preestablished statistically driven thresholds (e.g., high and low outliers were generally more than two standard deviations beyond the mean).

Additionally, the trend line slope was evaluated for “upward” versus “downward” trending, versus “no trend identified.” The subgroup of VA cardiac surgery programs with upwardly oriented trends identified (i.e., a trend toward increasing adverse event rates or increased resource use or problems with guideline compliance) or “high-outlier” status (potential challenges in overall performance) was identified for intensive review, with potential site visits performed when these indicators clustered in a manner to raise potential quality of care concerns. Summary reports across all quality metrics (called “dashboards”) were developed, as the number of quality indicators increased. These dashboards provided a quick and easy identification of the subgroup of VA cardiac surgery programs with challenges identified. Similarly, a focus was placed on identifying exemplary performance, that is, when clusters of positive performance indicators were identified, particularly if positive trends over time were identified, as well as sustained positive performance over time (Marshall et al. 1998).

With the expanded focus on multidimensional quality reports, the original CICSP report had grown from six pages to over 200 pages. The use of dashboards addressed the information

overload by providing summaries of the findings identified in these detailed process, structure, outcome, and resource reports (Shroyer et al. 2008). Based on the dashboard reports, very busy VA Central Office leadership team members, regional directors, hospital directors, and local VA cardiac surgery program directors could coordinate informed data-driven decisions to address any challenges identified, as well as work proactively to improve future VA cardiac surgery program quality of care. Thus, as an infrastructure quality reporting resource, the VA CICSP-X program set forth a dashboard framework that continues today as part of the consolidated VA Surgical Quality Improvement Program (VA SQIP), setting the VA as a leader in identifying, monitoring, and reporting quality for cardiac surgical care. As an example of this, Fig. 6 documents that there was a statistically significant downward trend observed for 30-day CABG operative mortality (a 2.1% reduction) from 1988 to 2007, indicative of continuing improvements over time for the CABG-only in-hospital surgical care and early post-discharge care provided.

As the VA historically invested substantial support at both the national level (in the



**Fig. 6** Example: Veterans Affairs time series monitors of outcome summary report evaluating trends in observed/expected ratios over time

CICSP-X and NSQIP programs) and at the local level (for the local nurses or data coordinators used originally to gather the data required), it is important to pause to evaluate the return on this investment. Based upon VA findings to date, these quality improvement endeavors appear to have positively impacted short-term and longer-term rates of adverse cardiac surgical outcomes, with dramatic improvements and statistically significant downward-sloping trends in the mortality and morbidity rates over the 20+-year period reported (Grover et al. 2001; Shroyer et al. 2008). Based on the trends in risk-adjusted outcomes reported, moreover, these positive improvements do not seem to be related to the VA taking on easier cardiac surgical cases, as the risk profile for veterans basically remained the same (with the exception that the average age of the veterans served increased slightly over the period of time evaluated) (Shroyer et al. 2008). Moreover, the markers of VA efficiency similarly documented substantial improvements, with same-day surgery rates rising from 0% (1987) to 40% (1997).

Although no causal impact could be identified (as many changes in both surgical practices and medical management of ischemic heart disease occurred during these same periods), these positive trends in risk-adjusted outcomes support the continuation of quality improvement efforts and the expansion of these programs beyond cardiac surgical patient care.

---

## Implementation of National Quality Improvement Programs

Under the leadership and guidance of Dr. Richard E. Clark, the Society of Thoracic Surgeons (STS) initiated the National Adult Cardiac Surgery Database (ACSD) in February 1991 with 330 surgeon members at 81 centers throughout the United States participating initially in this quality improvement endeavor (The Society of Thoracic Surgeons 2014c). Although the original goal was to initiate databases also for Congenital Heart Surgery (CHSD) and General Thoracic Surgery (GTSD), the development of these two databases was delayed until the full implementation of the

STS Adult Cardiac Surgery Database was successfully coordinated.

As background, the purpose of the STS Adult Cardiac Surgery Database was to gather data on mortality, morbidity, and resource-use outcomes, as well as patient risk factors, to allow the evaluation of risk-adjusted cardiac surgical outcomes across providers and to report trends over time. By 1995, Dr. Clark had reported that the Adult Cardiac Surgery Database had grown to include 1500 surgeons at 706 centers across 49 states, with decreasing postoperative length of stay trends documented and modest reductions in operative mortality rates in spite of increasing patient risk over time (Clark 1995).

By the late 1990s, a wide variety of STS initiatives had been coordinated related to the enhancement of the Adult Cardiac Surgery Database and the initiation of the Congenital Heart Surgery and General Thoracic Surgery endeavors. The STS databases were distributed to the participants by means of licensed software products via vendors, with centralized database management, analysis, and reporting functions coordinated by the Duke Clinical Research Institute (DRCI) team. Long-term goals were preestablished for the STS databases to become the main repositories to support improvements in local clinical decision-making, cardiac surgery program management, and policy decisions. Toward these goals, expansions of the existing database data forms and definitions were expanded to ensure that 1595 future comparisons might be coordinated across a broader array of outcomes (e.g., health-related quality of life, functional status, longer-term survival, and costs of care). Additionally, comparisons of cardiac surgical procedures to alternative treatments (e.g., cardiology-based interventions, such as the placement of stents) were planned.

By the early 2000s, the STS Adult Cardiac Surgery Database was viewed as the largest clinical repository of data available in the country, used to guide both health policy discussions and debates on reimbursement at congressional hearings. Database reports were generated semiannually, with local site reports compared to regional and national profiles. As STS National Database

Committee members, Drs. Bruce Keogh (United Kingdom) and Paul Sargent (Belgium) worked with their European colleagues to build upon the STS Adult Cardiac Surgery Database structure a new European Association for Cardio-Thoracic Surgery Adult Cardiac Surgery Database (EACTS), transforming the STS template into a structure that could be used to support quality improvement efforts globally. As of late 2008, this database was reported to include over one million patient records from 366 hospitals across 29 countries in Europe (Nashef et al. 1999; Head et al. 2013).

The STS worked with the National Quality Forum and the American Medical Association's Performance Improvement Physician's Consortium to coordinate new quality of care metrics for national reporting from 1999 to 2001. These external collaborations, beyond the STS-based quality reporting endeavors, were very important to establish the external credibility of the STS Adult Cardiac Surgery Database. Even today, the National Quality Forum metrics reported for adult cardiac surgery include the STS Adult Cardiac Surgery Database-based metrics used widely in program-based quality of care assessments (The Society of Thoracic Surgeons 2014a).

Focused upon the importance of high-quality, accurate, and reliable STS data to generate reports, the STS Adult Cardiac Surgery Database Committee (chaired by Dr. Rich Prager) began a new quality improvement process in 2006, randomly selecting STS participating sites to audit and validate the number of cardiac surgical records and outcomes submitted by participating surgeons and sites. For a random sampling of Adult Cardiac Surgery Database participating sites from 2007 to 2013, each audited sites' submitted risk, operative procedure, and outcome data were compared with data obtained independently by an external audit company. The number of Adult Cardiac Surgery Database sites audited increased from 24 in 2007 (3% of sites) to 86 in 2013 (8% of sites). Over 92% of audited STS sites provided positive audit feedback, noting that the audit process had positively impacted their data accuracy. Across all risk, process of care, and outcome variable categories, the aggregate

agreement rates ranged from 94.5% (2007) to 97.2% (2012), with improvements in the variable-specific agreement trends over time. Although the operative mortality agreement rate was reportedly lower in earlier years, the rate of reliability for death reporting has consistently remained above 95% since 2008. The STS external audit process established that Adult Cardiac Surgery Database data integrity is high, with data concordance reported at 97.2% (2012). By means of this external audit process, the STS Adult Cardiac Surgery Database can be interpreted with confidence, with independent external auditor verifications confirming that the data submitted by STS participating surgeons and centers is of the highest integrity (Winkley Shroyer et al. 2015, Member of STS Adult Cardiac Surgery Database Workgroup, "personal communication").

---

## Uncovering Quality Trends

Important quality improvement trends over time have been documented using the STS Adult Cardiac Surgery Database, including procedure-specific or population-specific reductions in the rate of adverse events reported. Overall rate of reoperations and correspondingly the rate of 30-day operative death have been documented to be diminishing (6.0% down to 3.4% and 6.1% down to 4.6%, respectively) over the 10-year period of 2000–2009 (The Society of Thoracic Surgeons 2014c). Importantly, the field of cardiothoracic surgery has documented substantial quality improvements over time, with diminishing rates of mortality and morbidity (Ferguson et al. 2002). As noted by Dr. Ferguson, remarkable strides to improve cardiac surgical care have been initiated by the surgeons (e.g., the use of new techniques for improved myocardial perseveration) and the pharmaceutical industry (providing new medications). Other improvements include the implementation of care pathways, the formation of cardiac surgery dedicated teams (e.g., including a dedicated cardiac anesthesiologist), better approaches used for patient selection, as well as innovations to improve the efficiency of care (e.g., "fast-track" cardiac surgery early

extubation protocols). Even though the population of cardiac surgical patients has grown older and sicker over time, risk-adjusted outcomes have improved. Another major change over time was the growing reliance of the STS Adult Cardiac Surgery Database by key national US-based decision-makers, including legislators. The STS Adult Cardiac Surgery Database was used to identify, monitor, report, and target future cardiac surgical improvements, shifting the national quality debates from a conceptual framework to data-driven patient care, program management, and policy discussions (Ferguson et al. 2002).

As a major transformation to multidimensional quality metrics, the STS has led the way in the development of composite scores, which were adopted by the National Quality Forum as new quality metrics in 2008. Specifically, Dr. David Shahian and the STS National Database Committee worked to coordinate an STS coronary artery bypass graft (CABG) composite score. The composite score was comprised of risk-adjusted mortality, risk-adjusted morbidity, a surgeon-related process of care metric (i.e., the use of the internal mammary artery as a conduit), and a facility-related process of care metric (i.e., the use of beta-blocker medications perioperatively) (O'Brien et al. 2007). In combination, these multidimensional composite metrics are used to categorize STS facilities and surgeons into "star ratings" for quality, based on a three-star, two-star, and single-star rating system, differentiating high-versus low-quality centers based on the composite metric (The Society of Thoracic Surgeons 2014b). Based on the success of the CABG-only composite score, an isolated aortic valve replacement (AVR) composite score was designed and implemented in 2012, as well as a combined aortic valve replacement-CABG composite score in 2014.

Most recently, the STS has added new modules to enhance focused quality endeavors for high-risk patient subgroups. For example, a new module related to prophylaxis and treatment of cardiac surgery patients that experience atrial fibrillation was added. As atrial fibrillation is a very common post-cardiac surgical complication, its prevention and early treatment is an important quality

consideration. Studies on atrial fibrillation have demonstrated that certain prophylactic measures (e.g., amiodarone, beta-blockers, magnesium, atrial pacing) do significantly reduce the rate of postoperative atrial fibrillation after cardiac surgery, as well as shorten hospital stays and decrease the cost of hospital care by over \$1,200. No significant effects on mortality or the incidence of stroke have been demonstrated, however (Arsenault et al. 2013). Similarly, a new module related to documentation of the details of cardiac anesthesiology was added in July 2013 to identify the anesthesiology-related processes of care that may be targeted for future quality improvement initiatives (The Society of Thoracic Surgeons 2013). Most importantly, the focus on STS cardiac and thoracic procedural outcomes has been shifted to evaluate long-term outcomes, such as long-term survival. Toward this goal, database matches with the national death registry were performed, with the first long-term follow-up risk models predicting survival completed in 2012.

---

### **The Michigan Society of Thoracic and Cardiovascular Surgeons Quality Collaborative**

The Michigan Society of Thoracic and Cardiovascular Surgeons Quality Collaborative (MSTCVS-QC), as an example of a regional STS initiative, is led by Dr. Richard Prager. The MSTCVS-QC is a consortium of 33 cardiac surgery programs throughout the state of Michigan focused on identifying intraoperative and postoperative opportunities to improve the quality of cardiac surgical care. As one of their recent endeavors, they examined the use of blood transfusions as a potential quality of care metric, examining the relationship between blood product use and clinical outcomes. The MSTCVS-QC found that quality collaborative educational approaches may have very positive impacts, as the blood product utilization was documented to decrease dramatically after routine quarterly reporting of program-identified transfusion rates was implemented. The quarterly MSTCVS-QC incorporated very frank



discussions about the potential adverse effects (i.e., increased risk of mortality and morbidity) associated with transfusions. Under the leadership of Dr. Prager, the Michigan team's persistent and continued focus on this topic has dramatically revised clinical practice and enhanced blood product conservation approaches used throughout Michigan State (Paone et al. 2013).

Another MSTCVS-QC recent endeavor looked at how to reduce hospital-acquired infections (HAI) related to CABG procedures. Hospital-acquired infections include complications such as pneumonia, sepsis, septicemia, wound-related infections, as well as other infections reported. As of early 2008, Medicare has not reimbursed hospitals for post-CABG mediastinitis-related treatments, as infections (such as mediastinitis) are perceived to be directly related to a lower quality of surgical care provided during the initial CABG hospitalization. Interestingly, Dr. Prager and his MSTCVS-QC colleagues found that on average 5.1% of CABG patients developed hospital-acquired infection postoperatively. Moreover, there was a tremendous variation in the reported rates of post-CABG hospital-acquired infections (ranging from 0.9% to 19.1%). Differences in cardiac surgery program-based patient risk characteristics did not account for much of this dramatic difference in program-based hospital-acquired infection rates observed. Within this analysis, four centers appeared to be high outliers (i.e., had a hospital-acquired infection O/E ratio that was statistically significantly higher than 1.0). Based on in-depth evaluations of the CABG care rendered at these four "high-outlier" centers, the MSTCVS-QC team concluded that the largest variations were found for pneumonia and multiple infection end points. Based on their reviews, they thought a multidisciplinary care team approach was needed to address the challenges identified, ideally to bridge across traditional specialty-based silos of care, facilitating future heart patient team-based care approaches in the future. Working collaboratively as an STS regional society, therefore, the MSTCVS-QC team provides research on quality improvements that extend beyond the STS Adult Cardiac Surgery Database capabilities, enhancing the data-

driven approaches used to assess and to improve cardiac surgical patient's quality of care (Shih et al. 2014).

---

### **The American College of Surgeons' Private Sector Initiative**

As a separate endeavor, the American College of Surgeons (ACS) coordinated an NSQIP Private Sector initiative, building upon the VA-based historical work by Dr. Shukri Khuri's team. The first step in this process was a feasibility study conducted in 1999 at three non-VA hospitals (University of Kentucky, University of Michigan, and Emory University) (Fink et al. 2002). Based on the initial success of this feasibility project, the NSQIP was expanded in 2001 to include 18 centers as part of a pilot project funded by the Agency for Healthcare Research and Quality (AHRQ) (Hall et al. 2009). Subsequently, the American College of Surgeons' NSQIP pilot was expanded in 2004 to include other private hospitals' reporting.

As background, the VA-based NSQIP had been documented to improve risk-adjusted mortality and morbidity across a diversity of surgical disciplines. For the period from 1991 to 2004, the surgical 30-day operative mortality rate improved by 31%, and the surgical 30-day perioperative major morbidity rate improved by 45% (Khuri 2005). During this time period, the VA NSQIP findings reported were deemed to be the "best in the nation" by the Institute of Medicine in 2003 for evaluating the quality of surgery across a broad range of surgical specialties (Khuri 2005). The "Patient Safety in Surgery" (PSS) study was initiated during 2001–2004 to evaluate the impact of a uniform quality improvement system and to compare VA and non-VA-based outcomes of care (where care-related details were gathered contemporaneously using a standardized set of data forms, definitions, and analyses). With nearly 185,000 surgical patient records gathered across 128 VA medical centers and 14 private sector hospitals, there were significant differences in the types of surgical procedures performed and patient baseline risk characteristics across the VA

versus non-VA hospitals. In spite of these differences in patient risk factors and procedures performed, the O/E ratios for 30-day operative death were remarkably similar between the VA and non-VA facilities (correlation coefficient = 0.98). Similar to the VA trends identified earlier, the non-VA private sector hospitals had an 8.7% decrease in major perioperative complications over the 3 years of the study, documenting an important and substantive quality improvement (Khuri et al. 2008).

The Agency for Healthcare Research and Quality provided a grant to Dr. Khuri's team, based in part upon these promising findings, to evaluate the "Structures and Processes of Surgical Care Study" in late 2003 to relate the processes and structures of surgical care to postoperative risk-adjusted outcomes. For this NSQIP-based endeavor, surveys were sent out to the 123 VA sites and 14 private sector sites that participated in the Patient Safety in Surgery study earlier. The survey included many questions, but specifically asked for information as to the organization of the preoperative, intraoperative, and postoperative care services. Additionally, there was information gathered on hospital-specific surgical program-based characteristics such as surgical program size, surgeon-specific volumes at the VA and non-VA affiliates, patterns in surgical staffing ratios, the nature of the organizational structure, the use of local facility-based quality improvement efforts, the types of novel equipment/technology available (e.g., ultrasonography used in the operating room, the use of a harmonic scalpel, the use of radio-frequency ablation, or availability of ultrasound-guided aspiration devices), available information systems, the use of coordination/communication processes, as well as residency training program characteristics. The published results from the VA-based surveys (with responses sent back by the local Chiefs of Surgical Service) identified that there were tremendous variations in the processes and structures of general surgical care. As documented by the descriptive survey findings, the process and structure variables that appeared to be associated with risk-adjusted morbidity (14 variables) and risk-adjusted mortality (four variables) were

preliminarily identified. Specifically, a higher O/E ratio (a potential marker for quality of metrics concerns) was found to be associated with several factors including anesthesia organized as a separate service, a larger number of operating rooms, more frequent reports of short staffing, and a higher rate for staff surgeons to be paid in part by the affiliated medical center. As a key process of care identified, changes in the anesthesia provider during the case (i.e., from across the pre-, intra-, and postoperative time periods) were associated with worse risk-adjusted mortality rates. A negative relationship between surgical volume (e.g., fewer cases per surgeon per month) and risk-adjusted morbidity (e.g., higher rates of perioperative complications) was identified. Overall, the self-reported survey findings for processes and structures of care appeared to be more strongly associated with the risk-adjusted morbidity rates observed, rather than risk-adjusted mortality rates documented. Importantly, the VA self-survey findings identified that a more integrated surgical service appeared to improve communication and coordination of surgical care, as well as the effectiveness of surgical team performance. Thus, these preliminary survey findings provided an impetus for the documentation of surgery-specific processes and structures of care, as well as the development of a more comprehensive set of quality metrics that are currently evaluated by NSQIP (Main et al. 2007).

In 2004, the Private Sector Study (conducted at 14 academic non-VA hospitals) was expanded and opened to other private sector hospitals. By 2008, the American College of Surgeons' NSQIP market penetration for private hospitals included over 200 facilities with diverse characteristics located throughout the United States. The initial evaluation of the first 3 years (2005–2007) documented dramatic improvements in quality of surgical care rendered, with 66% of the hospitals documenting improved risk-adjusted mortality rates and 82% of the hospitals documenting improved risk-adjusted morbidity rates. In spite of the increasing patient risk characteristics reported (e.g., average patient age increased over time), the results were impressive, with 9,598 potential complications avoided at 183 private sector hospitals (Hall et al. 2009).

Although many factors likely contributed to these important and positive changes, the use of a data-driven quality improvement initiative was identified as a major factor that appeared to lead to better outcomes, cost savings, as well as improvements in safety across patient subgroups (Maggard-Gibbons 2014). Several publications were coordinated evaluating the usefulness of different types of process and structural interventions. Reducing the rate of adverse clinical outcomes, the documented set of effective interventions included the use of protocols to manage postoperative blood glucose for diabetic patients, the use of venous thrombosis risk evaluations for high-risk patient subgroups, standardized approaches for wound care management, the use of physician order entry templates, the helpfulness of clinical pathways (e.g., a standardized approach to remove Foley catheters), enhanced tracking, and the use of more detailed patient tracking/monitoring tools for postoperative pulmonary management. Hence, changes in Medicare payment reforms were initiated to provide positive reimbursement incentives for surgeons and hospitals to participate in national quality improvement reporting endeavors such as ACS-NSQIP and the STS national database endeavors. Most importantly, the use of *clinical databases developed by surgeons for surgeons' use in self-assessment and self-improvement endeavors* gained momentum; with clinician-leaders rising to the ranks of government organizations (e.g., Dr. Jeff Rich, a cardiothoracic surgeon taking on a top-level leadership role with the Centers for Medicare and Medicaid Services) to advance the science of quality measurement and management.

---

### Implementation Challenges: Dilemmas Faced by Quality Measurement Projects

In evaluating the optimal quality metric or set of metrics to use for a project, researchers must consider many factors. The purpose of the project as well as the type of questions raised will direct which types of assessments are most important (e.g., process, structure, and/or outcomes). To

evaluate outcomes, there must be a plausible conceptual relationship (if not actual data) that would identify any other quality of care factors that could be associated with the outcomes selected for evaluation.

Different clinical fields are at different stages of maturation in selecting the “best” quality metrics. For surgical services, it has been demonstrated that the use of processes, structures, and risk-adjusted outcomes (as a comprehensive set of quality metrics) would be the most appropriate to consider. In other fields (e.g., psychiatry), however, simply defining the frequency of a broad array of clinical outcomes (along with the variety of risk factors that may be related to these outcomes) may be a more appropriate starting place for a project.

A good outcomes assessment instrument should be:

- **Valid** (reflect variations in quality that are consistent with expectations)
- **Reliable** (have reproducible findings across multiple raters for similar assessments of quality of care)
- **Timely** (measure a sufficient time 2057 sequence to evaluate the impact of medical 2058 care provided)
- **Sensitive to change** (reflect changes associated with the care impacts provided)
- **Feasible to implement** (reasonable to capture given time and cost constraints)
- **Clinically relevant** (reflect “best practice” and be useful to guide clinical decisions and/or actions)

(MacDermid et al. 2009). Additionally, the accurate documentation of risk factors is critical to allow risk-adjusted outcomes for meaningful comparison across provider subgroups, facilities, or patient subgroups (Shahian et al. 2004).

Although many advancements have been made in identifying approaches to implement Dr. Donabedian’s triad for assessing quality of patient care, many challenges remain that cause difficulties in achieving these goals. Specifically, there are issues related to handling missing data (Hamilton et al. 2010; Parsons et al. 2011). Although different statistical approaches can be

used to address missing data challenges, the distribution of missing data is unlikely to be random. Based on the nature and distribution of the missing data, therefore, it may be appropriate to clinically substitute specific values. For example, substituting negative findings for missing complications may be appropriate, as the medical chart does not uniformly document complications that did not occur. Pending the need for a statistical imputation approach, there are ways to reduce uncertainty associated with imputation. Whatever the approach used, the assumptions and methodological details should be documented. Where possible, sensitivity analyses should be conducted to evaluate the impact of the different imputation approaches upon the study-specific findings (as well as potential decisions to be drawn from these findings) (Hamilton et al. 2010).

Another challenge that arises in quality of care assessments is differentiating between planned and unplanned processes or structures of care, as well as to what degree these processes were coordinated in response to interim outcomes. For example, Dr. Guyatt and his team conducted a systematic review and meta-analysis of the factors associated with unplanned readmission for randomized, controlled, clinical trials of heart failure interventions (Gwadry-Sridhar et al. 2004). They found that targeted heart failure patients who received an educational intervention experienced a significantly decreased rate of unplanned hospital readmissions. As part of their review and analyses, they identified that unplanned readmission (as an adverse process of care that occurred relatively infrequently following targeted heart failure interventions) was a potential quality of care metric that was clinically relevant to monitor. However, unplanned readmission for congestive heart failure patients who received targeted educational interventions did not correspond with a decrease in longer-term patient survival (in the 6 months to 1 year post-intervention period). Thus, appropriate treatments coordinated at the time of the unplanned readmission may have mitigated any adverse impact upon the longer-term survival end point. In summary, unplanned processes of care that occur may be related to interim outcomes and

may not necessarily result in adverse longer-term outcomes.

Additional difficulties in evaluating quality of patient care may be related to the uncertainty in documenting the sequence and timing of events. As a case in point, the NSQIP database was used to evaluate the impact of the timing of major perioperative complications upon mortality. Interestingly, early wound infections resulted in a higher risk of mortality, in spite of adjusting for patient risk factors and other complication burdens. Somewhat surprisingly, the early occurrence of cardiac arrest or unplanned intubation was associated with lower risk of mortality after adjustment for other factors. However, late occurrence of pneumonia, acute myocardial infarction, or cerebrovascular accident was associated with higher risk of mortality (Wakeam et al. 2014). Although these study findings were preliminarily based on NSQIP database records, the timing and sequence of perioperative complications does appear to matter when identifying the interrelationships of different adverse events, such as complications and mortality.

Finally, there are many factors that impact patient longer-term outcomes including both medical events and nonmedical factors that occur after the main medical intervention studied. Specifically, the VA PSOCs study evaluated the factors that influenced 6-month mortality and 6-month health-related quality of life (Rumsfeld et al. 2001, 2004). The variations in the occurrence of interval events following post-CABG discharge, including both medical and nonmedical life events, were substantial. Similarly, Dr. Murphy and colleagues found that living alone following CABG surgery was a major risk factor for readmission, when such solitary patients were compared to those who were married or lived with others (Murphy et al. 2008).

---

## Summary

In summary, the goal of improving quality of care is an elusive one. The end point may appear to be in sight but, like a distant horizon, it cannot be reached. Great achievements have been

accomplished in implementing Dr. Donabedian's framework, particularly in the cardiac surgery and general surgery fields. This process for defining, measuring, and improving the quality of patient care is the mechanism that advances best practices and approaches optimum outcomes.

In a pluralistic society, the top priorities for quality of care initiatives are often difficult to ascertain. Clinical outcomes may not correspond with patients' self-reported outcomes, and the demand for cost containment may conflict with both.

Future electronic medical record systems (with a greater proportion of encoded data elements) may provide enhanced information, and statistical data reduction techniques combined with more sophisticated risk modeling analyses may identify the details for the best practices to improve patient outcomes. The next generation of clinicians and scientists will advance the frontier, with multidisciplinary, collaborative investigative teams leading the way. Ultimately, the focus may be expanded beyond the simplistic avoidance of major adverse events to encompass more subtle aspects of healing and health.

**Acknowledgments** This book chapter was supported, in part, by the Offices of Research and Development at the Northport and the Eastern Colorado Health Care System, Department of Veterans Affairs Medical Centers, as well as by the Stony Brook University School of Medicine's Department of Surgery and the Stony Brook University Health Science Center Library. Additionally, special thanks are extended to Ms. Sarah Miller (University of Colorado at Denver), Ms. Carol Wollenstein (Nursing Editor, Stony Brook University), and Ms. Jennifer Lyon (Reference Librarian, Stony Brook University) for their proofreading and editorial assistance.

## References

- Blackstone EH. Comparing apples and oranges. *J Thorac Cardiovasc Surg.* 2002;123(1):8–15.
- Blumberg MS. Risk adjusting health care outcomes: a methodologic review. *Med Care Rev.* 1986;43(2):351–93.
- Codman EA. The classic: the registry of bone sarcomas as an example of the end-result idea in hospital organization. 1924. *Clin Orthop Relat Res.* 2009;467(11):2766–70.
- Donabedian A. Explorations in quality assessment and monitoring vol. 1. The definition of quality and approaches to its assessment. Ann Arbor: Health Administration Press; 1980.
- Donabedian A. Criteria and standards for quality assessment and monitoring. *QRB Qual Rev Bull.* 1986;12(3):99–108.
- Donabedian A. The quality of care. How can it be assessed? *JAMA.* 1988;260(12):1743–8.
- Donabedian A, Wheeler JR, Wyszewianski L. Quality, cost, and health: an integrative model. *Med Care.* 1982;20(10):975–92.
- Ferguson TB Jr, Hammill BG, Peterson ED, DeLong ER, Grover FL, S. T. S. N. D. Committee. A decade of change – risk profiles and outcomes for isolated coronary artery bypass grafting procedures, 1990–1999: a report from the STS National Database Committee and the Duke Clinical Research Institute. *Society of Thoracic Surgeons. Ann Thorac Surg.* 2002;73(2):480–9; discussion 489–90.
- Grover FL, Hammermeister KE, Burchfiel C. Initial report of the Veterans Administration Preoperative Risk Assessment Study for Cardiac Surgery. *Ann Thorac Surg.* 1990;50(1):12–26; discussion 27–18.
- Grover FL, Shroyer AL, Hammermeister K, Edwards FH, Ferguson Jr TB, Dziuban Jr SW, Cleveland Jr JC, Clark RE, McDonald G. A decade's experience with quality improvement in cardiac surgery using the Veterans Affairs and Society of Thoracic Surgeons national databases. *Ann Surg.* 2001;234(4):464–72; discussion 472–464.
- Khuri SF, Henderson WG, Daley J, Jonasson O, Jones RS, Campbell Jr DA, Fink AS, Mentzer Jr RM, Neumayer L, Hammermeister K, Mosca C, Healey N, S. Principal Investigators of the Patient Safety in Surgery. Successful implementation of the Department of Veterans Affairs' National Surgical Quality Improvement Program in the private sector: the Patient Safety in Surgery study. *Ann Surg.* 2008;248(2):329–36.
- O'Brien MM, Shroyer AL, Moritz TE, London MJ, Grunwald GK, Villanueva CB, Thottapurathu LG, MaWhinney S, Marshall G, McCarthy Jr M, Henderson WG, Sethi GK, Grover FL, Hammermeister KE, S. Va Cooperative Study Group on Processes and S. Outcomes of Care in Cardiac. Relationship between processes of care and coronary bypass operative mortality and morbidity. *Med Care.* 2004;42(1):59–70.
- Shroyer AL, London MJ, VillaNueva CB, Sethi GK, Marshall G, Moritz TE, Henderson WG, McCarthy Jr MJ, Grover FL, Hammermeister KE. The processes, structures, and outcomes of care in cardiac surgery study protocol. *Med Care.* 1995;33(10 Suppl):OS17–25.
- Shroyer AL, McDonald GO, Wagner BD, Johnson R, Schade LM, Bell MR, Grover FL. Improving quality of care in cardiac surgery: evaluating risk factors, processes of care, structures of care, and outcomes. *Semin Cardiothorac Vasc Anesth.* 2008;12(3):140–52.

Sobolev B, Fradet G. Delays for coronary artery bypass surgery: how long is too long? *Expert Rev Pharmacoecon Outcomes Res.* 2008;8(1):27–32.

## Further Readings

- Arsenault KA, Yusuf AM, Crystal E, Healey JS, Morillo CA, Nair GM, Whitlock RP. Interventions for preventing post-operative atrial fibrillation in patients undergoing heart surgery. *Cochrane Database Syst Rev.* 2013;1, CD003611.
- Blackstone EH. Breaking down barriers: helpful breakthrough statistical methods you need to understand better. *J Thorac Cardiovasc Surg.* 2001;122(3):430–9.
- Clark RE. The STS Cardiac Surgery National Database: an update. *Ann Thorac Surg.* 1995;59(6):1376–80; discussion 1380–1371.
- Dowsey MM, Petterwood J, Lisik JP, Gunn J, Choong PF. Prospective analysis of rural–urban differences in demographic patterns and outcomes following total joint replacement. *Aust J Rural Health.* 2014;22(5):241–8.
- Ferguson TB Jr, Coombs LP, Peterson ED. Internal thoracic artery grafting in the elderly patient undergoing coronary artery bypass grafting: room for process improvement? *J Thorac Cardiovasc Surg.* 2002;123(5):869–80.
- Fihn SD, Gardin JM, Abrams J, Berra K, Blankenship JC, Dallas AP, Douglas PS, Foody JM, Gerber TC, Hinderliter AL, King 3rd SB, Kligfield PD, Krumholz HM, Kwong RY, Lim MJ, Linderbaum JA, Mack MJ, Munger MA, Prager RL, Sabik JF, Shaw LJ, Sikkema JD, Smith Jr CR, Smith Jr SC, Spertus JA, Williams SV, Anderson JL, F. American College of Cardiology Foundation/American Heart Association Task. 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS guideline for the diagnosis and management of patients with stable ischemic heart disease: a report of the American College of Cardiology Foundation/American Heart Association task force on practice guidelines, and the American College of Physicians, American Association for Thoracic Surgery, Preventive Cardiovascular Nurses Association, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons. *Circulation.* 2012;126(25):e354–471.
- Fink AS, Campbell Jr DA, Mentzer Jr RM, Henderson WG, Daley J, Bannister J, Hur K, Khuri SF. The National Surgical Quality Improvement Program in non-veterans administration hospitals: initial demonstration of feasibility. *Ann Surg.* 2002;236(3):344–53; discussion 353–344.
- Garbuz DS, Xu M, Duncan CP, Masri BA, Sobolev B. Delays worsen quality of life outcome of primary total hip arthroplasty. *Clin Orthop Relat Res.* 2006;447:79–84.
- Goldman S, Zadina K, Moritz T, Ovitt T, Sethi G, Copeland JG, Thottapurathu L, Krasnicka B, Ellis N, Anderson RJ, Henderson W, V. A. C. S. Group. Long-term patency of saphenous vein and left internal mammary artery grafts after coronary artery bypass surgery: results from a Department of Veterans Affairs Cooperative Study. *J Am Coll Cardiol.* 2004;44(11):2149–56.
- Gwady-Sridhar FH, Flintoft V, Lee DS, Lee H, Guyatt GH. A systematic review and meta-analysis of studies comparing readmission rates and mortality rates in patients with heart failure. *Arch Intern Med.* 2004;164(21):2315–20.
- Hall BL, Hamilton BH, Richards K, Bilimoria KY, Cohen ME, Ko CY. Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: an evaluation of all participating hospitals. *Ann Surg.* 2009;250(3):363–76.
- Hamilton BH, Ko CY, Richards K, Hall BL. Missing data in the American College of Surgeons National Surgical Quality Improvement Program are not missing at random: implications and potential impact on quality assessments. *J Am Coll Surg.* 2010;210(2):125–39, e122.
- Head SJ, Howell NJ, Osnabrugge RL, Bridgewater B, Keogh BE, Kinsman R, Walton P, Gummert JF, Pagano D, Kappetein AP. The European Association for Cardio-Thoracic Surgery (EACTS) database: an introduction. *Eur J Cardiothorac Surg.* 2013;44(3):e175–80.
- Hiratzka LF, Eagle KA, Liang L, Fonarow GC, LaBresh KA, Peterson ED, C. Get With the Guidelines Steering. Atherosclerosis secondary prevention performance measures after coronary bypass graft surgery compared with percutaneous catheter intervention and nonintervention patients in the Get With the Guidelines database. *Circulation.* 2007;116(11 Suppl):I207–12.
- Itani KM. A celebration and remembrance. *Am J Surg.* 2009a;198(5 Suppl):S1–2.
- Itani KM. Fifteen years of the National Surgical Quality Improvement Program in review. *Am J Surg.* 2009b;198(5 Suppl):S9–18.
- Karthik S, Fabri BM. Left internal mammary artery usage in coronary artery bypass grafting: a measure of quality control. *Ann R Coll Surg Engl.* 2006;88(4):367–9.
- Khuri SF. The NSQIP: a new frontier in surgery. *Surgery.* 2005;138(5):837–43.
- Khuri SF, Daley J, Henderson W, Hur K, Demakis J, Aust JB, Chong V, Fabri PJ, Gibbs JO, Grover F, Hammermeister K, Irvin 3rd G, McDonald G, Passaro Jr E, Phillips L, Scamman F, Spencer J, Stremple JF. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. National VA Surgical Quality Improvement Program. *Ann Surg.* 1998;228(4):491–507.
- Lee PJ, MacLennan A, Naughton NN, O'Reilly M. An analysis of reintubations from a quality assurance database of 152,000 cases. *J Clin Anesth.* 2003;15(8):575–81.

- MacDermid JC, Grewal R, MacIntyre NJ. Using an evidence-based approach to measure outcomes in clinical practice. *Hand Clin.* 2009;25(1):97–111, vii.
- Mack MJ. If this were my last speech, what would I say? *Ann Thorac Surg.* 2012;94(4):1044–52.
- Maggard-Gibbons M. The use of report cards and outcome measurements to improve the safety of surgical care: the American College of Surgeons National Surgical Quality Improvement Program. *BMJ Qual Saf.* 2014;23(7):589–99.
- Magno G. *The healing hand; man and wound in the ancient world.* Cambridge, MA: Harvard University Press; 1975.
- Main DS, Henderson WG, Pratte K, Cavender TA, Schiffner TL, Kinney A, Stoner T, Steiner JF, Fink AS, Khuri SF. Relationship of processes and structures of care in general surgery to postoperative outcomes: a descriptive analysis. *J Am Coll Surg.* 2007;204(6):1157–65.
- Malenka DJ, O'Connor GT. The Northern New England Cardiovascular Disease Study Group: a regional collaborative effort for continuous quality improvement in cardiovascular disease. *Jt Comm J Qual Improv.* 1998;24(10):594–600.
- Marshall G, Shroyer AL, Grover FL, Hammermeister KE. Time series monitors of outcomes. A new dimension for measuring quality of care. *Med Care.* 1998;36(3):348–56.
- Mavroudis C, Mavroudis CD, Jacobs JP, Siegel A, Pasquali SK, Hill KD, Jacobs ML. Procedure-based complications to guide informed consent: analysis of society of thoracic surgeons-congenital heart surgery database. *Ann Thorac Surg.* 2014;97(5):1838–49; discussion 1849–51.
- Miller T, Leatherman S. The National Quality Forum: a 'me-too' or a breakthrough in quality measurement and reporting? *Health Aff (Millwood).* 1999;18(6):233–7.
- Milstein A, Galvin RS, Delbanco SF, Salber P, Buck Jr CR. Improving the safety of health care: the leapfrog initiative. *Eff Clin Pract.* 2000;3(6):313–6.
- Murphy BM, Elliott PC, Le Grande MR, Higgins RO, Ernest CS, Goble AJ, Tatoulis J, Worcester MU. Living alone predicts 30-day hospital readmission after coronary artery bypass graft surgery. *Eur J Cardiovasc Prev Rehabil.* 2008;15(2):210–5.
- Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg.* 1999;16(1):9–13.
- Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, Lockowandt U. EuroSCORE II. *Eur J Cardiothorac Surg.* 2012;41(4):734–44; discussion 744–735.
- National Committee for Quality Assurance. About NCQA. 2014a. Retrieved 21 Nov 2014, from <http://www.ncqa.org/AboutNCQA.aspx>
- National Committee for Quality Assurance. Persistence of beta-blocker treatment after a heart attack. 2014b. Retrieved 21 Nov 2014, from <http://www.ncqa.org/ReportCards/HealthPlans/StateofHealthCareQuality/2014TableofContents/BetaBlockers.aspx>
- National Institutes of Health. PROMIS: Patient Reported Outcomes Measurement Information System. 2014. Retrieved 25 Oct 2014, from <http://www.nihpromis.org/>
- National Quality Forum. National voluntary consensus standards for cardiac surgery. Washington, DC: National Quality Forum; 2004.
- National Quality Forum. NQF patient safety terms and definitions. Washington, DC: National Quality Forum; 2009.
- Nawal Lutfiyya M, Bhat DK, Gandhi SR, Nguyen C, Weidenbacher-Hoper VL, Lipsky MS. A comparison of quality of care indicators in urban acute care hospitals and rural critical access hospitals in the United States. *Int J Qual Health Care.* 2007;19(3):141–9.
- Nielsen ME. The legacy of Ernest A. Codman in the 21st century. *J Urol.* 2014;192(3):642–4.
- Norcini JJ, Boulet JR, Dauphinee WD, Opalek A, Krantz ID, Anderson ST. Evaluating the quality of care provided by graduates of international medical schools. *Health Aff (Millwood).* 2010;29(8):1461–8.
- O'Brien SM, Shahian DM, DeLong ER, Normand SL, Edwards FH, Ferraris VA, Haan CK, Rich JB, Shewan CM, Dokholyan RS, Anderson RP, Peterson ED. Quality measurement in adult cardiac surgery: part 2 – statistical considerations in composite measure scoring and provider rating. *Ann Thorac Surg.* 2007;83(4 Suppl):S13–26.
- Osnabrugge RL, Speir AM, Head SJ, Jones PG, Ailawadi G, Fonner CE, Fonner E Jr, Kappetein AP, Rich JB. Cost, quality, and value in coronary artery bypass grafting. *J Thorac Cardiovasc Surg.* 2014a;148(6):2729–35.
- Osnabrugge RL, Speir AM, Head SJ, Jones PG, Ailawadi G, Fonner CE, Fonner Jr E, Kappetein AP, Rich JB. Prediction of costs and length of stay in coronary artery bypass grafting. *Ann Thorac Surg.* 2014b;98(4):1286–93.
- Ozker E, Saritas B, Vuran C, Yoruker U, Ulugol H, Turkoz R. Delayed sternal closure after pediatric cardiac operations; single center experience: a retrospective study. *J Cardiothorac Surg.* 2012;7:102.
- Paone G, Brewer R, Likosky DS, Theurer PF, Bell GF, Cogan CM, Prager RL, T. Membership of the Michigan Society of and S. Cardiovascular. Transfusion rate as a quality metric: is blood conservation a learnable skill? *Ann Thorac Surg.* 2013;96(4):1279–86.
- Papanikolaou PN, Christidi GD, Ioannidis JP. Patient outcomes with teaching versus nonteaching healthcare: a systematic review. *PLoS Med.* 2006;3(9), e341.
- Parsons HM, Henderson WG, Ziegenfuss JY, Davern M, Al-Refaie WB. Missing data and interpretation of cancer surgery outcomes at the American College of Surgeons National Surgical Quality Improvement Program. *J Am Coll Surg.* 2011;213(3):379–91.
- Public Law 99–166. Veterans' Administration Health-Care Amendments of 1985. *Public Law.* 1985;99–166.
- Puskas JD, Kilgo PD, Thourani VH, Lattouf OM, Chen E, Vega JD, Cooper W, Guyton RA, Halkos M. The society of thoracic surgeons 30-day predicted risk of

- mortality score also predicts long-term survival. *Ann Thorac Surg.* 2012;93(1):26–33; discussion 33–25.
- Rodkey GV, Itani KM. Evaluation of healthcare quality: a tale of three giants. *Am J Surg.* 2009;198(5 Suppl):S3–8.
- Rumsfeld JS, Magid DJ, O'Brien M, McCarthy Jr M, MaWhinney S, Scd ALS, Moritz TE, Henderson WG, Sethi GK, Grover FL, Hammermeister KE, S. Department of Veterans Affairs Cooperative Study in Health Services: Processes and S. Outcomes of Care in Cardiac. Changes in health-related quality of life following coronary artery bypass graft surgery. *Ann Thorac Surg.* 2001;72(6):2026–32.
- Rumsfeld JS, Ho PM, Magid DJ, McCarthy Jr M, Shroyer AL, MaWhinney S, Grover FL, Hammermeister KE. Predictors of health-related quality of life after coronary artery bypass surgery. *Ann Thorac Surg.* 2004;77(5):1508–13.
- Shahian DM, Blackstone EH, Edwards FH, Grover FL, Grunkemeier GL, Naftel DC, Nashef SA, Nugent WC, Peterson ED, S. T. S. w. o. e.-b. surgery. Cardiac surgery risk models: a position article. *Ann Thorac Surg.* 2004;78(5):1868–77.
- Shann KG, Giacomuzzi CR, Harness L, Myers GJ, Paugh TA, Mellas N, Groom RC, Gomez D, Thuys CA, Charette K, Ojito JW, Tinius-Juliani J, Calaritis C, McRobb CM, Parpard M, Chancy T, Bacha E, Cooper DS, Jacobs JP, Likosky DS. Complications relating to perfusion and extracorporeal circulation associated with the treatment of patients with congenital cardiac disease: consensus definitions from the Multi-Societal Database Committee for Pediatric and Congenital Heart Disease. *Cardiol Young.* 2008;18 Suppl 2:206–14.
- Sharp LK, Bashook PG, Lipsky MS, Horowitz SD, Miller SH. Specialty board certification and clinical outcomes: the missing link. *Acad Med.* 2002;77(6):534–42.
- Shih T, Zhang M, Kommareddi M, Boeve TJ, Harrington SD, Holmes RJ, Roth G, Theurer PF, Prager RL, Likosky DS, T. Michigan Society of and C. Cardiovascular Surgeons Quality. Center-level variation in infection rates after coronary artery bypass grafting. *Circ Cardiovasc Qual Outcomes.* 2014;7(4):567–73.
- Sobolev B, Mercer D, Brown P, FitzGerald M, Jalink D, Shaw R. Risk of emergency admission while awaiting elective cholecystectomy. *CMAJ.* 2003;169(7):662–5.
- The Joint Commission. Ernest Amory Codman Award. 2014. Retrieved 23 Oct 2014, from <http://www.jointcommission.org/codman.aspx>
- The Society of Thoracic Surgeons. Consumer Reports and STS Public Reporting. 2012. Retrieved 27 Oct 2014, from <http://www.sts.org/news/consumer-reports-and-sts-public-reporting>
- The Society of Thoracic Surgeons. Adult Cardiac Anesthesia Module. 2013. Retrieved 30 Oct 2014, from <http://www.sts.org/sts-national-database/adult-cardiac-anesthesia-module>
- The Society of Thoracic Surgeons. NQF# 0696: The STS CABG Composite Score. NQF: Quality Positioning System. 2014a. Retrieved 30 Oct 2014, from <http://www.qualityforum.org/QPS>
- The Society of Thoracic Surgeons. STS CABG Composite Score. 2014b. Retrieved 30 Oct 2014, from <http://www.sts.org/sts-public-reporting-online/cabg-composite-score>
- The Society of Thoracic Surgeons. STS National Database. 2014c. Retrieved 30 Oct 2014, from <http://www.sts.org/national-database>
- Tran C, Wijeyesundera HC, Qui F, Tu JV, Bhatia RS. Comparing the ambulatory care and outcomes for rural and urban patients with chronic ischemic heart disease: a population-based cohort study. *Circ Cardiovasc Qual Outcomes.* 2014;7(6):8, 35–43.
- Tricoci P, Allen JM, Kramer JM, Califf RM, Smith Jr SC. Scientific evidence underlying the ACC/AHA clinical practice guidelines. *JAMA.* 2009;301(8):831–41.
- van Kasteren ME, Mannien J, Ott A, Kullberg BJ, de Boer AS, Gyssens IC. Antibiotic prophylaxis and the risk of surgical site infections following total hip arthroplasty: timely administration is the most important factor. *Clin Infect Dis.* 2007;44(7):921–7.
- Veterans Health Administration. VHA handbook 1102.3: criteria and standards for cardiac surgery programs. Washington, DC: Veterans Health Administration; 2008.
- Veterans Health Administration and Department of Defense. VA/DoD clinical practice guideline for the management of ischemic heart disease. Washington, DC: Veterans Health Administration, Department of Defense; 1997.
- Veterans Health Administration CARE-GUIDE Working Group, Denver VA Medical Center CARE-GUIDE Coordinating Team and United States Veterans Health Administration Office of Quality Management, Denver VA Medical Center CARE-GUIDE Coordinating Team, United States Veterans Health Administration Office of Quality Management. Veterans Health Administration CARE-GUIDE for ischemic heart disease. Washington, DC: Department of Veterans Affairs; 1996.
- VillaNueva CB, Ludwig ST, Shroyer AL, Deegan NI, Steeger JE, London MJ, Sethi GK, Grover FL, Hammermeister KE. Variations in the processes and structures of cardiac surgery nursing care. *Med Care.* 1995;33(10 Suppl):OS59–65.
- Wakeam E, Hyder JA, Tsai TC, Lipsitz SR, Orgill DP, Finlayson SR. Complication timing and association with mortality in the American College of Surgeons' National Surgical Quality Improvement Program database. *J Surg Res.* 2014,193(1):77–87.
- Winkley Shroyer AL, Bakaeen F, Shahian DM, Carr BM, Prager RL, Jacobs JP, Ferraris V, Edwards F, Grover FL. The society of thoracic surgeons adult cardiac surgery database: the driving force for improvement in cardiac surgery. *Semin Thorac Cardiovasc Surg.* 2015 Summer;27(2):144–51. PubMed PMID: 26686440.
- Yasa H, Lafci B, Yilik L, Bademci M, Sahin A, Kestelli M, Yesil M, Gurbuz A. Delayed sternal closure: an effective procedure for life-saving in open-heart surgery. *Anadolu Kardiyol Derg.* 2010;10(2):163–7.





# Health Services Information: Data-Driven Improvements in Surgical Quality: Structure, Process, and Outcomes

# 7

Katia Noyes, Fergal J. Fleming, James C. Iannuzzi, and  
John R. T. Monson

## Contents

<b>Introduction</b> .....	142
<b>Stakeholders for Surgical Outcome Assessment</b> .....	144
<b>Types of Data for Surgical Outcome Assessment</b> .....	145
Existing Data Sources .....	145
Data Quality .....	147
Changes in Surgical Procedures and Practices Over Time .....	150
Individual Surgeon Variation (Preferences, Techniques, and Skills) .....	152
Timing of Complications .....	152
Limited Information on Socioeconomic Drivers of Health .....	152
Need for Linked Data .....	153
Data Management and Big Data .....	154
<b>Structure-Process-Outcome Assessment in Surgery</b> .....	154
Theoretical Framework of Quality Assessment in Healthcare .....	154
Structure .....	156
Process .....	157
Surgical Outcomes .....	158
Risk Adjustment .....	161

---

K. Noyes (✉)

Department of Surgery, University of Rochester Medical  
Center, Rochester, NY, USA

e-mail: [katia\\_noyes@urmc.rochester.edu](mailto:katia_noyes@urmc.rochester.edu)

F. J. Fleming · J. C. Iannuzzi

University of Rochester Medical Center, Rochester, NY,  
USA

J. R. T. Monson

Florida Hospital System Center for Colon and Rectal  
Surgery, Florida Hospital Medical Group Professor of  
Surgery, University of Central Florida, College of  
Medicine, Florida Hospital, Orlando, FL, USA

e-mail: [john.monson.md@flhosp.org](mailto:john.monson.md@flhosp.org)

<b>From Data to Quality Improvement</b> .....	161
Understanding Hospital Billing Data .....	161
Focusing on Modifiable Factors .....	161
Identifying Actionable Goals .....	162
Presenting Results .....	163
<b>References</b> .....	165

## Abstract

The barriers to surgical quality improvement in the United States are significant. Fee-for-service reimbursement approach does not encourage provider communication and drives volume, not value. Quality report cards and pay-for-performance strategies have been implemented to reflect performance of individual providers at specific healthcare settings, but they have not been very effective at enforcing continuity of care and integration. In this chapter we describe how, using Donabedian approach to quality assessment, one can develop reliable and useful quality indicators for surgical services. We review main sources of relevant data and discuss practical implications of working with each of the databases. Finally, we provide an overview of current knowledge gaps and challenges affecting surgical care delivery and provide recommendation for future research and quality improvement interventions.

## Introduction

Quality assessment and public reporting are powerful approaches to improve quality of care whether it is preventive services, acute surgical care, and chronic illness management. We can learn a lot from the 20 years of coronary artery bypass grafting (CABG) surgery report cards experience (Hannan et al. 2012). It is also widely recognized that the chief factor of the success of the cardiac surgery report cards is the development of the New York State (and then national) coronary angioplasty reporting system to ensure collection of high-quality clinical data, including data elements not routinely available from administrative databases. Establishment of the

Percutaneous Coronary Interventions Reporting System (PCIRS) in 1992 allowed for development and ongoing calibration of the cardiac surgery risk-adjusted mortality model which in turn provides meaningful reports that local practices can use to compare their performance with similar groups and national benchmarks, without the fear of being penalized for treatment high-risk patients. In the last 20 years, greatly due to the publically available CABG Reports Cards, the outcomes of CABG and over cardiac surgical procedures have improved dramatically (Mukamel and Mushlin 1998; Hannan et al. 1994, 1995).

In recent decades, as life expectancy has continued to grow around the world, the illness profile of highly populated countries in the Middle East and Asia has undergone an epidemiologic transition from predominantly infectious diseases to primarily chronic illness, vastly expanding the role and importance of surgical services. Surgical procedures that were previously extremely rare as well as “simple, ineffective, and relatively safe” became common, “complex, effective, and potentially dangerous” (Chantler 1999). On average, an American patient is expected to undergo about 10 surgical procedures in a lifetime, translating into an estimated 234 million operations annually worldwide (Weiser et al. 2008; Lee et al. 2008). While surgery can be extremely beneficial, often saving lives, surgical procedures are also associated with the risk of complications, infection, and death. Furthermore, surgical interventions are the key treatment modalities for many prevalent conditions including cancer, trauma, and obstetrics, positioning surgical quality and safety as one of the top public health concerns.

Public worry and focus on medical outcomes is entirely warranted. The Institute of Medicine (IOM) in the landmark 1999 patient safety report

“To Err is Human” concluded that the healthcare in the United States is not as safe as it should be. One of the report’s main revolutionary conclusions was that the majority of medical errors in the United States did not result from individual recklessness. More commonly errors are caused by faulty systems, processes, and underlying conditions that lead people to either make mistakes or fail to prevent them. The report advocated reducing harm through system-based initiatives rather than increasing pressure on individual providers (Brown and Patterson 2001). A focus on surgical outcomes is thus even more paramount where any small slip can quickly lead to disastrous consequences.

While the IOM report led to some system-level improvements, including expansion of health insurance coverage through PPACA in 2010, many problems remained or even worsened. In 2013, the IOM convened a committee of experts to examine the quality of cancer care in the United States and formulate recommendations for improvement. *Delivering High-Quality Cancer Care: Charting a New Course for a System in Crisis* presented the committee’s findings and recommendations. The committee concluded that the cancer care delivery system is in crisis due to a growing demand for cancer care, increasing treatment complexity (including surgical procedures), a shrinking workforce, and rising costs (Levit et al. 2013).

While it is widely recognized and accepted that assessment of surgical quality and outcomes should be a continuous process alongside care delivery, there is no clear consensus on how, when, and what outcomes should be measured. The problem is fueled by the fact that quality’s definition changes depending on the stakeholder’s perspective. For instance, surgeons evaluate each other’s quality based on technical skills, board certifications, and morbidity which is under their perceived direct control, characteristics that are often invisible and hence meaningless to patients. Instead, patients prefer clinicians with excellent communication skills who are always on time, regardless of whether or not the surgeon is a board-certified Fellow of the American College of Surgeons (FACS). Similar

discrepancies and misalignments can be observed with respect to surgical outcomes. The vast majority of surgical oncologists will consider clean margins as synonymous with being “cured of cancer,” despite the fact that a patient may still have to endure many months of exhausting and toxic chemotherapy and radiation, temporary or permanent colostomy, fatigue, depression, and undesirable cosmetic changes. Successful quality improvement in clinical practice requires a common vision, multidisciplinary plans, and cooperation among all involved stakeholders, across the spectrum of all clinical providers including healthcare administrators, payers, social services, community organizations, and patient advocates.

Hurtado (Hurtado et al. 2001) defines quality as “the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge,” but such broad definitions can have limited direct applications. A more useful definition of quality measures it over six domains: effectiveness, timely access, capacity, safety, patient centeredness, and equity (Leatherman and Sutherland 2003). Within each of these domains, it is possible to measure various elements, and so from this paradigm, a picture of a service’s quality of care can be outlined. However, such comprehensive assessment can be too burdensome and thus not practical for frequent monitoring and real-time evaluation.

In addition, there have been significant efforts to identify and assess important elements of care pathways, rather than individual procedures, which may lead to better outcomes and higher quality (Donabedian 1966; Hurtado et al. 2001; Maxwell 1984; Schiff and Rucker 2001; Sitzia and Wood 1997). Many countries have made significant progress with the implementation of national quality programs (Department of Health Office 1995; Department of Health 2000) including NSQIP (Agency for Healthcare Research and Quality 2009; Australian Commission on Safety and Quality in Healthcare 2008; American College of Surgeons 2014a), but further research is required to accurately and affordably improve assessments of surgical quality.

## Stakeholders for Surgical Outcome Assessment

There are many stakeholders that actively participate in surgical quality initiatives. When there is common purpose between these groups, progress can easily be made; however, often agendas do not align making advancement difficult. Understanding the key stakeholder, their perspective, and roles is fundamental to quality improvement.

Medical societies and professional groups have long been the leaders in developing clinical practice guidelines, supporting provider accreditation, and both auditing and providing clinical training as well as continuing medical education activities. While heavily dominated by surgeons, the field of surgical outcome assessment also includes medical and radiation oncologists, imaging scientists, primary care providers, other advanced care partners, and allied health professionals. These include, but are not limited to, the American College of Surgeons (ACS), the Commission on Cancer (CoC) the Consortium for Optimizing Surgical Treatment of Rectal Cancer (OSTRiCh), American Society of Colon and Rectal Surgeons (ASCRS), Society for Surgery for the Alimentary Tract (SSAT), Society of Surgical Oncology, and others (American College of Surgeons 2014b, c; Optimizing the Surgical Treatment of Rectal Cancer 2014; Society for Surgery of the Alimentary Tract 2016; Society for Surgical Oncology 2014).

The provider stakeholder structure can take many forms and can work at every level of the healthcare system. For instance, the American College of Surgeons represents an umbrella organization that pushes an overarching quality agenda. Its purpose is to be broad, as the organization spans multiple disciplines. While ACS includes lobbying initiatives in congress, it also has recently employed benchmarking for hospitals and now individual providers through data collection and risk adjustment. Other broad organizations, such as the National Comprehensive Cancer Network (NCCN), release specific consensus guidelines aimed at improving care through utilizing the best available evidence. Other societies with a narrower focus also

contribute to determining guidelines aimed at standardizing care for specific biologic systems as demonstrated by the American Society of Colon and Rectal Surgeons who release guidelines about colon screening recommendations, prophylaxis, and other elements of cancer care. There are also disease-specific groups such as the Consortium for Optimizing Surgical Treatment of Rectal Cancer (OSTRiCh), or regional groups such as the Upstate New York Quality Initiative (UNYSQI), which currently focuses on improving the quality of colon resections. Quality improvement at the hospital and surgical division level also occurs aimed at more specific interventions such as thromboprophylaxis protocols or surgical site infection prevention bundles that are more applicable to single providers or individual hospital systems. This hierarchical structure, however, is not partitioned or independent with extensive overlap between organizations, societies, disease-specific coalitions, and locoregional initiatives. Collaborations between all groups can propel initiatives; however, their recommendations are not always aligned with one another with nuanced differences that can create confusion and can potentially hinder quality improvement efforts.

In the current environment post-PPACA, accountable care organizations are frequently the key drivers of clinical quality improvement. This is because according to the Triple Aim principle developed by Don Berwick and the IHI, high-quality care overall is less expensive than poor care. Accountable Health Partners LLC is one of the accountable care organizations in the Greater Rochester area. It was organized to create a partnership between URMC and community physicians, to enable them to succeed in the looming era of value-based contracts by creative initiatives to deliver high-quality care at a lower cost. The goals and interests of the AHP are parallel to those of PPACA: to engage specialty providers in the delivery of integrated care pathways; to establish efficient communication between care managers in medical homes, primary care, and specialist practices; to develop an integrated information system capable of monitoring quality of care measures; and to

develop a payment mechanism to facilitate such engagement.

Other community-based stakeholders may include medical societies, public health and safety providers and agencies, social and aging services, and educational organizations. Stakeholders outside of the healthcare system and non-for-profit world may include patient support groups and organizations, payers, large self-insured corporations, and business alliances who are also interested in improving overall community health at a lower cost (Blackburn 1983; Brownson et al. 1996; Group 1991; Fawcett et al. 1997; Goodman et al. 1995; Howell et al. 1998; Johnston et al. 1996; Mayer et al. 1998; Zapka et al. 1992; Roussos and Fawcett 2000). In Upstate New York, the Greater Rochester and Finger Lakes regions are well recognized for their long history of community-wide collaborations including University of Rochester Medical Center, Finger Lakes Health Systems Agency (FLHSA), Monroe County Medical Society (MCMS), Rochester Business Alliance, Rochester regional office of American Cancer Society (2014), local payers (e.g., Excellus Blue Cross Blue Shield), accountable care organizations, and others. The FLHSA is an independent community health planning organization working collaboratively with multi-stakeholder groups to improve healthcare quality and access and eliminate healthcare disparities in the nine-county Finger Lakes region. Its mission is to bring into focus community health issues via data analysis and community engagement and to implement solutions through community collaboration and partnership. It has become the convener and facilitator of multi-stakeholder community initiatives to measure and improve the health, healthcare, and cost of care. In the initial round of the CMMI Innovation Challenge, the FLSHA was awarded with a \$26.6 million initiative “Transforming the Delivery of Primary Care: A Community Partnership.”

Excellus Blue Cross Blue Shield is a nonprofit health plan, whose mission is to work collaboratively with local hospitals, doctors, employers, and community leaders to offer affordable healthcare products. For instance, Excellus administers its managed care products for

Medicaid eligible individuals through its partnering organization, Monroe Plan. Over the years, Excellus partnered with many other community stakeholders (e.g., Kodak, MCMS, URMC) to lead several area-wide initiatives aimed to improve quality of care and population health and reduce necessary variation in care and services overuse.

---

## Types of Data for Surgical Outcome Assessment

### Existing Data Sources

There are multiple types of medical data available, and each have their own set of complexities that while answering important questions also leave gaps that require further analysis from alternative perspectives found through other data sources. Typical datasets are comprised of the following: hospital discharges, claims, registry, and survey results. Other administrative types of data include hospital discharge data or billing data as recorded and provided by the hospital itself. These datasets are highly dependent on local practices and can vary between institutions. It can be linked with other subject data providing an in-depth chart review; however, it is limited by the cases performed at an individual hospital. Some states have statewide discharge census data, including California and New York (Hannan et al. 1994, 1995, 2012, CA Society of Thoracic Surgeons 2014). These datasets provide billing data at a larger level, which includes ICD-9 codes by diagnosis, with the ability to track hospital and surgeon level variation, subject linking longitudinally across in-state and charges (in contrast to claims paid out) (Table 1).

Claims data are available at a national as well as local levels and include Medicare data that can be linked to other datasets and insurance claims (i.e., Excellus-blue shield, large self-insured corporations (Xerox, Kodak), and data warehouses (Thompson Reuters)). Registry data can be quite detailed, albeit specific to the registry’s purpose. Examples of registry datasets include tumor registries like SEER that can be linked to Medicare

**Table 1** Types of data used to assess surgical outcomes, quality, and safety

Types of data	Databases	Examples
Cancer registry	SEER, NCDB	(Mack et al. 2013; Rutter et al. 2013)
Hospital registry	Case series	(Sinclair et al. 2012; Aquina et al. 2014b)
Observational	SPARCS, Statewide data, Medicare/Medicaid, UHC	(Rickles et al. 2013; Aquina et al. 2014a)
Randomized controlled trials	CEA/CAS (NASCET) Colonoscopy trial, Breast cancer z0011	(Ferguson et al. 1999; Grube and Giuliano 2001; Whitlock et al. 2008; Atkin 2003)
Cost-Data	PharMetrics, hospital billing, Medicare Charges, Tufts Cost-Effectiveness Registry	(Iannuzzi et al. 2014b; Jensen et al. 2012; Tufts 2014)
Process measures	SCIP, WHO Surgical checklist, inpatient smoking, VTE prophylaxis	(The Joint Commission Core Measure Sets 2014a; American College of Surgeons, Commission on Cancer, Surgical Care Improvement Project 2014b; Safety 2008)
Satisfaction	HCAHPS, Press Ganey	(Systems 2014; Press Ganey Associates 2014)
Benchmarking	ACS-NSQIP observed to expected mortality ratio (United States, thoracic, transplant; United Kingdom, all surgeons), hospital compare, creating centers of excellence (Medicaid Centers of Excellence for breast cancer)	(Centers for Medicare and Medicaid Services 2014; Department of Health 2000; Cohen et al. 2009a, b; Medicare.gov 2014)
	AMA provider survey	(Etzioni et al. 2010, 2014)
	AHA (ICU/staffing/nursing)	(Nallamotheu et al. 2006; Solomon et al. 2002)

*SEER* Surveillance, Epidemiology, and End Results Tumor Registry, *NCDB* National Cancer Data Base, *SPARCS* New York Statewide Planning and Research Cooperative System, *UHC* University HealthSystem Consortium, *SCIP* Surgical Care Improvement Project, *VTE* venous thromboembolism, *HCAHPS* Hospital Consumer Assessment of Hospital Providers and Systems, *ACS-NSQIP* American College of Surgeons National Surgical Quality Improvement Project, *CMS* Center for Medicare and Medicaid Services, *AHA* American Hospital Association, *AMA* American Medical Association, *ICU* intensive care unit

for more robust analysis, NCDB that expands cancer data beyond the identified cancer centers that are included within SEER, and the National Surgical Quality Improvement Program (NSQIP) registry that samples approximately 20% of all cases performed at participating hospitals. Other registries include those maintained by provider organizations (AMA, AHA). Finally, survey data can provide the patient perspective that is lacking from other large dataset analyses. Two prime examples are the Medicare Current Beneficiary Survey and the Hospital Consumer Assessment of Hospital Providers and Systems (HCAHPS) Survey.

The first database for surgical outcomes was developed in NYS for cardiothoracic surgery (Hannan et al. 1990) leading to substantial quality improvement, facilitating development of the field of quality assessment and risk adjustment in

medicine, and challenged the traditional approach of confidential reporting of adverse events. Based on its success, this was expanded to the STS National Database established in 1989. The STS states that “physicians are in the best position to measure clinical performance accurately and objectively” (Surgeons 2014), serving as a mandate for surgeon participation in these initiatives.

While cardiac surgery has long maintained a similar database for tracking quality, this approach was expanded nationally to help improve surgical outcomes. The National Surgical Quality Improvement Program (NSQIP) has been a major development within the surgical community as it provides more detailed surgical information at a national level than was ever previously available. The main purpose of this program was to improve quality through benchmarking, where hospitals were given risk-

adjusted data comparing outcomes nationally to other hospitals of similar size. Based on the depth of data, numerous research studies have been conducted, describing surgical risk factors and comparing operative approaches. While this has been very useful for expanding our understanding of surgical quality as a whole, it was quickly realized that different operations needed specific in-depth data in order to design meaningful quality improvement strategies. One approach to providing more detailed data has been the roll out of procedure targeted variables, in which institutions can add to the traditional NSQIP data for additional cost. This approach allows for a more detailed approach to individual procedures. This was first made available with the release of the 2012 NSQIP dataset, and the impact remains to be seen. Targeted variables have required consensus from experts that can be difficult to obtain and be limited in its scope. This in-depth approach also requires more resources limiting participation.

Another specialty-specific approach includes the Organ Procurement and Transplantation Network (OPTN) database aimed at monitoring transplant programs nationally. This is monitored and run by the US Department of Health and Human Services (National Cancer Institute 2014). The desire for more detailed data has led to a number of subspecialty datasets modeled after NSQIP. A few examples include a vascular surgery-specific dataset, the Vascular Quality Initiative (2014), Pediatric NSQIP, and an endocrine surgery-specific dataset (Collaborative Endocrine Surgery Quality Improvement Collective 2014). The methods of data collection vary, NSQIP employs a clinical nurse reviewer, and CESQIP does not yet have the same infrastructure, requiring the surgeon or the surgeon's designee to input data.

Another approach has been the creation of regional collaboratives, which requires a high level of collaboration with both academic and nonteaching hospitals alike. Regional collaboratives will likely play a role in decreasing unnecessary variability and tracking quality at a more manageable, regional level, where it is easier to implement change than at the national level. Thus far, the regional approach has been seen in both Michigan and Central New York. The central

New York collaborative, called UNYSQI (Upstate New York Surgical Quality Initiative), has focused predominantly on colorectal surgery and more specifically at addressing the question of readmissions. NSQIP allows for 40 additional variables, and given this narrow limitation, specific questions must be addressed.

Participation in data collection programs is promoted as it meets criteria for both maintenance of certification (MOC) and Physician Quality Reporting System (PQRS) as part of CMS (EHealth University: Centers for Medicare & Medicaid Services 2014). This section for maintaining credentials requires that providers evaluate their performance based upon specialty-established requirements which must include national benchmarking. The MOC outlines six core competencies, one of which is practice-based learning and improvement. Part IV of the process for continuous learning includes practice performance assessment. For the American Board of Surgery, diplomats must participate in a national, regional, or local surgical outcome database or quality assessment program. The PQRS is a part of CMS and is the second specific incentive promoting the use of outcome data collection programs as it uses both payment adjustments to penalize, as well as incentive payments to ensure providers report quality data (Table 2).

## Data Quality

A common saying in large database analysis is "garbage in garbage out," and while there are methods to account for missing data, a major limitation remains with extensive missing data points. One approach might be to limit case inclusion to only those with a full set of data; however, this quickly limits patient inclusion. This approach may be appropriate for some major data points such as sex, where it can be assumed that if subject sex is not included then other variables are likely to be of questionable quality. Missing data may also be secondary to the data collection process. For instance, in NSQIP, preoperative laboratory values are gathered; however,

**Table 2** Databases and outcomes used to assess surgical outcomes, quality, and safety

Dataset	Description	Sample and outcomes
ACS-NSQIP <a href="http://site.acsnsqip.org/">http://site.acsnsqip.org/</a>	Maintained by the American College of Surgeons. Participation through annual fees by hospital	30-day data based on postoperative outcomes. Provides benchmarking
Pediatric NSQIP <a href="http://www.pediatric.acsnsqip.org/">http://www.pediatric.acsnsqip.org/</a>	Subset of overall NSQIP	30-day follow-up for surgical procedures performed on pediatric patients
VQI (Vascular Quality Initiative) <a href="http://www.vascularqualityinitiative.org">www.vascularqualityinitiative.org</a>	Vascular procedure-specific data (including those performed by radiologists, cardiologists, and vascular surgeons). Follow-up through 1 year. Governed by the Society of Vascular Surgeons (SVS) Patient Safety Organization	255 participating centers. Uses cloud computing to allow multiple users to enter data and does not depend on full-time data entry specialist. Can be integrated into electronic medical records
CESQIP (Collaborative Endocrine Surgery Quality Improvement Program) <a href="http://cesqip.org/">http://cesqip.org/</a>	Since 2012, through the American Association of Endocrine Surgeons (AAES)	Patient-centered data collection, ongoing performance feedback to clinicians, and improvement based on analysis of collected data and collaborative learning
STS National Database <a href="http://www.sts.org/national-database">http://www.sts.org/national-database</a>	Society of Thoracic Surgeons run program that makes quality scores available to institutions and the public at large. National data for research requires specific application to the STS and is not released to participating hospitals by virtue of inclusion in data gathering	Focuses on three areas: adult cardiac, general thoracic, and congenital heart surgery
The Surveillance, Epidemiology, and End Results (SEER) program funded by the National Cancer Institute <a href="http://seer.cancer.gov/about/overview.html">http://seer.cancer.gov/about/overview.html</a>	1973–2011 cancer incidence and survival data from population-based cancer registries covering approximately 28 % of the US population	Includes data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and 12-month survival
<i>Hospital discharge data</i>		
Statewide Planning and Research Cooperative System (SPARCS) California Patient Discharge Dataset National Inpatient Sample (US) <a href="http://www.hcup-us.ahrq.gov/nisoverview.jsp">http://www.hcup-us.ahrq.gov/nisoverview.jsp</a> Hospital Episode Statistics (UK) <a href="http://www.hscic.gov.uk/hes">http://www.hscic.gov.uk/hes</a>	Comprehensive all-payer data reporting system. The system was initially created to collect information on discharges from hospitals	Patient-level data on patient characteristics, diagnoses and treatments, services, and charges for each hospital inpatient stay and outpatient (ambulatory surgery, emergency department, and outpatient services) visit, and each ambulatory surgery and outpatient service visit to a hospital extension clinic and diagnostic and treatment center licensed to provide ambulatory surgery services
The Centers for Medicare & Medicaid Services (CMS) claims and survey data <a href="http://www.resdac.org/cms-data/file-directory">http://www.resdac.org/cms-data/file-directory</a>	CMS is responsible for administering the Medicare, Medicaid, and State Children's Health Insurance Programs. CMS gathers and formats about Medicare beneficiaries, Medicare claims, Medicare providers, clinical data, and Medicaid eligibility and claims. CMS also collects additional survey data on health behavior and utilization Medicare &	Data on acute, psychiatric and skilled nursing inpatient admissions, outpatient services, procedures and tests, use of prescription medications, skilled nursing, durable medical equipment, and hospice

(continued)



**Table 2** (continued)

Dataset	Description	Sample and outcomes
	Current Beneficiary Survey (MCBS) and satisfaction with care Consumer Assessment of Healthcare Providers & Systems (CAHPS)	
American Hospital Association (AHA) Annual Hospital Survey <a href="http://www.aha.org/research/rc/stat-studies/data-and-directories.shtml">http://www.aha.org/research/rc/stat-studies/data-and-directories.shtml</a>	Hospital-specific data on approximately 6,500 hospitals and 400-plus systems	1,000 data fields covering organizational structure, personnel, hospital facilities and services, and financial performance
American Medical Association (AMA) Physician Masterfile <a href="http://www.ama-assn.org/ama/pub/about-ama/physician-data-resources/physician-masterfile.page">http://www.ama-assn.org/ama/pub/about-ama/physician-data-resources/physician-masterfile.page</a>	Established in 1906, current and historical data for more than 1.4 million physicians, residents, and medical students in the United States, including approximately 411,000 graduates of foreign medical schools	Information about demographics, practice type, significant education, training and professional certification on virtually all Doctors of Medicine (MD) and Doctors of Osteopathic Medicine (DO)

there remains extensive variation in timing of preoperative labs, as well as whether a specific blood level is checked at all. One particular example is albumin level. Albumin level has demonstrated associations with nutrition and overall health status. Studies have shown associations with surgical outcomes as well; however, this laboratory value is not always checked preoperatively. In fact, there may be a bias of checking this value in patients that may be at risk for malnutrition or have other major comorbidities. This fact may bias results leading to concern about its inclusion in multivariable analysis, even though it holds clinical value. Some suggest it should not be included at all, while others suggest it requires a more nuanced approach. Albumin, for instance, is reported as a continuous variable, but can be transformed into a binary variable using clinically meaningful cutoffs previously described as 3.5 g/dl. By assuming all missing values fall within the normal range, one creates a differential misclassification that underestimates the true effect as some in this group may in fact have low albumin levels. Thus, if an observed association is found, it likely is true, albeit an underestimate. The data can then still be useful for clinical decision making even though many values are in fact missing. Another approach to this same problem can be assessing whether those in the missing dataset are different with respect to the endpoint than the others. This is specifically testing whether there is differential misclassification. If there is, then one can treat the

missing data group as its own categorical level without making any assumptions if there is an observed effect compared to subjects with data. Another method includes imputation of data. These methods are beyond the scope of this chapter, but briefly involve separate analysis predicting that specific data point based on the subject's other characteristics.

Missing data of the first type (missing sex) can be avoided through auditing processes. Many data collection programs employ auditing processes to ensure quality data and sites are not included if they demonstrate inability to conform to predetermined standards.

Another major limitation to all large datasets is changing variable definitions over time. While this process is necessary to some extent as clinically meaningful definitions may change with time, it can drastically limit the subject numbers available for analysis for that endpoint. One such example is postoperative transfusion within NSQIP. Initially, the number of transfused units was included intraoperatively and postoperatively defined as greater than 4 units. Researchers were able to then describe this endpoint as major postoperative bleeding and specifically describe the extent of intraoperative blood loss. This changed in 2011 when the number of intraoperative units of blood was removed altogether and postoperative transfusion was changed to 2 units or more of packed red blood cells. The first limitation is the danger of merging datasets across years without

understanding these changes. First, if ignored, researchers may erroneously code these missing intraoperative transfusions as no transfusion given and make assumptions upon it which will clearly be mistaken. Secondly, it poses a challenge in the second instance as the postoperative transfusion variable in the newer dataset has a different clinical meaning. Two units of blood can be given for merely low hematocrit levels with comorbidities meant to optimize patients and no longer representing a postoperative bleeding event. These two variables of transfusion are not comparable over time, given the changes limiting analysis.

### **Changes in Surgical Procedures and Practices Over Time**

Other issues regarding data collection include the constantly evolving process of case definition and even the addition of new surgical procedures over time. For instance, the change from ICD-9 to ICD-10 is looming, and how this will impact data collection remains to be seen. The nuanced changes between the two systems will likely impact some areas more than others, and a deep understanding of these nuances will be necessary to compare cases between these two time periods. The last major ICD coding change was in 1975, and the medical arena has changed dramatically in that time including the advent of the electronic record.

Some databases only include ICD-9 coding where numerous different procedures may be relevant for repair of that diagnosis, for instance, appendicitis can be treated by an open approach making an incision in the right lower quadrant or can be treated using laparoscopic techniques, using three small incisions and a camera for appendix extraction. Where only ICD-9 codes are available such datasets lack discrimination preventing comparison of operative approach.

The introduction of laparoscopic procedures is one example of how surgical procedures change over time; while the first report of laparoscopic appendectomy was published in 1981, this practice did not become ubiquitous until the turn of the

century and now represents the preferred technique (Korndorffer et al. 2010).

These changes can significantly impact research as each procedure has specific complications; however, there may be limits in the available data due to changes not captured by the coding systems. For instance, CPT coding does not capture robotic techniques lumping them with laparoscopic procedures. This has limited observational studies comparing or even tracking robotics usage over the past decade. Another example on the limits of CPT coding include the absence of transanal endoscopic microsurgery (TEMS) codes used for distal rectal cancer resections that are of sufficiently minimal rectal wall invasion. This approach is a minimally invasive one that spares the rectum and the sphincter allowing for essentially full rectal function in low-grade tumors; however, they are lumped in with other rectal cancer resections which often include complete rectal resections with end colostomy or loss of sphincter. The difference in quality of life and even the types of complications are huge. While it clearly makes it impossible to perform observational studies on TEMS within large datasets, it also adds variation and error into any assumptions about outcomes after low rectal cancer resections. There are some ways to exclude TEMS from dataset by selecting cases where the tumor stage was sufficiently high to make TEMS contraindicated; however, this does not help elucidate specifically the advantages of TEMS. Another example where CPT coding fails is differentiating between some specific laparoscopic approaches. Although open inguinal hernia repair has been a bread-and-butter surgical operation, within the last decade, increasingly surgeons are applying their laparoscopic skills to hernia repair. There are two available laparoscopic approaches: totally extraperitoneal (TEP) or transabdominal preperitoneal (TAPP). The TAPP approach enters the abdominal cavity in standard laparoscopic fashion repairing the hernia from the inside using tacks, whereas the TEP approach enters a space above the peritoneum placing the mesh between layers and usually does not require tacks to keep the mesh in place. Both approaches

may have different risk profiles and long-term sequelae; however, observational evaluation is limited since there is no differentiation by CPT codes in the ICD-9 system.

There also remain many processes that are not coded in most databases. This includes many data points that may impact outcomes, such as patient follow-up strategies, staffing, utilization of trainees, and even postdischarge medications. While large datasets evolve, opportunities to expand the data as research questions arise may be available. UNYSQI is one example where through the ACS-NSQIP institutions can track their own specific data points which may help answer specific questions.

The surgical field is constantly progressing, not just specifically with new procedures but also with the introduction of entirely new specialties. For example, endocrine surgery is starting to become a major surgical subspecialty; although not yet a board-certified specialty, the presence of these more specialized surgeons may impact outcomes. Other major changes in surgery may also impact outcomes, which have not been included in current databases. For example, resident work hour restrictions by the ACGME continue to change and become increasingly strict. Previously, it was not unheard of for surgical residents to work 120–100 h weekly, where now work hours are capped at 80 per week and interns are prevented from taking 24-h call. These changes have drastically changed patient coverage and in some cases required supplementing staffing through advanced practice providers or moonlighters. These changes have not been tracked and it is unclear how changing the workforce structure has impacted outcomes. Although controversial, this question holds some urgency as more and more restrictions are being implemented. In fact, a new randomized controlled trial will observe how these restrictions impact care; one arm of the trial will require surgical residents to follow the new regulations, while the other will function without work hour restrictions. However, such data is largely absent from current datasets.

Other major changes include the advent of telemedicine, and with robotics, even remote

operations are now possible with the first transatlantic cholecystectomy or so-called “Lindbergh” operation was performed in 2001 (Marescaux et al. 2002). These changes were only possible through improvements in electronic communication that decreased the lag time sufficiently to allow such an operation.

The role that virtual communication will have in the future remains unclear, but will likely increase in frequency in the coming decades. Currently, such approaches are not tracked; however, including such practices in large healthcare databases may be useful in understanding their uptake and impact on clinical care. Other adjunct advances also impact surgical care, although largely unappreciated, such as major advances and availability in high-quality imaging. Where 20 years ago computed tomography was limited, it is now ubiquitous and high-quality scans are available within minutes. These findings change the diagnostic paradigms and the quality of surgical decision making, although availability of such high-quality CT scans is not included in databases, even those that track whether CT scanning was done at all. Other technological advances include intraoperative imaging through 3D laparoscopy and the development of new instruments that make previously unthinkable operative approaches possible such as single incision surgery or natural orifice transluminal endoscopic surgery that allows surgeons to perform cholecystectomy through the vagina.

There are many other changes to the structure of healthcare that may drastically impact outcomes including advances in patient monitoring or quality of care in the intensive care unit. While it would be onerous to include all of these changes into any given dataset, it is important to remember the many forces that impact outcomes. Much like a projectile in physics has many forces that alter its course such as friction, rotation, and wind forces, and many of these forces can be ignored to provide the overall picture using the major forces of velocity and gravity on the object to provide an estimated course; however, keeping these other forces in mind remains important as they may have potential to be key forces in surgical care.

## Individual Surgeon Variation (Preferences, Techniques, and Skills)

Even if there is a single code and agreed-upon surgical treatment or practice, the implementation of this can vary considerably. Laparoscopic cholecystectomy, for instance, one of the most commonly performed operations, has considerable variation in the way the procedure itself is performed. The absence of this precise detail is an obstacle to standardizing procedures nationally. There are statistical techniques for controlling for variation at the surgeon level, specifically hierarchical modeling with random effects. Hierarchical random effect modeling also addresses the issue that most multivariable models ignore; independence assumptions are voided in healthcare studies as patients are treated by surgeons within hospitals which have been shown to impact quality. Surgeon volume is one surgeon factor that was initially noted in 1979, where complex procedures such as pancreatectomy and coronary artery bypass graft have better outcomes when performed by higher-volume surgeons (Solomon et al. 2002; Birkmeyer et al. 2002; Katz et al. 2004). This may in part reflect standardization of technique, evidence-based practice, and skill, which may be a function of practice. Teasing out how outcomes are dependent on technique variation is virtually impossible in current large dataset, although one could argue this variation might explain quality to a much greater degree than even risk adjustment based on patient factors.

## Timing of Complications

Even if a reasonable outcome is chosen, it is essential to understand the interplay of that complication with the hospital course. Incorrect assumptions about this can lead to incorrect answers. Recent studies on readmissions have suffered from major errors when they attempt to include complications as risk factor for readmission (Aquina et al. 2014b). Some studies suggest that complications are the biggest risk factor for readmission, and while this may seem reasonable, they often confuse the reason the

patient was admitted with a risk factor for readmission. This has led to disastrous consequences as inclusion of such reasons for readmission in the model can make all other risk factors no longer statistically significant, and in one model, the authors came to the incorrect conclusion that the only risk factor for readmission was postoperative complications, although subsequent studies have demonstrated this to be false. This can be avoided by using complication timing to define complications as during the inpatient stay as compared to at postdischarge. While predischarge complications have been associated with readmissions, the effect estimates have been much lower than previously described when all complications are considered together.

## Limited Information on Socioeconomic Drivers of Health

Analyses of patterns and outcomes of care require an assessment of the complex relationships among patient characteristics, treatments, and outcomes. Furthermore, according to the Andersen healthcare utilization model (Aday and Andersen 1974), usage of health services (including inpatient care, outpatient physician visits, imaging, etc.) is determined by three dynamics: predisposing factors, enabling factors, and need. Predisposing factors can be characteristics such as race, age, and health beliefs. For instance, an individual who believes surgery is an effective treatment for cancer is more likely to seek surgical care. Examples of enabling factors could be familial support, access to health insurance, one's community, etc. Need represents both perceived and actual need for healthcare services. To conduct and interpret outcome analyses properly, researchers should both understand the strengths and limitations of the primary data sources from which these characteristics are derived and have a working knowledge of the strategies used to translate primary data into the categories available in public databases. For instance, SEER-Medicare documents details on individual cancer diagnoses, demographics, (age, gender, race), Medicare eligibility and program enrollment by month, and

aggregate measures of the individual's "neighborhood" (e.g., average income and years of education presented at the zip-code and census-tract level) as determined through a linkage to recent US Census data. However, census level data do not allow for assessment of differences among those zip-code areas.

Many analyses of large databases focus on the patient's race or ethnicity as a confounder or a predictor of outcome or a marker for other unobserved factors (disadvantaged geographic area or low health literacy). Information on race is generally available, while information on ethnicity is often missing or inappropriately coded. While most of the US data surveys allow only one category for Hispanic ethnicity (yes/no), the NCDB classifies cancer patients into seven categories (Mexican, Cuban, Puerto-Rican, Dominican, South/Central American, Hispanic by name, and Other). In our analysis of treatment patterns for Hispanic cancer patients in NCDB, we demonstrated persistent disparities in receipt of guideline-recommended care. The care in Hispanic group as a whole was not significantly different from non-Hispanic, while individual subgroups demonstrated significant differences, highlighting a critical need of acknowledging Hispanic subgroups in outcome research.

## Need for Linked Data

Surgical safety and quality are multifactorial issues with more than one risk factor and hence multiple potential mechanisms for improvement. For instance, reduction in postsurgical complications could be partially achieved by more efficient patient education about early symptoms, improvement in surgeon's skills, changes in nursing and hospital practices, use of surgical visiting nurse services, and other interventions. Similarly, one quality improvement intervention may have impact on multiple stakeholders including patients and their caregivers, clinic personnel, and health insurance. Hence, a comprehensive evaluation may require information about all involved parties. Such data are rarely available in one dataset, and therefore, many surgical

outcomes and quality improvement studies are using multiple merged sources of data.

*The SEER-Medicare data* is a product of a linkage between two large population-based datasets: Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute and beneficiaries healthcare claims data collected by the Center for Medicare and Medicaid Services for billing and enrollment purposes. The linked dataset includes Medicare beneficiaries with cancer from selected states participating in SEER Program, with unit of observation being one healthcare utilization event. This includes all Medicare-covered healthcare services from the time of a person's Medicare eligibility (before or after cancer diagnosis) until their death. Because of complex sampling design, number of included variables, and specific data reporting practices for tumor characteristics and services utilization, the investigator considering a SEER-Medicare-based study or a proposal should spend time understanding SEER-Medicare data limitations (National Institute of Health 2014) and learning about data layout and coding (manuals and training are available at the NCI and other cancer research organizations).

*The Medicare Current Beneficiary Survey (MCBS)* is a longitudinal survey of a nationally representative sample of the Medicare population. The MCBS contains data about sociodemographics, health and medical history, healthcare expenditures, and sources of payment for all services for a randomly selected representative sample of Medicare beneficiaries (Centers for Medicare and Medicaid Services 2014). For every calendar year, there are two separate MCBS data files released: Access to Care and Cost and Use files which can be ordered directly from the CMS with assistance from the Research Data Assistance Center at the University of Minnesota (Research Data Assistance Center 2014).

*MCBS Access to Care file* contains information on beneficiaries' healthcare access, healthcare satisfaction, and their usual sources of care (Goss et al. 2013; Research Data Assistance Center 2014). *MCBS Cost and Use file* offers a complete summary of all healthcare expenditure and source of payment data on all healthcare services including

expenditures not covered by (CMS Research Data Assistance Center 2015). The information collected in the surveys is combined with the claims data on the use and cost of services. Medicare claims data includes information on the utilization and cost of a broad range of costs including inpatient hospitalizations, outpatient hospital care, skilled nursing home services, and other medical services. In order for the Cost and Use file to collect, summarize, and validate accurate payment informations, the release of C&U file is usually delayed by 2 years compared to the MCBS AC file.

In addition to publically available merged datasets, individual investigators can create their own aggregated databases by linking together information from multiple sources and combining existing data with prospectively collected and patient-reported information. Examples of such studies include a NSQIP-based evaluation of pre-operative use of statins and whether it is associated with decreased postoperative major noncardiac complications in noncardiac procedures (Iannuzzi et al. 2013c), a study of recipients of abdominal solid organ transplant (ASOT) using additional data from patient medical records (Sharma et al. 2011), and a retrospective review of the data from medical records of patients diagnosed with hepatocellular carcinoma compared to patients in the California Cancer Registry (CCR) (Atla et al. 2012).

## Data Management and Big Data

More and more data are being collected for different purposes and are available to be linked together including electronic memberships, online purchasing and consumer behavior records, electronic transactions and others. The datasets become so large and complex that it becomes difficult to manage using traditional resources, and organizations have to increase their resources in order to be able to manage them. Before we know what to do with it, we have entered into a new era of big data. Big data is high-volume, high-velocity, and/or high-variety information assets that require new forms of

processing to enable enhanced decision making, insight discovery, and process optimization (Gartner 2013). The challenges of working with big data include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations, among many others. Innovative solutions such as cloud computing chip away at some challenges while remaining limited by others. For instance, cloud computing outside services such as Amazon ec2, box, dropbox, internet2, etc. provide storage or processing capabilities, but without internal infrastructure or agreements with the outside services, there is the potential for privacy violations. Yet, just like with the administrative data several decades earlier, the opportunities provided by big data potentially outweigh the risks and, in time, may become data-driven analytics as routine as EMR and digital image sharing.

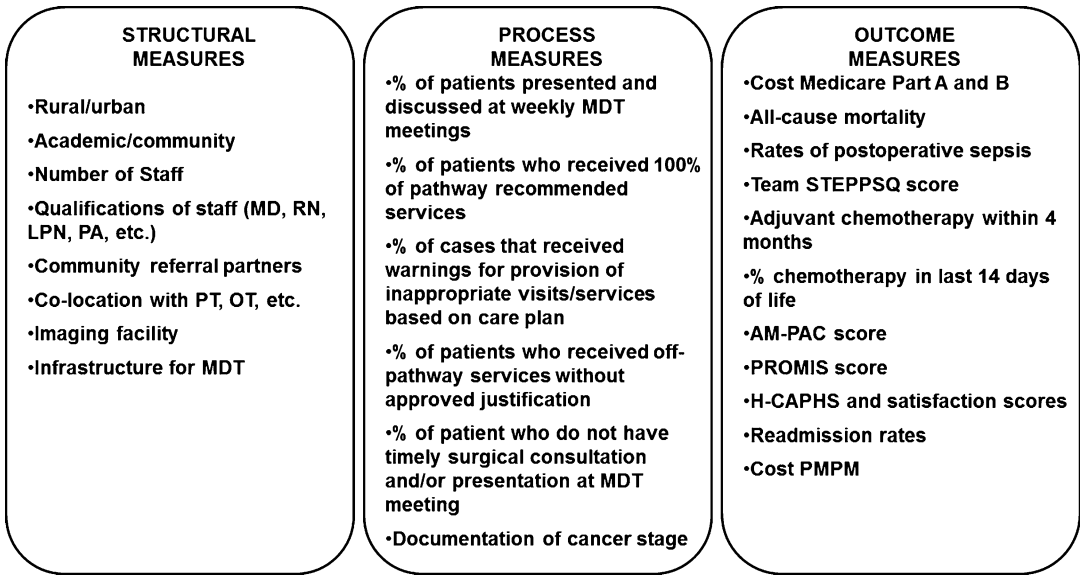
---

## Structure-Process-Outcome Assessment in Surgery

### Theoretical Framework of Quality Assessment in Healthcare

According to Donabedian (1966), if there is evidence that good structure leads to appropriate processes which in turn result in good outcomes, quality of healthcare intervention could be measured in terms of either structures (S), processes (P), or outcomes (O) (Fig. 1).

These indicators can be measured using electronic, readily available, data from the organizational health information systems, data collected by cancer trackers, and other regional data systems, like Rochester RHIO. It is important to work closely with each hospital's clinical quality assessment team, to avoid redundancy in data collection and other quality assessment and reporting initiatives (e.g., Hospital Scorecard, the Clinical Service Scorecard, and the Management Plan Tracking Reports, SCIP, HCAHPS), and others (Hospital Consumer Assessment of Healthcare Providers and Systems 2014; The Joint Commission Core Measure Sets 2014a). Additional financial and pre- and postadmission



**Fig. 1** Donabedian approach for evaluating outcomes

cost and utilization information about patients can be obtained from CMS claims data for Medicare fee-for-service beneficiaries and Excellus BCBS claims for commercially insured and Medicare HMO patients (Medicare Health Insurance Claim (HIC) number or health insurance ID will be abstracted from the patients' medical charts).

The bundles of care for surgical patients can be defined by multidisciplinary care teams for specific diagnoses and surgical service lines. A care bundle identifies a set of key interventions from evidence-based guidelines that, when implemented, are expected to improve patient outcomes (Institute for Healthcare Improvement 2006). The aim of care bundles is to change patient care processes and thereby encourage guideline compliance in a number of clinical settings (Brown et al. 2002; Burger and Resar 2006; Pronovost et al. 2006). Using regional or national healthcare utilization and expenditure data with Medicare or private plan reimbursement schedule, clinicians and hospital administrators can estimate annual cost of care for surgical patients receiving various care bundles, by disease stage. These bundled cost estimates can be used internally (e.g., for budgeting projections or to calculate return on investment for new programs and interventions) or externally, to provide a foundation for contract negotiations

with payers, regional healthcare systems, and accountable care organizations (Froimson et al. 2013; Ugiliweneza et al. 2014).

While it is tempting to seek out a single perfect metric of surgical quality, anybody familiar with the complexity and variation in patient risks and the delivery of surgical care would agree that such metric could not possibly exist. More suitable would be a multidimensional measure similar to the six-domain definition of healthcare quality suggested by the World Health Organization (WHO). These dimensions require that healthcare be:

- **Effective:** delivering healthcare that is adherent to an evidence based and results in improved health outcomes for individuals and communities

*Example:* each cancer case is reviewed by a specialty multidisciplinary team at least once before the final decision about treatment is reached.

- **Efficient:** delivering healthcare in a manner that maximizes resource use and avoids waste

*Example:* avoid unnecessary imaging for colorectal cancer (CRC) patients such as PET scans or multiple CT scans.

- **Accessible:** delivering healthcare that is timely, geographically reasonable, and provided in a setting where skills and resources are appropriate to the medical need  
*Example:* providing a hub-and-spoke model for chemotherapy delivery for CRC patients residing far from major cancer centers
- **Acceptable/patient centered:** delivering healthcare which takes into account the preferences and aspirations of individual service users and the cultures of their communities  
*Example:* offering palliative care to all patients with advanced cancer
- **Equitable:** delivering healthcare that does not vary in quality because of personal characteristics such as gender, race, ethnicity, geographical location, or socioeconomic status  
*Example:* providing financial assistance to low-income cancer patients assuring that out-of-pocket expenses do not represent a barrier for adequate treatment
- **Safe:** delivering healthcare that minimizes risks and harm to service users  
*Example:* following WHO surgical checklist to minimize the risk of surgical complications and never events

As illustrated by the examples above, this definition of healthcare quality provides the link between the organization of care, care processes, surgical quality, and outcomes. Hence, it enables all participating stakeholders (e.g., clinicians, researchers, payers, and hospital administrators) to rely on Donabedian's framework when assessing quality of surgical services. According to Donabedian, if there is evidence that good structure leads to appropriate processes which in turn results in good outcomes, quality of healthcare intervention could be measured based on presence of appropriate structures (S) or processes (P).

Below we provide several examples of evidence-based measures of quality in surgical care.

## Structure

Lord Darzi, international expert on quality and innovation in cancer care, world-leading colorectal

surgeon, the former Minister of Health in the United Kingdom, and the lead author of the UK Darzi Plan to redesign care delivery, encouraged healthcare agencies to “localize care where possible, and centralize services where necessary” for efficacy and safety. This implies that routine healthcare, like cancer survivorship services, should take place as close to home as possible, while more complex care, like active cancer treatment, should be centralized to ensure it is carried out by the most skilled professionals with cutting-edge equipment and high volume/experience.

There exist several validated care delivery models to improve access to specialty care for patients with complex chronic disease living in underserved or remote communities (for instance, using videoconferencing technology for enhanced care coordination). There is a large body of literature demonstrating that standardized care pathways, use of multidisciplinary teams (MDTs), resident involvement (Iannuzzi et al. 2013a, b), availability of specialized providers (e.g., board-certified surgical specialists, surgical nurses, and PA) and services (e.g., stoma care, wound care, surgical ICU), and receiving care in a high-volume center of excellence are associated with better outcomes (Reames et al. 2014; Howell et al. 2014).

Evidence that hospital volume influences outcomes has been verified in nearly every major type of surgery (Begg et al. 1998; Birkmeyer et al. 2002; Katz et al. 2004). This body of work highlighted important and previously unrecognized variations in hospital performance and ignited efforts to improve surgical quality among poorly performing hospitals. In an effort to reduce these variations among hospitals, new health policy and quality improvement initiatives, such as public reporting, pay-for-performance, and surgical checklists, have been implemented to promote best practice and improve standards of care (Hannan et al. 1990, 2012; Haynes et al. 2009; Lindenauer et al. 2007). Over the last decade, surgical mortality rates have significantly decreased throughout the country, possibly due to such measures (Weiser et al. 2011; Finks et al. 2011; Birkmeyer 2012). While surgical/facility volume is easy to measure, the mechanism of



association between procedure volume and outcomes remains to be poorly understood. Possible explanations highlight the importance of surgical expertise, specialized services, and infrastructure that tend to be associated with large-volume centers.

Patient management following multidisciplinary principles consistently leads to superior outcomes at much lower costs. Published supporting evidence for improved cancer-specific outcomes with the use of multidisciplinary teams is available for a range of cancers, including breast, lung, head and neck, esophageal, and colorectal (Chang et al. 2001; Coory et al. 2008; Gabel et al. 1997; Stephens et al. 2006; Wille-Jorgensen et al. 2013; Burton et al. 2006).

## Process

Many factors that constitute the structure and organization of surgical services contribute to the processes of care and, ultimately, affect patient outcomes. For instance, in addition to knowing structural features, such as whether a hospital has a surgical ICU, it is also important to identify processes of care, such as how the ICU is staffed and what policies, regulations, and checklists the SICU personnel adhere to, including failure to rescue, escalation of care, communication, use of imaging and antibiotics, and patient nutritional protocols. If a residence program is housed in a hospital (structure), what, when, and how surgical residents are required to perform during cases (processes) may vary by institution and has serious impact on institutional outcomes.

There is also a growing interest regarding the potentially detrimental impact of interruptive operating room (OR) environments on surgical performance (Healey et al. 2006; Wiegmann et al. 2007). Previous investigations showed that interruptions occur frequently in ORs, across various surgical specialties (Weigl et al. 2015).

In an effort to improve surgical outcomes and potentially lower costs, recent attention has been placed on efficiency of care delivery and the surgical volume-outcome relationship. Luft et al. first explored this concept in 1979 showing that there

was a relation between hospital volume and mortality for complex procedures such as open-heart surgery or coronary bypass (Luft et al. 1979). Since then, Birkmeyer et al. expanded on this idea by showing a significant relationship between both hospital volume and surgeon volume and operative mortality for many different procedures, including resections for lung, bladder, esophageal, and pancreatic cancer (Birkmeyer et al. 2002). Subsequent surgical oncology studies have shown an association between volume and negative margin status, superior nodal harvest, and both short-term and long-term survival. Recently, volume-outcome relationship has been demonstrated even for less specialized procedures, such as incisional hernia repair (Aquina et al. 2014a).

Evidence of the volume-outcome relationship, along with financial pressures, implementation of surgical bundled payments, and shift to accountable care organizations brought to light the importance of efficient and coordinated models of care delivery. With the increase in the number of surgical subspecialties and nonsurgical specialties performing surgical procedures (e.g., intervention radiology and cardiology, urogynecology), there is an increase in the involvement of advanced practice providers in patient care delivery (e.g., nurse practitioners (NP), physician assistants (PA), technicians, and therapists) and growing acceptance of multidisciplinary care pathways (oncology, geriatrics, orthopedics, among others). For example, high-volume bariatric surgery practices can hire psychologists, nutritionists, exercise therapists, and specialty nurses to provide additional supportive services. This approach can free surgeon's time and improve care coordination and patient experience. There are other situations when the specialty and training of provider is important – for the procedures that could be performed by different types of providers, for instance, inferior vena cava filter (IVC filter), a type of vascular filter that is implanted to prevent life-threatening pulmonary emboli (PEs). IVC filters could be placed by a number of different types of providers (vascular surgeons, general surgeons, cardiologists, interventional radiologists) for various indications. The outcomes of the intervention

(mortality, complications, PE) could potentially depend on the specialty and skill of the provider.

In general, clinic staff rarely bill for their services and often are employed by the institution. Multidisciplinary consultations for cancer patients are also not reimbursable and often count toward “academic time” for faculty physicians. As a result, these services may be “invisible” from insurance claims or medical records. In fact, only one provider can be associated with each billable service (procedure or hospital admission). For any service delivered by more than one provider (e.g., resident participating in a surgical case, several APPs involved in hospital discharge process), additional data may need to be included (e.g., operating notes, individual provider claims).

## Surgical Outcomes

A choice of optimal outcome for each study or evaluation depends on the goal of the assessment as well as factors that may be driving this outcome (causal pathway) and resources available to the investigators as some of the outcome collection processes may be very costly and time consuming (e.g., health utility and quality of life measurement) (Drummond et al. 2005; Iezzoni 2004). Below we describe some of the most common types of outcomes used in surgical outcome research and quality assessment and discuss their applications, limitations, and sources of data.

## Clinical Outcomes

*Mortality:* When defining mortality, it is important to be specific about the duration of the observation period (e.g., in-hospital vs. 30-day mortality) as well as the starting point for the observation period (e.g., day when the procedure was performed for 30-day postsurgical mortality versus 30 days after hospital discharge for 30-day hospital mortality). Using hospital discharge abstracts and publicly available software, one can measure in-hospital mortality using the most appropriate definitions for the needs of the project. For instance, if there is a significant variation in the hospital length of stay between patients in the study, it may be more accurate to define hospital

mortality based on the 30-day postadmission interval rather than postdischarge time (Borzecki et al. 2010; Hannan et al. 1990, 2013).

*Cancer Survival:* For surgical oncology studies, cancer survival rate is often more appropriate outcome metric than surgical mortality because the vast majority of cancer patients receive multimodal therapy. Cancer survival is reported by most tumor registries or can be calculated from pathology reports. Cancer survival is defined as a percentage of people who have survived a certain type of cancer for a specific amount of time (e.g., 12 months, 2 or 5 years). Certain cancers can recur many years after first being diagnosed and treated (e.g., breast cancer). During this time, a former cancer patient (also called survivor) may die from a different condition (oncologic or benign), and hence, the most appropriate choice of reported statistics in this case would be tumor site-specific mortality. For instance, patient may be successfully treated for thyroid cancer but die from colon cancer 20 years later. Other types of survival rates that give more specific information include disease-free survival rate (the amount of cancer patients who are cancer-free), progression-free survival rate (the amount of cancer patients who are not cured but their cancer is not progressing), and cancer recurrence (cancer that has returned after treatment and after a period of time during which the cancer was not detected). Sometimes without detailed pathology data, it is impossible to distinguish cancer recurrence from cancer progression. An example of recurrence versus progression dilemma could be observed in rectal cancer patients who received nonsurgical neoadjuvant treatment. Following neoadjuvant chemoradiotherapy (CRT) and interval proctectomy, 15–20% of patients are found to have a pathological complete response (pCR) to combined multimodal therapy, but controversy persists about whether this yields a survival benefit (Martin et al. 2012).

*Surgical Complications: Incisional Hernia.* Incisional hernia is abdominal wall fascia that fails to heal. Incisional hernia is a common postoperative complication following major abdominal surgery. Data on incidence of incisional hernia is highly variable with reported values ranging

from 0% to 91%. Diagnosis for incisional hernias is typically within the first 3 years after initial laparotomy (Yahouchy-Chouillard et al. 2003; Rosen et al. 2003; Rea et al. 2012); however, it may take up to 10 years to become evident after the initial surgery (LeBlanc et al. 2000; Akinici et al. 2013). This large amount of variation in the reported rates of incisional hernia is not unforeseen, given the wide assortment of the group of patients included into the studies, the executed surgery, and the amount of time during the follow-up (Caglià et al. 2014). Several outcome measures could be appropriate for a study on incisional hernia including incidence, prevalence, rates of hospital admission, and reoperation.

*Surgical Complications: Surgical Site Infection (SSI)* (Schweizer et al. 2014). In addition to pain, discomfort, and high risk for readmission, surgical site infections (SSIs) are identified with an excessive amount of morbidity and mortality. The costs of SSIs have been the focus of quality improvement and safety efforts ever since the Centers for Medicare and Medicaid have halted compensation for the growing costs linked with SSIs after some surgical operations (so-called potentially preventable infections) (Aquina et al. 2014b). Prior studies have reported cost of hospitalizations after SSIs in the range from \$24 000 to \$100 000 (Schweizer et al. 2014).

### **Patient-Reported Outcome Measures (PROMs)**

*Patient-Reported Outcomes Measurement Information System (PROMIS®)*: Measures included in PROMIS® are intended for standardized assessment of various patient-reported outcome domains – including pain, fatigue, emotional distress, physical functioning, and social role participation (Devlin and Appleby 2010). PROMIS® is a new set of tools intended to be used in routine clinical practice as a part of electronic medical record (EMR) (Cella et al. 2007) system. PROMIS® was established in 2004 with funding from the National Institutes of Health (NIH). PROMIS measures are based on common validated metrics to ensure computerized and burden-free data collection process in any

healthcare setting that yields accurate measurement of patient health status domains over time with few items (National Institute of Health 2015a).

*Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS)* (Systems 2014): Just like with any other consumer goods and services, many providers and organizations have collected information on patient satisfaction with healthcare. However, prior to HCAHPS, there was no national standard for collecting and publicly reporting patients' perspectives on their healthcare experience that would enable valid comparisons to be made across providers. In May 2005, the National Quality Forum (NQF), an organization responsible for standardization of healthcare quality measurement and reporting, formally endorsed the CAHPS® Hospital Survey (Press Ganey Associates Inc 2014).

The HCAHPS survey is mailed to a random sample of hospital patients after a recent discharge. The survey asks patients to rate 21 aspects of their hospital care combined into nine key topics: communication with patients and doctors, communication between patients and nurses, responsiveness of the hospital staff, pain management, communication with patients about medicines, discharge information, hospital's cleanliness, hospital environment's noise levels, and transition of care. Patients' perception of care is a key performance metric and is used to determine payments to hospitals (Hospital Consumer Assessment of Healthcare Providers and Systems 2014). The Hospital Compare database (4605 hospitals) can be used to examine complication rates and patient-reported experience for hospitals across the nation. Prior studies have demonstrated an inverse relationship between patient experience and complication rates. This negative correlation suggests that reducing these complications can lead to a better hospital experience. Overall, these results suggest that patient experience is generally correlated with the quality of care provided.

Depending on the type of surgery and patient population, other outcome measures may be also relevant (e.g., pain, functional status, and cognitive ability). Quality of life is a multidomain

indicator that combines all aspects of health relevant to patients and, hence, may serve as an aggregate outcome measure.

*Quality of Life and Subjective Well-Being* (Lee et al. 2013): Quality continues to be placed at the heart of discussions about healthcare. This raises important questions how quality of care should be measured and from whose perspective, patient's, provider's, or payer's. Subjective well-being (SWB) is a measure of the overall "wellness" of an individual and as such has the potential to be used as this global marker for how treatments affect people in the experience of their lives. SWB links all stages in the treatment and care process, thus allowing the overall quality of care to be determined and valued according to its direct effect on people's lives. SWB has been shown to have an effect on outcomes at all stages of the treatment experience, and improved health and quality outcomes are shown to consistently enhance SWB (Lee et al. 2013). Furthermore, SWB measures have been shown to be a suitable method to value the impact of healthcare on the families and caregivers of patients and, in this way, can join up health outcomes to show wider effects of treatment on patients' lives. Measuring an individual's SWB throughout his or her treatment experience can enable a full appraisal of the quality of care that they receive. This could facilitate service improvements at the microlevel and help value treatments for resource allocation purposes at the macrolevel.

### Surrogate Outcomes

Although everybody recognizes the importance of measuring patient outcomes and several valid and accurate measures (as described above) are available, there are several practical barriers to measuring patient outcomes. These include time (waiting for cancer recurrence or mortality to occur while maintaining regular follow-up with a patient), personnel costs (to perform routine surveillance and follow-ups), and patient burden (repeated follow-up, evaluations, and surveys). One of the potential solutions to these problems is use of surrogate outcomes. A **surrogate outcome** (or **endpoint**) is a measure of effect of a specific treatment that may substitute for a *real*

clinical endpoint but does not necessarily have a guaranteed relationship (Cohn 2004). Surrogate markers are also used when the number of events is very small, thus making it impractical to conduct a clinical trial to detect a statistically significant effect (e.g., instead of measuring VTE events which have an incidence of less than 1%, studies often use ultrasound-detected blood clots which are much more prevalent but do not always result in PE or VTE) (Fleming and DeMets 1996). A correlate does not make a surrogate. It is a common misconception that if an outcome is a correlate (i.e., correlated with the true clinical outcome), it can be used as a valid surrogate endpoint (i.e., a replacement for the true clinical outcome). However, proper justification for such replacement requires that the effect of the intervention on the surrogate endpoint predicts the effect on the clinical outcome – a much stronger condition than correlation. Other examples of commonly used surrogate outcomes in surgery include costs of care as a measure of poor outcomes and disability, positive surgical margins, carcinoembryonic antigen (CEA), and number of lymph nodes retrieved as a measure of long-term cancer recurrence and mortality (Nussbaum et al. 2014).

### Composite Outcomes: Episode of Care or Care Bundles

The value of quality reporting in surgical care, however, is limited by problems with existing measures of quality, mainly, that existing quality indicators are designed to measure the quality of a specific facility (e.g., hospital) or a specific provider (e.g., surgeon). This, however, does not reflect the current paradigm of care delivery when a patient may be diagnosed in the community, referred to a regional center of excellence for neoadjuvant chemoradiation, followed up for 6 months by an academic colorectal surgeon, before returning back to the community for years of posttreatment surveillance. Regional standardized pathways of care and multidisciplinary team (MDT) approach has been recommended by all clinical societies to better identify, coordinate, deliver, and monitor the optimal treatment on an individual patient-by-patient basis (Chang et al.

2001; Coory et al. 2008; Stephens et al. 2006; Abbas et al. 2014; Wille-Jorgensen et al. 2013; Morris et al. 2006; Gatt et al. 2005; Adamina et al. 2011).

## Risk Adjustment

Risk adjustment is a set of analytic tools used for an array of functions in the healthcare (Iezzoni and Long-Bellil 2012; Schone and Brown 2013). One of the primary uses of risk adjustment is providing fair comparison between different patient populations, providers, or programs. Risk adjustment is also necessary to set costs for health plans to suggest expected treatment expenses of their specific membership group. Because of discrepancy in everyone's health and treatment needs, the cost and outcomes of healthcare may differ from person to person. Without risk adjustment, plans or providers have an enticement to enroll and treat healthier patients (so-called cream skimming or cherry-picking) and avoid sick, frail, or complex patients. After appropriate risk adjustment, plans and providers receive a larger amount of reimbursement for members with numerous chronic illnesses than for members with a small amount of or no health problems at all. In addition to costs, risk adjustment is also applied to health outcomes when comparing performance across providers (e.g., risk-adjusted mortality is reported by the STS National Database and NSQIP, CABRG Report Cards NYS, UK surgical mortality (National Health Services 2015); The Society of Thoracic Surgeons National Database 2014). The methodology used to risk adjustment varies, depending in part on healthcare market regulations, the populations served, and the source of payments. Risk adjustment is used in all major public programs offering health coverage in the United States – including Medicare Advantage (MA), Medicare Part D, and state Medicaid managed care programs. The STS National Database, with its three million patient records, has long used risk adjustment to provide more accurate patient outcomes. If not risk adjusted, the records of surgeons who perform operations on higher-risk patients would always look worse than the records

of surgeons who treat low- or average-risk patients.

---

## From Data to Quality Improvement

### Understanding Hospital Billing Data

For many hospital and outpatient services, there is a wide difference between billed charges and the amounts that providers expect to receive for services. Hospital charges are usually determined by hospital administrators depending on prior history and demand. Reimbursement rates, on the other hand, or the payments that hospitals are actually willing to accept for a specific service or product, vary by payer and specific plan. On average, hospitals billed Medicare 3.77 times (standard deviation = 1.83) what they were actually reimbursed, with a range of 0.42 to 16.23 (Muhlestein 2013). The ratio may vary for private payers.

High hospital charges, though, do have some important consequences. First, since the charges do not correlate with the amount being paid and hospital expenditures required to produce a specific service (i.e., true cost), it becomes difficult, if not impossible, to compare process between hospitals, and draw conclusions about financial sustainability of various service lines. Second – and potentially devastating for some – those who are uninsured who receive care at a hospital, or those who are insured and receive care at an out-of-network hospital, may face a bill that greatly exceeds by many times the negotiated price paid by any payer.

### Focusing on Modifiable Factors

One of the major paradoxes that limits our ability to improve practice based on the results of published studies is that most available predictors are not modifiable (readmissions: patient severity, comorbidities), while most modifiable factors are not routinely collected through standard clinical data systems (SES, organizational structure). Furthermore, the reported statistical associations not equal causation (but often assumed) and hence,

modifying predictor may not result in a desired change in the outcome of interest. Let's consider the example below.

Failure to rescue (FTR) refers to the mortality among patients with serious complications (Johnston et al. 2014; Pucher et al. 2014; Almoudaris et al. 2013). Typically, it is hospitals with greater FTR rates (not greater complication rates) that have the greatest rate of mortality. Thus although complications may occur, outcomes can still be improved by optimizing the quality of care provided to the patient post-complication. Although there have been several studies highlighting the importance of FTR as a marker for quality of care, these have only considered organizational aspects of healthcare. Few have explored the underlying human factors that lead up to this critical event. Two main factors may contribute toward an FTR event: first, a failure to recognize a sick patient and, second, a failure to act promptly once deterioration has been detected. In both situations, an escalation of care (EOC) process is required if FTR is to be avoided.

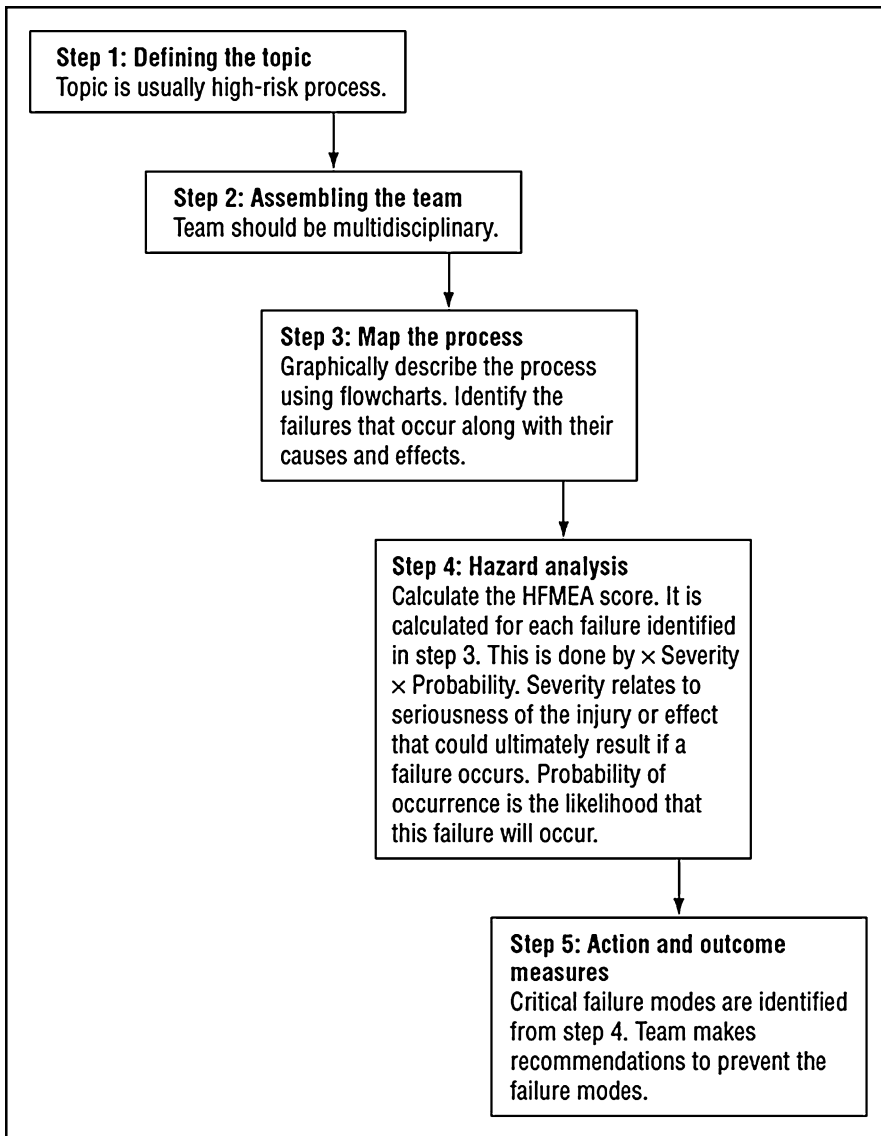
EOC involves a nurse recognizing a change in patient status and communicating it to a postgraduate year 1 (PGY1) resident, who subsequently reviews the patient and then escalates care further for advice and/or management. Escalation is a difficult process, as the first doctor called by the nurses will usually be the most junior; this is the traditional hierarchy. After initial assessment, the junior doctor must then contact his or her senior to explain why they need help and the urgency of response required. All of this places a premium on the value of communication between team members. However, failures in communication are ubiquitous and frequent in the postoperative phase. Although this EOC process lies at the center of FTR and is critically important for safety and quality of surgical care, it remains difficult to measure and quantify and, hence, relatively unexplored in the research literature.

## Identifying Actionable Goals

Despite the most sound study design and state-of-the-art statistical methodology, outcome studies

do not always lead to meaningful improvement in care quality and patient outcomes. Is this the ground for skepticism? Not at all. Just like many investigations in basic biomedical sciences, outcomes and quality assessment projects often fall short of their potential impact by simply reporting barriers to high-quality care without considering strategies for systematically overcoming these limitations and obstacles. Other common mistake is assuming that just because some risk factors are statistically associated with poor quality or outcomes, they represent a target for improvement. For instance, if low patient education is associated with poor cancer prognosis, it may be naïve to assume that more education would improve outcomes in cancer patients without a high school diploma. In this case, low education is likely to be a marker for social and economic deprivation in this demographic group. Addressing this issue may require developing a system-wide solution like providing a care navigator, graphics rather than text-based decision support tools, and phone- rather than internet-based communication with care providers.

Sometimes when large administrative dataset are used for the analysis, statistically significant risk factors are not necessary clinically significant. Before considering any change in clinical practice, it may be beneficial to review the results for face validity with all stakeholders involved in care process. One approach is to use a systematic quantitative validated method to assess risks in the process of information transfer across all phases of surgical care. The method is known as failure mode and effect analysis (FMEA) and was originally developed by engineers to accomplish proactive risk analyses (McDermott et al. 1996). The National Center for Patient Safety of the US Department of Veterans Affairs adjusted FMEA for use in healthcare, resulting in healthcare FMEA (HFMEA) (DeRosier et al. 2002). Healthcare FMEA is a multistep process (Fig. 2) that uses a multidisciplinary team to proactively evaluate a healthcare process. The team uses process flow diagrams, hazard scoring, and decision trees to identify potential vulnerabilities and to assess their potential effect on patient care. The method captures the likelihood of risks, the severity of consequences, and the probability that they



**Fig. 2** Main steps in surgical healthcare failure mode and effect analysis (*HFMEA*) (Adapted from the Veterans Affairs National Center for Patient Safety, DeRosier et al. 2002)

may be detected and intercepted before causing harm. Healthcare FMEA has so far been applied to medication administration (Fletcher 1997; McNally et al. 1997; Kunac and Reith 2005; Weir 2005), intravenous drug infusion (Adachi and Lodolce 2005; Apkon et al. 2004; Wetterneck et al. 2006), blood transfusions (Burgmeier 2002), equipment problems (Weinstein et al. 2005; Wehrli-Veit et al. 2004), and surgery (Nagpal et al. 2010).

## Presenting Results

Quality outcome research results may be presented in a variety of ways depending in part upon the endpoint and how that data will be used. Standard statistical approaches using student's *t*-test for continuous and chi-square for categorical data, for instance, have long been noted to have biased results based on patient factor distribution. This is particularly

important for observational studies using data where patients have not been randomized. Higher-level statistical packages using multivariable approaches to adjust for patient-level factors are now readily available, providing adjusted estimated effects in terms of odds ratios. Despite the ubiquity of such methods, if not well thought out, results can be drastically skewed. Only confounding factors and covariates not on the causal pathway should be included. If one controls for factors on the causal pathway, one may find that no presumed risk factors are associated with the outcome, because they have been effectively controlled for in the multivariable analysis. This will be discussed further below. Confounders such as comorbidities may also be highly collinear, and grouping or using already established practices for comorbidity adjustment may be helpful in decreasing the number of variables, particularly if the research question is regarding comparing two different surgical approaches where one only desires to adjust for comorbidities rather than ascertain their independent contribution to risk for poor outcome.

While multivariable analyses are presented with odds ratios, even this relatively straightforward result presentation requires some additional thought in terms of the desired interpretation. One particular nuance is whether using a reference group that makes the odds ratio greater than one, in other words suggesting increased risk, or such that the odds ratio suggests a protective effect. It is often more intuitive to present odds ratios suggesting increased risk; however, this is not always appropriate.

As quality data becomes more prevalent, multiple metrics reportedly measuring the same poor outcome may exist. Auditing these results and comparing which approach is more reliable and measures the underlying disease state is of utmost importance, particularly if this data is to lead to clinical change. For instance, using Pearson's correlation coefficient, a study of NSQIP data when compared to regional data measuring anastomotic leaks found that the traditional approach of "organ space infection"

poorly correlated with the more specific anastomotic leak variable as more specifically defined. These findings suggest that prior reports are based on identifying organ space infection as an anastomotic leak in colorectal surgery.

Odds ratios may be difficult to put into clinically meaningful terms other than demonstrating relative importance. Another approach to taking multivariable analysis to the next step is the creation of risk scores aimed at guiding clinical decision making. This approach effectively operationalizes the data available in multivariable analysis by weighting risk factors. The approach to these analyses is slightly different as they are aimed at predicting an event, rather than identifying all potential risk factors. This changes in which variables are included in analysis, as only those that improve the predictive ability should be used. There may be a high degree of crossover; however, risk scores are most useful when they are simple and so one may desire to make a parsimonious model, that is, a model with the fewest number of covariates while maximizing the predictive power of the model (Iannuzzi et al. 2013d, 2014a; Kelly et al. 2014a). In order to perform a predictive analysis, data should be split into a development and validation dataset so the risk score can be tested on naive subjects estimating its ability to be applied to novel patients. Another similar approach is the use of nomograms, which is simply another way to organize risk score-type data.

With the advent of the electronic record, some of this risk scoring can now be integrated directly into the clinical record, alerting physicians about high-risk patients for readmissions or high-risk DVT patients prompting some action such as prophylaxis prescription. This approach has increased the use of guideline-based approaches and may be an effective tool moving forward. NSQIP also provides individual patient risk calculators for many complications which allow in-office estimates of risk based on individual patient factors. This tool anecdotally has a high degree of satisfaction for patients and providers alike and likely improves the consent process.



## References

### Primary Sources

- American College of Surgeons (ACS). National Surgical Quality Improvement Program. American College of Surgeons. 2014a. <http://site.acsnsqip.org/>. Accessed 19 Sept 2014.
- Andersen R, Newman J. Societal and individual determinants of medical care utilization in the United States. *Milbank Q*. 2005;83(4):1–28.
- Birkmeyer J. Progress and challenges in improving surgical outcomes. *Br J Surg*. 2012;99(11):1467–9.
- Cohen M, Dimick J, Bilimoria K, Clifford K, Richards K, Hall B. Risk adjustment in the American College of Surgeons National Surgical Quality Improvement Program: a comparison of logistic versus hierarchical modeling. *J Am Coll Surg*. 2009a;209(6):687–93.
- Donabedian A. Evaluating the quality of medical care. *Milbank Mem Fund Q*. 1966;44:166–206.
- Fleming F, Thomas R, DeMets D. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med*. 1996;125(7):605–13.
- Hospital Consumer Assessment of Healthcare Providers and Systems. HCAHPS: Hospital Consumer Assessment of Healthcare Providers and Systems. 2014. <http://www.hcahpsonline.org/home.aspx>. Accessed 5 May 2015.
- Maxwell R. Quality assessment in health. *Br Med J*. 1984;288(6428):1470.
- Medicare.gov. The official U.S. Government Site for Medicare. Safe Surgery Checklist Use. In: Medicare.gov. 2014. <http://www.medicare.gov/hospitalcompare/hospital-safe-surgery-checklist.html?AspxAutoDetectCookieSupport=1>. Accessed 4 May 2015.
- Pucher P, Rajesh A, Pritam S, Ara D. Enhancing surgical performance outcomes through process-driven care: a systematic review. *World J Surg*. 2014;38(6):1362–73.
- Schiff GD, Rucker D. Beyond structure–process–outcome: Donabedian’s seven pillars and eleven buttresses of quality. *Jt Comm J Qual Patient Saf*. 2001;27(3):169–74.
- Sinclair A, Schymura M, Boscoe F, Yung R, Chen K, Roohan P, Tai E, Schrag D. Measuring colorectal cancer care quality for the publicly insured in New York State. *Cancer Med*. 2012;1(3):363–71. <https://doi.org/10.1002/cam4.30>.
- The Joint Commission Core Measure Sets. 2014a. [http://www.jointcommission.org/core\\_measure\\_sets.aspx](http://www.jointcommission.org/core_measure_sets.aspx). Accessed 19 Sept 2014.
- The Society of Thoracic Surgeons National Database. The Society of Thoracic Surgeons National Database. 2014. <http://www.sts.org/national-database>. Accessed 19 Sept 2014.
- Tufts Medical Center. Cost-Effectiveness Analysis Registry. In: Cost-Effectiveness Analysis Registry. 2014. <https://research.tufts-nemc.org/cear4/> Accessed 5 May 2015.

### Secondary Sources

- Abbas MA, Chang GJ, Read TE, Rothenberger DA, Garcia-Aguilar J, Peters W, Monson JR, Sharma A, Dietz DW, Madoff RD, Fleshman JW, Greene FL, Wexner SD, Remzi FH. Optimizing rectal cancer management: analysis of current evidence. *Dis Colon Rectum*. 2014;57(2):252–9. <https://doi.org/10.1097/dcr.000000000000020>.
- Adachi W, Lodolce AE. Use of failure mode and effects analysis in improving the safety of iv drug administration. *Am J Health-Syst Pharm*. 2005;62(9):917–22.
- Adamina M, Kehlet H, Tomlinson G, Senagore A, Delaney C. Enhanced recovery pathways optimize health outcomes and resource utilization: a meta-analysis of randomized controlled trials in colorectal surgery. *Surgery*. 2011;149(6):830–40. <https://doi.org/10.1016/j.surg.2010.11.003>.
- Aday L, Andersen R. A framework for the study of access to medical care. *Health Serv Res*. 1974;9(3):208.
- Agency for Healthcare Research and Quality. National Healthcare Quality & Disparities Report, 2008. US Department of Health and Human Services. 2009.
- Akinci M, Yilmaz KB, Kulah B, Seker GE, Ugurlu C, Kulacoglu H. Association of ventral incisional hernias with comorbid diseases. *Chirurgia*. 2013;108:807–11.
- Almoudaris A, Burns E, Bottle A, Aylin P, Darzi A, Vincent C, Faiz O. Single measures of performance do not reflect overall institutional quality in colorectal cancer surgery. *Gut*. 2013;62(3):423–9.
- American Cancer Society. What is cancer recurrence? In: When cancer comes back: cancer recurrence. 2014. <http://www.cancer.org/treatment/survivorshipduringandaftertreatment/understandingrecurrence/whenyourcancercomesback/when-cancer-comes-back-what-is-recurrence>. Accessed 7 Jul 2016.
- American College of Surgeons. American College of Surgeons (ACS). In: American College of Surgeons. 2014b. <https://www.facs.org/>. Accessed 19 Sept 2014.
- American College of Surgeons, Commission on Cancer, Surgical Care Improvement Project. Core measure dets. In: The Joint Commissions. 2014b. [http://www.jointcommission.org/surgical\\_care\\_improvement\\_project/](http://www.jointcommission.org/surgical_care_improvement_project/). Accessed 10 May 2015.
- American College of Surgeons. Commission on Cancer. In: American College of Surgeons. 2014c. <https://www.facs.org/quality-programs/cancer>. Accessed 19 Sept 2014.
- Apkon M, Leonard J, Probst L, DeLizio L, Vitale R. Design of a safer approach to intravenous drug infusions: failure mode effects analysis. *Qual Saf Health Care*. 2004;13(4):265–71.
- Aquina C, Kelly K, Probst C, Noyes K, Langstein H, Monson JR, Fleming F. Surgeon and facility volume play significant role in hernia recurrence and reoperation after open incisional hernia repair. SSAT 55th annual meeting, Chicago;2014a. 2–6 May 2014.
- Aquina C, Rickles A, Iannuzzi JC, Kelly K, Probst C, Noyes K, Monson JR, Fleming FJ. Centers of

- excellence have lower ostomy-related Nsquip. Tripartite Birmingham; 2014b. 30 June 30–3 July 2014.
- Atkin W. Options for screening for colorectal cancer. *Scand J Gastroenterol.* 2003;38(237):13–6.
- Atla P, Sheikh M, Mascarenhas R, Choudhury J, Mills P. Survival of patients with hepatocellular carcinoma in the San Joaquin Valley: a comparison with California Cancer Registry data. *Ann Gastroenterol.* 2012;25(2):138.
- Australian Commission on Safety and Quality in Health Care. *Windows Into Saf and Quality in Health Care.* 2008.
- Begg C, Cramer L, Hoskins W, Brennan M. Impact of hospital volume on operative mortality for major cancer surgery. *JAMA.* 1998;280(20):1747–51.
- Birkmeyer JD, Siewers AE, Finlayson EVA, Stukel TA, Lee Lucas F, Batista I, Gilbert Welch H, Wennberg DE. Hospital volume and surgical mortality in the United States. *N Engl J Med.* 2002;346(15):1128–37.
- Blackburn H. Research and demonstration projects in community cardiovascular disease prevention. *J Public Health Policy.* 1983;4:398–421.
- Borzecki A, Christiansen C, Chew P, Loveland S, Rosen A. Comparison of in-hospital versus 30-day mortality assessments for selected medical conditions. *Med Care.* 2010;48(12):1117–21.
- Brown A, Patterson D. To err is human. In: *Proceedings of the first workshop on evaluating and architecting system dependability (EASY'01).* 2001.
- Brown M, Riley G, Schussler N, Etzioni R. Estimating health care costs related to cancer treatment from SEER-Medicare data. *Med Care.* 2002;40(8):IV104–17. <https://doi.org/10.2307/3767931>.
- Brownson R, Smith C, Pratt M, Mack N, Jackson-Thompson J, Dean C, Dabney S, Wilkerson. Preventing cardiovascular disease through community-based risk reduction: the Bootheel Heart Health Project. *Am J Public Health.* 1996;86(2):206–13.
- Burger CD, Roger RK. “Ventilator bundle” approach to prevention of ventilator-associated pneumonia. *Mayo Clin Proc.* 2006;81(6):849–50. <https://doi.org/10.4065/81.6.849>.
- Burgmeier J. Failure mode and effect analysis: an application in reducing risk in blood transfusion. *Jt Comm J Qual Patient Saf.* 2002;28(6):331–9.
- Burton S, Brown G, Daniels I, Norman A, Mason B, Cunningham D. MRI directed multidisciplinary team preoperative treatment strategy: the way to eliminate positive circumferential margins? *Br J Cancer.* 2006;94(3):351–7.
- CA Society of thoracic Surgeons. California Cardiac Surgery and Intervention Project (CCSIP). In: California Cardiac Surgery Intervention project. 2014. <http://www.californiacardiacsurgery.com/CCSIP-2012/index.html>. Accessed 19 Sept 2014.
- Caglià P, Tracia A, Borzi L, Amodeo L, Tracia L, Veroux M, Amodeo C. Incisional hernia in the elderly: risk factors and clinical considerations. *Intern J Surg.* 2014;12(Suppl 2):S164–9.
- Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Ader D, Fries J, Bruce B, Rose M. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Med Care.* 2007;45(5 Suppl 1):S3.
- Cella D, Gershon R, Bass M, Rothrock N. What is assessment center. In: *Assessment Center.* 2014. <https://www.assessmentcenter.net/>. Accessed 7 July 2016.
- Centers for Medicare and Medicaid Services (CMS). Physician Quality Reporting System (PQRS): maintenance of Certification Program Incentive. In: *eHealth University.* 2014. [https://www.cms.gov/eHealth/downloads/eHealthU\\_PQRSMaintenanceCertification.pdf](https://www.cms.gov/eHealth/downloads/eHealthU_PQRSMaintenanceCertification.pdf). Accessed 6 Jul 2016.
- Chang J, Vines E, Bertsch H, Fraker D, Czerniecki B, Rosato E, Lawton T, Conant E, Orel S, Schuchter L, Fox K, Zieber N, Glick J, Solin L. The impact of a multidisciplinary breast cancer center on recommendations for patient management: the University of Pennsylvania experience. *Cancer.* 2001;91(7):1231–7.
- Chantler C. The role and education of doctors in the delivery of health care\*. *The Lancet.* 1999;353(9159):1178–81. [https://doi.org/10.1016/S0140-6736\(99\)01075-2](https://doi.org/10.1016/S0140-6736(99)01075-2).
- Cohen M, Bilimoria K, Ko C, Hall B. Development of an American College of Surgeons National Surgery Quality Improvement Program: morbidity and mortality risk calculator for colorectal surgery. *J Am Coll Surg.* 2009b;208(6):1009–16.
- Cohn JN. Introduction to surrogate markers. *Circulation.* 2004;109(25 Suppl 1):IV-20–21.
- Collaborative Endocrine Surgery Quality Improvement Collective. Collaborative Endocrine Surgery Quality Improvement Program (CESQIP). In: *The American Association of Endocrine Surgeons.* 2014. <http://cesqip.org/>. Accessed 19 Sept 2014.
- Coory M, Gkolia P, Yang I, Bowman R, Fong K. Systematic review of multidisciplinary teams in the management of lung cancer. *Lung Cancer.* 2008;60(1):14–21.
- Department of Health Office/Welsh. A policy framework for commissioning cancer services (Calman-Hine report). London: Department of Health; 1995.
- Department of Health. The NHS Cancer plan: a plan for investment, a plan for reform. In: *Publications.* 2000. [http://webarchive.nationalarchives.gov.uk/+www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyandGuidance/DH\\_4009609](http://webarchive.nationalarchives.gov.uk/+www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyandGuidance/DH_4009609). Accessed 19 Sept 2014.
- DeRosier J, Stalhandske E, Bagian JP, Nudell T. Using health care failure mode and effect analysis™: the VA National Center for Patient Safety’s prospective risk analysis system. *Jt Comm J Qual Patient Saf.* 2002;28(5):248–67. <http://www.patientsafety.va.gov/professionals/onthejob/hfmea.asp>.
- Devlin N, Appleby J. Getting the most out of PROMs: putting health outcomes at the heart of NHS decision-making. London: The King’s Fund; 2010.

- Drummond MF, Sculpher M, Torrance GW, O'Brien BJ, Stoddart G. *Methods for the economic evaluation of health care programmes*. 3rd ed. New York: Oxford University Press; 2005.
- EHealth University: Centers for Medicare & Medicaid Services. *Physician Quality Reporting System (PQRS): Maintenance of Certification Program Incentive*. CMS. 2014.
- Etzioni DA, Cannom RR, Madoff RD, Ault GT, Beart Jr RW. Colorectal procedures: what proportion is performed by American Board of Colon and Rectal Surgery-certified surgeons? *Dis Colon Rectum*. 2010;53(5):713–20.
- Etzioni DA, Young-Fadok TM, Cima RR, Wasif N, Madoff RD, Naessens JM, Habermann EB. Patient survival after surgical treatment of rectal cancer: impact of surgeon and hospital characteristics. *Cancer*. 2014;120(16):2472–81.
- Fawcett S, Lewis R, Paine-Andrews A, Francisco V, Richter K, Williams E, Copple B. Evaluating community coalitions for prevention of substance abuse: the case of project freedom. *Health Educ Behav*. 1997;24(6):812–28.
- Ferguson G, Eliasziw M, Barr H, Clagett P, Barnes R, Wallace C, Taylor W, Haynes B, Finan J, Hachinski V, Barnett H, for the North American Symptomatic Carotid Endarterectomy Trial Collaborators. The North American Symptomatic Carotid Endarterectomy Trial: surgical results in 1415 patients. *Stroke*. 1999;30(9):1751–8. <https://doi.org/10.1161/01.str.30.9.1751>.
- Finks J, Osborne N, Birkmeyer J. Trends in hospital volume and operative mortality for high-risk surgery. *N Engl J Med*. 2011;364(22):2128–37.
- Fletcher C. Failure mode and effects analysis: an interdisciplinary way to analyze and reduce medication errors. *J Nurs Adm*. 1997;27(12):19–26.
- Fromson M, Rana A, White R, Marshall A, Schutzer S, Healy W, Naas P, Daubert G, Lorio R, Parsley B. Bundled payments for care improvement initiative: the next evolution of payment formulations: AAHKS Bundled Payment Task Force. *J Arthroplasty*. 2013;28(8):157–65.
- Gabel M, Hilton N, Nathanson S. Multidisciplinary breast cancer clinics. Do they work? *Cancer*. 1997;79(12):2380–4.
- Gartner. *Big Data*. In: *It Glossary*. 2013. <http://www.gartner.com/it-glossary/big-data/>. Accessed 19 Sept 2014.
- Gatt M, Anderson A, Reddy B, Hayward-Sampson P, Tring I, MacFie J. Randomized clinical trial of multimodal optimization of surgical care in patients undergoing major colonic resection. *Br J Surg*. 2005;92(11):1354–62. <https://doi.org/10.1002/bjs.5187>.
- Goodman R, Wheeler F, Lee P. Evaluation of the Heart To Heart Project: lessons from a community-based chronic disease prevention project. *Am J Health Promot*. 1995;9(6):443–55.
- Goss P, Lee B, Badovinac-Crnjevic T, Strasser-Weippl K, Chavarri-Guerra Y, Louis J, Villarreal-Garza C, Unger-Saldana K, Ferreyra M, Debiassi M, Liedke P, Touya D, Werutsky G, Higgins M, Fan L, Vasconcelos C, Cazap E, Vallejos C, Mohar A, Knaul F, Arreola H, Batura R, Luciani S, Sullivan R, Finkelstein D, Simon S, Barrios C, Kightlinger R, Gelrud A, Bychkovsky V, Lopes G, Stefani S, Blaya M, Souza F, Santos F, Kaemmerer A, Azambuja E, Zorilla A, Murillo R, Jeronimo J, Tsu V, Carvalho A, Gil C, Sternberg C, Duenas-Gonzalez A, Sgroi D, Cuello M, Fresco R, Reis R, Maserà G, Gabus R, Ribeiro R, Knust R, Ismael G, Rosenblatt E, Roth B, Villa L, Solares A, Leon M, Torres-Vigil I, Covarrubias-Gomez A, Hernandez A, Bertolino M, Schwartzmann G, Santillana S, Esteva F, Fein L, Mano M, Gomez H, Hurlbert M, Durstine A, Azenha G. Planning cancer control in Latin America and the Caribbean. *Lancet Oncol*. 2013;14(5):391–436. [https://doi.org/10.1016/S1470-2045\(13\)70048-2](https://doi.org/10.1016/S1470-2045(13)70048-2).
- Group, COMMIT Research. *Community Intervention Trial for Smoking Cessation (COMMIT): summary of design and intervention*. *J Natl Cancer Inst*. 1991;83(22):1620–8.
- Grube B, Giuliano A. Observation of the breast cancer patient with a tumor-positive sentinel node: implications of the ACOSOG Z0011 trial. *Semin Surg Oncol*. 2001;20(3):230–7.
- Hannan E, Kilburn H, O'Donnell J, Lukacik G, Shields E. Adult open heart surgery in New York State: an analysis of risk factors and hospital mortality rates. *JAMA*. 1990;264(21):2768–74.
- Hannan E, Siu A, Kumar D, Kilburn H, Chassin M. The decline in coronary artery bypass graft surgery mortality in New York State: the role of surgeon volume. *JAMA*. 1995;273(3):209–13.
- Hannan E, Cozzens K, King S, Walford G, Shah N. The New York State cardiac registries history, contributions, limitations, and lessons for future efforts to assess and publicly report healthcare outcomes. *J Am Coll Cardiol*. 2012;59(25):2309–16.
- Hannan E, Farrell L, Wechsler A, Jordan D, Lahey S, Culliford A, Gold J, Higgins R, Smith C. The New York risk score for in-hospital and 30-day mortality for coronary artery bypass graft surgery. *Ann Thorac Surg*. 2013;95(1):46–52.
- Hannan EL, Kilburn H, Racz M, Shields E, Chassin MR. Improving the outcomes of coronary artery bypass surgery in New York State. *JAMA* 1994;271(10):761–6.
- Haynes A, Weiser T, Berry W, Lipsitz SR, Breizat A, Dellinger P, Herbosa T, Joseph S, Kibatala P, Lapitan M. A surgical safety checklist to reduce morbidity and mortality in a global population. *N Engl J Med*. 2009;360(5):491–9.
- Healey AN, Sevdalis N, Vincent CA. Measuring intraoperative interference from distraction and interruption observed in the operating theatre. *Ergonomics*. 2006;49(5–6):589–604.
- Hospital Consumer Assessment of Healthcare Providers and Systems. HCAHPS: Hospital consumer assessment of healthcare providers and systems. In: *Hospital Consumer Assessment of Healthcare Providers and*

- Systems. 2014. <http://www.hcahponline.org/home.aspx>. Accessed 19 Sept 2014.
- Howell E, Devaney B, McCormick M, Raykovich K. Back to the future: community involvement in the healthy start program. *J Health Polit Policy Law*. 1998;23(2):291–317.
- Howell A, Panesar S, Burns E, Donaldson L, Darzi A. Reducing the burden of surgical harm: a systematic review of the interventions used to reduce adverse events in surgery. *Ann Surg*. 2014;259(4):630–41.
- Hurtado M, Swift E, Corrigan J. Crossing the quality chasm: a new health system for the 21st century. Institute of Medicine, Committee on the National Quality Report on Health Care Delivery. 2001.
- Iannuzzi J, Chandra A, Rickles A, Kumar N, Kelly K, Gillespie D, Monson J, Fleming F. Resident involvement is associated with worse outcomes after major lower extremity amputation. *J Vasc Surg*. 2013a;58(3):827–31.e1. <https://doi.org/10.1016/j.jvs.2013.04.046>.
- Iannuzzi J, Rickles A, Deeb A, Sharma A, Fleming F, Monson J. Outcomes associated with resident involvement in partial colectomy. *Dis Colon Rectum*. 2013b;56(2):212–8. <https://doi.org/10.1097/DCR.0b013e318276862f>.
- Iannuzzi J, Rickles A, Kelly K, Rusheen A, Dolan J, Noyes K, Monson J, Fleming F. Perioperative pleiotropic statin effects in general surgery. 2013c. <https://doi.org/10.1016/j.surg.2013.11.008>.
- Iannuzzi J, Young K, Kim M, Gillespie D, Monson J, Fleming F. Prediction of postdischarge venous thromboembolism using a risk assessment model. *J Vasc Surg*. 2013d;58(4):1014–20.e1. <https://doi.org/10.1016/j.jvs.2012.12.073>.
- Iannuzzi J, Chandra A, Kelly K, Rickles A, Monson J, Fleming F. Risk score for unplanned vascular readmissions. *J Vasc Surg*. 2014a. <https://doi.org/10.1016/j.jvs.2013.11.089>.
- Iannuzzi J, Rickles A, Kelly K, Fleming F, Dolan J, Monson J, Noyes K. Defining high risk: cost-effectiveness of extended-duration thromboprophylaxis following major oncologic abdominal surgery. *J Gastrointest Surg*. 2014b;18(1):60–8. <https://doi.org/10.1007/s11605-013-2373-4>.
- Iezzoni L. Risk adjusting rehabilitation outcomes: an overview of methodologic issues. *Am J Phys Med Rehabil*. 2004;83(4):316–26.
- Iezzoni L, Long-Bellil L. Training physicians about caring for persons with disabilities: “Nothing about us without us!”. *Disabil Health J*. 2012;5(3):136–9.
- Institute for Healthcare Improvement. In: Raising the bar with bundles: treating patients with an all-or-nothing standard. Institute for Healthcare Improvement. 2006. [www.ihf.org/IHI/Topics/CriticalCare/Intensive](http://www.ihf.org/IHI/Topics/CriticalCare/Intensive). Accessed 9 June 2014.
- Jensen L, Prasad L, Abcarian H. Cost-effectiveness of laparoscopic vs open resection for colon and rectal cancer. *Dis Colon Rectum*. 2012;55(10):1017–23.
- Johnston J, Marmet P, Coen S, Fawcett S, Harris K. Kansas LEAN: an effective coalition for nutrition education and dietary change. *J Nutr Educ*. 1996;28(2):115–8.
- Johnston M, Arora S, King D, Stroman L, Darzi A. Escalation of care and failure to rescue: a multicenter, multiprofessional qualitative study. *Surgery*. 2014;155(6):989–94.
- Katz J, Barrett J, Mahomed N, Baron J, Wright J, Losina E. Association between hospital and surgeon procedure volume and the outcomes of total knee replacement. *J Bone Joint Surg*. 2004;86(9):1909–16.
- Kelly K, Iannuzzi J, Rickles A, Monson J, Fleming F. Risk factors associated with 30-day postoperative readmissions in major gastrointestinal resections. *J Gastrointest Surg*. 2014a;18(1):35–44. <https://doi.org/10.1007/s11605-013-2354-7>.
- Kelly KN, Fleming FJ, Aquina CT, Probst CP, Noyes K, Pegoli W, Monson JRT. Disease severity, not operative approach, drives organ space infection after pediatric appendectomy. *Ann Surg*. 2014b;260(3):466–73.
- Korndorffer J, Fellingner E, Reed W. SAGES guideline for laparoscopic appendectomy. *Surg Endosc*. 2010;24(4):757–61.
- Kunac D, Reith D. Identification of priorities for medication safety in neonatal intensive care. *Drug Saf*. 2005;28(3):251–61.
- Leatherman S, Sutherland K. The quest for quality in the NHS: a mid-term evaluation of the ten-year quality agenda. London: Stationary Office; 2003.
- LeBlanc K, Booth W, Whitaker J, Bellanger D. Laparoscopic incisional and ventral herniorrhaphy in 100 patients. *Am J Surg*. 2000;180(3):193–7.
- Lee P, Regenbogen S, Gawande A. How many surgical procedures will Americans experience in an average lifetime? Evidence from three states. Massachusetts Chapter of the American College of Surgeons 55th Annual Meeting, Boston. 2008.
- Lee H, Vlaev I, King D, Mayer E, Darzi A, Dolan P. Subjective well-being and the measurement of quality in healthcare. *Soc Sci Med*. 2013;99:27–34.
- Levit L, Balogh E, Nass S, Ganz P. Delivering high-quality cancer care: charting a new course for a system in crisis. Washington, DC: National Academies Press; 2013.
- Lindenaier P, Remus D, Roman S, Rothberg M, Benjamin E, Ma A, Bratzler D. Public reporting and pay for performance in hospital quality improvement. *N Engl J Med*. 2007;356(5):486–96.
- Luft H, Bunker J, Enthoven A. Should operations be regionalized? The empirical relation between surgical volume and mortality. *N Engl J Med*. 1979;301(25):1364–9.
- Mack J, Chen K, Boscoe F, Gesten F, Roohan P, Weeks J, Schymura M, Schrag D. Underuse of hospice care by medicaid-insured patients with stage IV lung cancer in New York and California. *J Clin Oncol*. 2013;2012(45):9271.
- Marescaux J, Leroy J, Rubino F, Smith M, Vix M, Simone M, Mutter D. Transcontinental robot-assisted remote

- telesurgery: feasibility and potential applications. *Ann Surg.* 2002;235(4):487.
- Martin S, Heneghan H, Winter D. Systematic review and meta-analysis of outcomes following pathological complete response to neoadjuvant chemoradiotherapy for rectal cancer. *Br J Surg.* 2012;99(7):918–28.
- Mayer J, Soweid R, Dabney S, Brownson C, Goodman R, Brownson R. Practices of successful community coalitions: a multiple case study. *Am J Health Behav.* 1998;22(5):368–77.
- Mayo Clinic Staff. Cancer survival rate: What it means for your prognosis. In: *Diseases and Conditions Cancer.* 2016. <http://www.mayoclinic.org/diseases-conditions/cancer/in-depth/cancer/art-20044517>. Accessed 7 Jul 2016.
- McDermott R, Mikulak R, Beauregard M. *The basics of FMEA: quality resources.* New York: CRC Press; 1996.
- McNally K, Page M, Sunderland V. Failure-mode and effects analysis in improving a drug distribution system. *Am J Health-Syst Pharm.* 1997;54(2):171–7.
- Morris E, Haward R, Gilthorpe M, Craigs C, Forman D. The impact of the Calman-Hine report on the processes and outcomes of care for Yorkshire’s colorectal cancer patients. *Br J Cancer.* 2006;95(8):979–85. <https://doi.org/10.1038/sj.bjc.6603372>.
- Muhlestein D. What type of hospitals have high charge-to-reimbursement ratios? In: *Health Affairs Blog.* 2014. 2013. <http://healthaffairs.org/blog/2013/07/15/what-types-of-hospitals-have-high-charge-to-reimbursement-ratios/>. Accessed 19 Sept 2014.
- Mukamel D, Mushlin A. Quality of care information makes a difference: an analysis of market share and price changes after publication of the New York State Cardiac Surgery Mortality Reports. *Med Care.* 1998;36(7):945–54.
- Nagpal K, Vats A, Ahmed K, Smith A, Sevdalis N, Jonannsson H, Vincent C, Moorthy K. A systematic quantitative assessment of risks associated with poor communication in surgical care. *Arch Surg.* 2010;145(6):582–8.
- Nallamotheu B, Bates E, Wang Y, Bradley E, Krumholz H. Driving times and distances to hospitals with percutaneous coronary intervention in the United States implications for prehospital triage of patients with ST-elevation myocardial infarction. *Circulation.* 2006;113(9):1189–95.
- National Cancer Institute. SEER-Medicare: Brief description of the SEER-medicare database. In: *Healthcare Delivery Research.* 2014. <http://healthcaredelivery.cancer.gov/seermedicare/overview/>. Accessed 5 May 2014.
- National Health Services. Consultant outcome data. In: *My NHS.* 2015. <http://www.nhs.uk/choiceintheNHS/Yourchoices/consultant-choice/Pages/consultant-data.aspx>. Accessed 14 May 2015.
- National Institute of Health. SEER-Medicare linked database. In: *Healthcare Delivery Research.* 2014. <http://appliedresearch.cancer.gov/seermedicare/>. Accessed 15 May 2015.
- National Institute of Health. PROMIS: patient-reported outcomes measurement information system. In: *Programs.* 2015a. <http://commonfund.nih.gov/promis/index>. Accessed 15 May 2015.
- National Institute of Health. SEER-Medicare: brief description of the SEER-Medicare database. In: *Healthcare Delivery Research.* 2015b. <http://healthcaredelivery.cancer.gov/seermedicare/overview/>. Accessed 7 Jul 2016.
- New York State Department of Health. Cardiovascular disease data and statistics. In: *New York State Department of Health.* 2014. <https://www.health.ny.gov/statistics/diseases/cardiovascular/>. Accessed 7 Jul 2016.
- Nussbaum D, Speicher P, Ganapathi A, Englum B, Keenan J, Mantyh C, Migaly J. Laparoscopic versus open low anterior resection for rectal cancer: results from the national cancer data base. *J Gastrointest Surg.* 2014;19(1):124–31. <https://doi.org/10.1007/s11605-014-2614-1>.
- Optimizing the Surgical Treatment of Rectal Cancer (OSTRiCh). OSTRiCh consortium. In: *OSTRiCh Consortium.* 2014. [http://www.ostrichconsortium.org/news\\_archive.htm#.VByQjBYglhU](http://www.ostrichconsortium.org/news_archive.htm#.VByQjBYglhU). Accessed 19 Sept 2014.
- Press Ganey Associates, Inc. HCAHPS Insights. In: *Our solutions.* 2014. <http://www.pressganey.com/ourSolutions/patient-voice/regulatory-surveys/hcahps-survey.aspx>. Accessed 19 Sept 2014.
- Pronovost P, Needham D, Berenholtz S, Sinopoli D, Chu H, Cosgrove S, Sexton B, Hyzy R, Welsh R, Roth G, Bander J, Kepros J, Goeschel C. An intervention to decrease catheter-related bloodstream infections in the ICU. *N Engl J Med.* 2006;355(26):2725–32. <https://doi.org/10.1056/NEJMoa061115>.
- Rea R, Falco P, Izzo D, Leongito M, Amato B. Laparoscopic ventral hernia repair with primary transparietal closure of the hernial defect. *BMC Surg.* 2012;12 Suppl 1:S33.
- Reames B, Ghaferi A, Birkmeyer J, Dimick J. Hospital volume and operative mortality in the modern era. *Ann Surg.* 2014;260(2):244–51.
- Research Data Assistance Center. MCBS access to care. In: *Research Data Assistance Center.* 2014. <http://www.resdac.org/cms-data/files/mcbs-access-care>. Accessed 19 Sept 2014.
- Research Data Assistance Center. MCBS cost and use. In: *Research Data Assistance Center.* 2015. <http://www.resdac.org/cms-data/files/mcbs-cost-and-use>. Accessed 5 May 2015.
- Rickles A, Iannuzzi J, Kelly K, Cooney R, Brown D, Davidson M, Hellenthal N, Max C, Johnson J, DeTraglia J, McGurrin M, Kimball R, DiBenedetto A, Galyon D, Esposito S, Noyes K, Monson J, Fleming F. Anastomotic leak or organ space surgical site infection: what are we missing in our quality improvement programs? *Surgery.* 2013;154(4):680–7. <https://doi.org/10.1016/j.surg.2013.06.035>; discussion 687–9.

- Rosen M, Brody F, Ponsky J, Walsh R, Rosenblatt S, Duperier F, Fanning A, Siperstein A. Recurrence after laparoscopic ventral hernia repair. *Surg Endosc Other Intervent Tech*. 2003;17(1):123–8.
- Roussos S, Fawcett S. A review of collaborative partnerships as a strategy for improving community health. *Ann Rev Public Health*. 2000;21(1):369–402. <https://doi.org/10.1146/annurev.publhealth.21.1.369>.
- Rutter C, Johnson E, Feuer E, Knudsen A, Kuntz K, Schrag D. Secular trends in colon and rectal cancer relative survival. *J Natl Cancer Inst*. 2013;105:1806–13.
- Schone E, Brown R. Risk adjustment: what is the current state of the art and how can it be improved? In: Robert Wood Johnson Foundation. 2013. <http://www.rwjf.org/en/library/research/2013/07/risk-adjustment—what-is-the-current-state-of-the-art-and-how-c.html>. Accessed 19 Sept 2014.
- Schweizer M, Cullen J, Perencevich E, Vaughan S. Costs associated with surgical site infections in veterans affairs hospitals. *JAMA Surg*. 2014. <https://doi.org/10.1001/jamasurg.2013.4663>.
- Sharma R, Hawley C, Griffin R, Mundy J, Peters P, Shah P. Cardiac surgical outcomes in abdominal solid organ (renal and hepatic) transplant recipients: a case matched study. *Heart Lung Circ*. 2011;20(12):804–5.
- Sitzia J, Wood N. Patient satisfaction: a review of issues and concepts. *Soc Sci Med*. 1997;45(12):1829–43.
- Society for Surgery of the Alimentary Tract. The society for surgery of the alimentary tract. In: The Society for Surgery of the Alimentary Tract. 2016. <http://www.ssat.com/>. Accessed 6 Jul 2016.
- Society for Surgical Oncology. SSO: Society for surgical oncology. In: Society for Surgical Oncology. 2014. <http://www.surgonc.org/>. Accessed 19 Sept 2014.
- Solomon D, Losina E, Baron J, Fossel A, Guadagnoli E, Lingard E, Miner A, Phillips C, Katz J. Contribution of hospital characteristics to the volume–outcome relationship: dislocation and infection following total hip replacement surgery. *Arthritis Rheum*. 2002;46(9):2436–44.
- Stephens M, Lewis W, Brewster A, Lord I, Blackshaw G, Hodzovic I, Thomas G, Roberts S, Crosby T, Gent C, Allison M, Shute K. Multidisciplinary team management is associated with improved outcomes after surgery for esophageal cancer. *Dis Esophagus*. 2006;19(3):164–71. <https://doi.org/10.1111/j.1442-2050.2006.00559.x>.
- U.S. Department of Health & Human Services. Data. In: Organ Procurement and Transplantation Network. 2014. <http://optn.transplant.hrsa.gov/data/>. Accessed 7 Jul 2016.
- Ugiliweneza B, Kong M, Nosova K, Huang BA, Babu R, Klad SP, Boakye M. Spinal surgery: variations in healthcare costs and implications for episode-based bundled payments. *Spine*. 2014;39:1235–42.
- Vascular Quality Initiative. Improving vascular care. In: Society for Vascular Surgery. 2014. <http://www.vascularqualityinitiative.org/>. Accessed 19 Sept 2014.
- Wang Y, Jiang C, Guan J, Yang G, Yue J, Chen H, Xue J, Xu Z, Qian Q, Fan L. Molecular alterations of EGFR in small intestinal adenocarcinoma. *Int J Colorectal Dis*. 2013. <https://doi.org/10.1007/s00384-013-1689-6>.
- Wayne A, Lodolce A. Use of failure mode and effects analysis in improving the safety of IV drug administration. *Am J Health-Syst Pharm*. 2005;62(9):917–22.
- Wehrli-Veit M, Riley J, Austin J. A failure mode effect analysis on extracorporeal circuits for cardiopulmonary bypass. *J Extra Corpor Technol*. 2004;36(4):351–7.
- Weigl M, Antoniadis S, Chiapponi C, Bruns C, Sevdalis N. The impact of intra-operative interruptions on surgeons' perceived workload: an observational study in elective general and orthopedic surgery. *Surg Endosc*. 2015;29(1):145–53.
- Weinstein R, Linkin D, Sausman C, Santos L, Lyons C, Fox C, Aumiller L, Esterhai J, Pittman B, Lautenbach E. Applicability of healthcare failure mode and effects analysis to healthcare epidemiology: evaluation of the sterilization and use of surgical instruments. *Clin Infect Dis*. 2005;41(7):1014–9.
- Weir V. Best-practice protocols: preventing adverse drug events. *Nurs Manage*. 2005;36(9):24–30.
- Weiser T, Regenbogen S, Thompson K, Haynes A, Lipsitz S, Berry W, Gawande A. An estimation of the global volume of surgery: a modelling strategy based on available data. *Lancet*. 2008;372(9633):139–44.
- Weiser T, Semel M, Simon A, Lipsitz S, Haynes A, Funk L, Berry W, Gawande A. In-hospital death following inpatient surgical procedures in the United States, 1996–2006. *World J Surg*. 2011;35(9):1950–6.
- Wetterneck T, Skibinski K, Roberts T, Kleppin S, Schroeder M, Enloe M, Rough S, Hundt A, Carayon P. Using failure mode and effects analysis to plan implementation of smart IV pump technology. *Am J Health-Syst Pharm*. 2006;63(16):1528–38.
- Whitlock E, Lin J, Liles E, Beil T, Fu R. Screening for colorectal cancer: a targeted, updated systematic review for the US Preventive Services Task Force. *Ann Intern Med*. 2008;149(9):638–58.
- Wiegmann D, ElBardissi A, Dearani J, Daly R, Sundt III T. Disruptions in surgical flow and their relationship to surgical errors: an exploratory investigation. *Surgery*. 2007;142(5):658–65.
- Wille-Jorgensen P, Sparre P, Glenthøj A, Holck S, Norgaard Petersen L, Harling H, Stub Hojen H, Bulow S. Result of the implementation of multidisciplinary teams in rectal cancer. *Colorectal Dis*. 2013;15(4):410–3. <https://doi.org/10.1111/codi.12013>.
- World Alliance for Patient Safety. WHO surgical safety checklist and implementation manual. In: World Health Organization. 2014. [http://www.who.int/patientsafety/safesurgery/ss\\_checklist/en/](http://www.who.int/patientsafety/safesurgery/ss_checklist/en/). Accessed 7 Jul 2016.
- Yahchouchy-Chouillard E, Aura T, Picone O, Etienne J, Fingerhut A. Incisional hernias. *Digest Surg*. 2003;20(1):3–9.
- Zapka J, Marrocco G, Lewis B, McCusker J, Sullivan J, McCarthy J, Birch F. Inter-organizational responses to AIDS: a case study of the Worcester AIDS Consortium. *Health Educ Res*. 1992;7(1):31–46.



# Health Services Information: From Data to Policy Impact (25 Years of Health Services and Population Health Research at the Manitoba Centre for Health Policy)

# 8

Leslie L. Roos, Jessica S. Jarmasz, Patricia J. Martens, Alan Katz, Randy Fransoo, Ruth-Ann Soodeen, Mark Smith, Joshua Ginter, Charles Burchill, Noralou P. Roos, Malcolm B. Doupe, Marni Brownell, Lisa M. Lix, Greg Finlayson, and Maureen Heaman

## Contents

<b>Introduction</b> .....	172
<b>The Deliverable Process</b> .....	172
What Is a Deliverable? .....	172
Negotiating the Deliverable Topics .....	173
The Approval Process .....	173
Meetings .....	173
Presentations During the Project .....	174
Deliverable Measures and Indicators .....	176
<b>Highlights of Selected Deliverables</b> .....	177
The “Need to Know” Team Deliverables .....	177

Patricia J. Martens: deceased.

L. L. Roos (✉) · J. S. Jarmasz · A. Katz · R. Fransoo · R.-A. Soodeen · M. Smith · C. Burchill · N. P. Roos · M. B. Doupe · M. Brownell  
Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada  
e-mail: [Leslie\\_Roos@cpe.umanitoba.ca](mailto:Leslie_Roos@cpe.umanitoba.ca)

P. J. Martens  
Winnipeg, MB, Canada

J. Ginter  
Montreal, QC, Canada

L. M. Lix  
Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, Canada

G. Finlayson  
Finlayson and Associates Consulting, Kingston, ON, Canada

M. Heaman  
College of Nursing, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB, Canada

Manitoba's Indigenous Population .....	182
Hospitals, Emergency Departments, ICUs, and Long-Term Care .....	182
Maternal and Child Health .....	184
<b>Knowledge Translation (KT)</b> .....	187
The Repository .....	187
Other KT Activities .....	188
<b>Impact of Large Integrated Data Repositories</b> .....	188
Looking Ahead .....	189
Summing Up .....	189
<b>References</b> .....	190

## Abstract

The impact of the Manitoba Centre for Health Policy (MCHP) on policy development has resulted from an integrated approach to knowledge translation (KT), combined with a close relationship between the proposed/ongoing research and those working on provincial programs. Under a 5-year funding agreement with Manitoba Health, the director of MCHP negotiates five new major projects (called “deliverables”) annually with the Deputy Minister of Health. Researchers interact among themselves and with the provincial government in several ways: through forums, advisory group meetings, knowledge translation workshops, and *Need to Know* (NTK) team meetings. *Need to Know* representatives are from all the regional health authorities (RHAs), from Manitoba Health, and also from MCHP staff. This and other activities related to knowledge translation are discussed.

This chapter outlines steps in the deliverable process. MCHP researchers retain publication rights over the content of the deliverable with government input being advisory only. Several deliverables over the past 15 years, and their program and policy impacts, are discussed.

Finally, linking information from various government departments with longitudinal and familial data has created a large, integrated data repository. Looking ahead, life stage analyses and intervention studies have great potential. In keeping with past success, MCHP believes information-rich environments should continue to facilitate opportunities for new types of research and policy analysis.

## Introduction

The Manitoba Centre for Health Policy's (MCHP) impact on policy and program development is the result, in large part, of an integrated approach to knowledge translation. This chapter focuses on this integrated approach which has become one of the key factors underlying MCHP's success. This chapter begins with a description of the deliverable process and the numerous ways researchers interact with provincial government personnel. MCHP enjoys an arm's-length relationship with the provincial government, which has no involvement in the interpretation of data or drafting of deliverables (reports), and MCHP retains rights to publish all of its work. Next, the impact several deliverables have had on government policies and programs will be highlighted. Following this, an overview of the knowledge translation (KT) activities that have resulted in so many of MCHP's impacts is provided. To conclude, important and interesting research opportunities as well as challenges that lie ahead for scientists using information-rich repositories like ours will be discussed.

## The Deliverable Process

### What Is a Deliverable?

MCHP works under a 5-year funding agreement with Manitoba Health to undertake five new major research projects a year plus KT events that ensure the research is understood by



policy-makers and planners. These projects – termed deliverables – address health and social questions that can best be answered using data from the Population Health Research Data Repository (Repository) which is developed, housed, and maintained at MCHP (see ► Chap. 2, “Health Services Data: Managing the Data Warehouse: 25 Years of Experience at the Manitoba Centre for Health Policy”).

Each deliverable takes approximately 2 years to complete. Deliverables are produced by teams that typically include a principal and co-principal investigator (PI and Co-PI), a research coordinator (RC), research support (RS), and data analysts (DAs). Team members are typically chosen based on their area of expertise. Teams typically meet weekly or biweekly throughout the course of a deliverable to discuss the direction and progress of the study, interpret results, and determine how best to “tell the stories” that emerge from the data. A few times over the course of a deliverable, the team also meets with an “advisory group” made up of representatives from government and other stakeholders who have relevant expertise and can provide valuable feedback at different points in the research process (see section “Meetings” for more details).

## Negotiating the Deliverable Topics

Topics for deliverables are jointly determined by the Deputy Minister of Manitoba Health in negotiation with the director of MCHP. Consultations with assistant deputy ministers, MCHP scientists, and regional health authorities (RHAs) are undertaken when appropriate. The final list of topics is signed off by the Minister of Manitoba Health.

Ideas are solicited from a broad range of stakeholders. If the research seems feasible using repository data, the idea is added to a list. Specific topics are also put forward by Manitoba Health and the Healthy Child Committee of Cabinet (*Health Child Manitoba* is Manitoba’s long-term, cross-departmental strategy for putting families and children first). Negotiations typically start in the fall with final decisions made by the

following spring. At that time, Manitoba Health provides MCHP with a brief description of each deliverable. These descriptions are posted on the MCHP website in the area called “Upcoming MCHP Reports” [http://umanitoba.ca/faculties/health\\_sciences/medicine/units/community\\_health\\_sciences/departamental\\_units/mchp/upcoming\\_deliverables.html](http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departamental_units/mchp/upcoming_deliverables.html).

The associate director of research at MCHP works with the director to assign the investigators for each project. Soon after, a similar process is undertaken by the lead research coordinator, the associate director of data access and use, and the research support coordinator to identify the remaining team members from their respective workgroups (research coordinators, data analysts, and research support). Occasionally, deliverable teams will include graduate students or members outside of MCHP because of their expertise or interest in the topic.

## The Approval Process

The PI works with the deliverable team to develop an initial analysis plan, which is then presented and critically reviewed in a research-scientist forum held at MCHP. This forum is attended by internal researchers and team members who help refine the plan. The RC, in collaboration with the PIs, then prepares and submits the Health Information Privacy Committee (HIPC) and Health Research Ethics Board (HREB) applications for approval. Depending on the datasets to be used in the deliverable, additional approvals from other data providers may also be required. Throughout the life of the project, changes to the analysis plan (“amendments”) and annual progress reports must be submitted to HREB in order for the project to maintain its approved status.

## Meetings

### Meetings of the Advisory Group

An advisory group (AG) is also formed for each deliverable. It includes data providers, clinicians,

health or social service experts, provincial planners, policy-makers, RHA representatives, and other stakeholders with an interest in the topic. This group meets two to three times over the life of the project to review progress, discuss findings, suggest alternative strategies or approaches where necessary, provide clarifications based on their area of expertise, and review the final draft of the deliverable. It is also not uncommon for AG members to be contacted between meetings for their advice on specific issues. A strong relationship with policy-makers and other stakeholders also facilitates access to data and other nonfinancial resources that are important for the success of the research MCHP conducts.

The AG is a critically important group for MCHP; many times the real expertise concerning issues of data collection, history, and use lies with members of the AG. Their input provides an important check on any assumptions the deliverable team may have formed. Occasionally, depending on their contributions, AG members may also be recognized with authorship on the final report.

### **Meetings with the Associate Director of Research**

Throughout the project, PIs meet with the associate director of research regularly to discuss their projects and enlist support if projects are progressing slowly or running into problems. Two common challenges addressed at these meetings include the acquisition of new data or human resource issues (lack of resources, inappropriate skills or expertise, workload conflicts, etc.). These meetings also help to ensure that steady progress is being made and that expectations concerning deadlines are achievable.

### **Meetings with the *Need to Know* (NTK) Team**

A small number of deliverables involve the *Need to Know* Team (NTK Team), a collaborative researcher/senior-level-planner group that includes representatives from all RHAs, several representatives of Manitoba Health, and MCHP staff. The NTK Team was established in 2001 through funding from the Canadian Institutes of Health Research (CIHR) and has continued with

support from various other sources. The NTK Team meets three times a year for 2-day workshops, together creating knowledge of relevance to regional planners, informing the research, building capacity among the partners, and devising dissemination and application strategies to promote research uptake. Its foundation and goals are simple; by having researchers work with decision-makers, research may be brought closer to policy. In other words, the hope is to smooth the transition between analysis and application, between paper and practice. In 2005 the national “CIHR Knowledge Translation Award” was awarded to the NTK Team for regional impact on health research.

### **Presentations During the Project**

During the life of a typical deliverable, there are numerous opportunities to discuss the project, present preliminary results, and report on progress. Such opportunities include:

- MCHP knowledge translation workshop days – where invited guests consisting of government stakeholders meet with MCHP scientists and support staff to discuss deliverables
  - Provincial RHA Day
  - Winnipeg RHA Day
  - Manitoba Health Day
  - Manitoba Government Day
- Research forums – meetings where invited participants discuss the substantive merits of various research proposals and progress updates
  - Held weekly on Wednesday afternoons at MCHP
- NTK meetings (held two to three times a year, as discussed above)
- MCHP Advisory Board meetings (held biannually)
  - The board consists of five deputy ministers plus leading experts, other academic representatives, and the MCHP executive group.

The main steps in the deliverable process are presented in Table 1.

**Table 1** Steps in the deliverable process

<b>Analysis plan and approvals</b>	<ul style="list-style-type: none"> <li>• Develop draft analysis plan</li> <li>• Present to researchers for discussion</li> <li>• Finalize analysis plan</li> <li>• Apply for approvals (HIPC, HREB, other data approvals as required)</li> </ul>
<b>Meetings</b> (ongoing throughout project)	<ul style="list-style-type: none"> <li>• Weekly/biweekly team meetings begin and continue until draft writing stage</li> <li>• Advisory group – two to three meetings over course of project</li> </ul>
<b>Data preparation and methodology planning</b> (ongoing throughout project)	<ul style="list-style-type: none"> <li>• Data cleaning/validation for new and established datasets</li> <li>• Start defining inclusion/exclusion criteria and defining outcome measures and independent variables, statistical methods, etc. (this is a somewhat iterative process throughout project)</li> </ul>
<b>Documentation</b> (ongoing throughout project)	<ul style="list-style-type: none"> <li>• DA(s) document SAS programs and output</li> <li>• RC documents methodology based on information provided in meetings, email correspondence, and annotated output</li> <li>• Team identifies concepts; DA/PI (with RC support as necessary) develops by end of project</li> <li>• RC/DA Identify and define key glossary terms</li> </ul>
<b>Presentations</b> (various times throughout project)	<ul style="list-style-type: none"> <li>• MCHP research-scientist forums (two to three per deliverable)</li> <li>• MCHP government knowledge transfer workshop days</li> <li>• Academic conferences</li> </ul>
<b>Draft report writing and prep for review</b>	<ul style="list-style-type: none"> <li>• Writing may be ongoing during the course of the deliverable but often occurs close to the end of the analysis</li> <li>• Internal review by deliverable team and senior reader (an MCHP researcher); feedback sent to PI and modified by PI</li> <li>• Identification of external reviewers</li> </ul>
<b>Delivery to Manitoba Health and External Review</b>	<p>Minimum 60 days before release</p> <ul style="list-style-type: none"> <li>• Finalized draft for Manitoba Health to review for factual accuracy and comments regarding the content</li> <li>• PI, in collaboration with the associate director of research, identifies external reviewers</li> <li>• Concurrently, copies sent to MCHP researchers, advisory group members, team members, senior reader, external reviewer(s), and relevant data providers</li> </ul>
<b>Four-page (or two-page) deliverable lay summary</b>	<p>Iterative process between PI and writer</p> <ul style="list-style-type: none"> <li>• Identify writer of lay summary</li> <li>• Copy of draft report provided to summary writer</li> </ul>
<b>Briefings</b>	<ul style="list-style-type: none"> <li>• Deputy Minister of Health</li> <li>• Manitoba Health – senior management and assistant deputy minister (other depts. also invited or they may request separate briefing)</li> <li>• Minister of Health (if requested)</li> <li>• WRHA (if their data were used and they requested a briefing)</li> <li>• Other briefings as requested or required</li> </ul>
<b>Editing and final report</b>	<ul style="list-style-type: none"> <li>• Draft revised per reviewers' feedback</li> <li>• In-house editor performs content edits and works with PI to finalize deliverable report</li> </ul>
<b>Final production</b>	<ul style="list-style-type: none"> <li>• Review of printers' proof; approval of printing</li> <li>• Similar process followed for deliverable lay summary</li> <li>• Layout and preparations for publishing by RS; report sent to printers</li> </ul>
<b>Release date and requirements</b>	<ul style="list-style-type: none"> <li>• PI and associate director, research consults with communication officer and research support lead to determine deliverable release date and if a media conference is necessary</li> <li>• Manitoba Health advised of release date</li> </ul>
<b>Dissemination</b>	<ul style="list-style-type: none"> <li>• Embargoed copies to Manitoba Health and select provincial government ministers ~1 week prior</li> <li>• Communications officer prepares media release</li> <li>• Release circulated to University's Public Affairs office and to local media</li> <li>• Communications Officer handles all media related requests for questions and interviews</li> <li>• Deliverable and all related content uploaded to MCHP website for public access</li> <li>• PI responds to interview requests</li> </ul>

An important component of the process is engaging communication during the course of each deliverable. The following summarizes most of the important KT activities:

#### 1. Research-based communication

- An advisory group consisting of academics, clinicians (where appropriate), and policy- or program-oriented stakeholders are involved in developing the content of each project.
- During final document review, at least one and possibly two external reviewers who are experts in the field are recruited to review the document.
- Presentations are made at academic conferences.

#### 2. Dissemination to key decision-makers

- Consultations with the Deputy Minister (DM) of Health, Healthy Child Committee of Cabinet, and KT forums are used to disseminate results and collect ideas for future research.
- Core research teams frequently include clinical and policy- or program-oriented content experts.
- The MCHP director briefs the Deputy Minister of Health during regular bimonthly meetings.
- Prior to release, the PI briefs the assistant DM, Manitoba Health, and other stakeholders
- During the project, numerous briefings are given at government KT workshops.

#### 3. Public dissemination

- A four-page (or two-page) deliverable summary aimed at a lay audience who may be interested in the project is developed.
- A one-page “physician briefing” may be developed if relevant.
- An infographic is designed and produced if relevant.
- A media release is prepared for the release date.
- The PI responds to media requests for information, comments, or interviews.

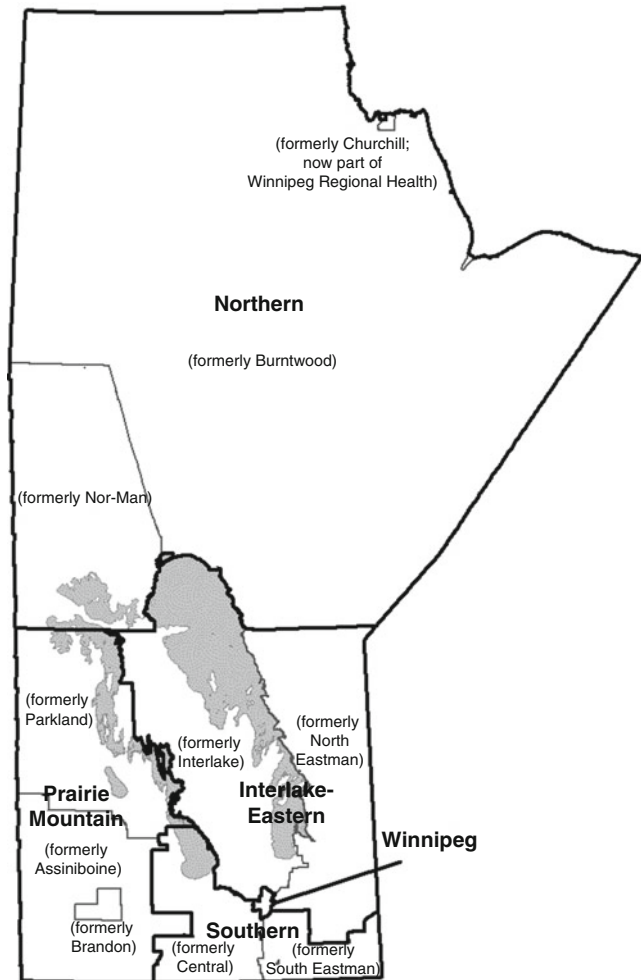
- Announcements are made through social media.
- All completed deliverables as well as research in progress are posted on the MCHP website: <http://mchp-appserv.cpe.umanitoba.ca/deliverablesList.html>

In addition, MCHP delivers the report to Manitoba Health at least 60 days before release. This gives the government time to prepare a response to the findings. From an academic perspective, MCHP researchers retain full publication rights over the contents of the deliverable once it has been released and any input from government is advisory only. This arm’s-length relationship with government helps to maintain academic rigor in the development of the final product. Once released, most deliverables form the basis of peer-reviewed articles in academic journals. Prior to any dissemination, all publications and presentations are reviewed by the Manitoba Government (through the HIPC coordinator) (and if necessary by other government departments who have provided data) for privacy and confidentiality issues.

### **Deliverable Measures and Indicators**

Generally, MCHP analyzes data at the population level. This provides an opportunity to present results at a geographic level (i.e., by RHA and/or Winnipeg Community Areas, see Figs. 1 and 2). RHAs are given important information that allows them to improve practices, policies, and healthcare services in their particular region and to make comparisons between regions or with the province as a whole. Reports often consist of common indicators of population health status, healthcare use, and quality of care that are presented by socioeconomic status (SES). This allows policy-makers to compare populations that are less well-off (low socioeconomic status) to those who are better-off (high socioeconomic status) and to design programs and

**Fig. 1** Manitoba's Five Regional Health Authorities (RHAs) (former RHAs are shown in *brackets*)



practices that address inequities in the healthcare system. Table 2 provides a list of frequently included study indicators.

### Highlights of Selected Deliverables

This section provides an overview of MCHP deliverables (see Table 3) that have had specific or ongoing impacts on policy and programs in the Manitoba community. The deliverables highlighted were published within the last 15 years (2000-2014) and there were no major criteria for their selection. Only deliverables with a concrete example of impact on policy and programs in Manitoba were described.

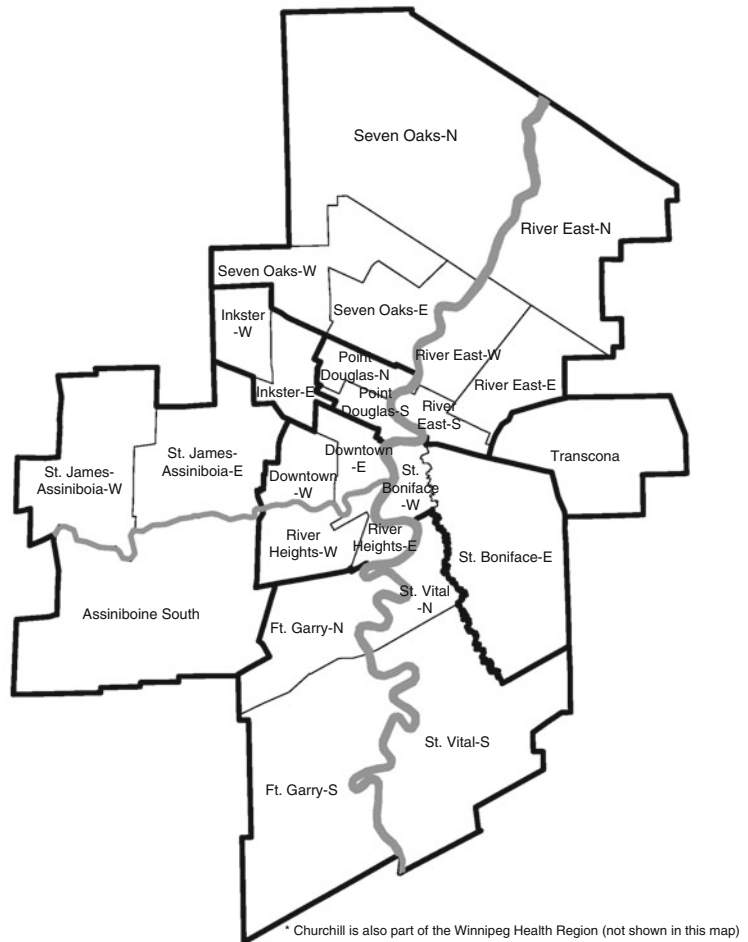
### The “Need to Know” Team Deliverables

As described above, a small number of all deliverables involve the *Need to Know Team* (NTK Team), a collaborative researcher/senior-level-planner group that includes representatives from all RHAs, several representatives of Manitoba Health, and MCHP staff.

### The RHA Indicators Atlas Reports

The NTK Team is an important component of the RHA Atlas deliverables. The Manitoba RHA Indicators Atlas reports provide regional and subregional data on over 50 indicators of population health status, health service use,

**Fig. 2** Winnipeg  
Community Areas (WCAs)



and quality of care. These reports provide RHAs with information on which to plan, increasing the likelihood that they will achieve their goals, and allow all RHAs to compare their health status with regional and provincial averages. The three atlases (see Table 3 a–c) were commissioned by Manitoba Health to inform the Comprehensive Community Health Assessment (CHA) reports required by provincial legislation every 5 years.

The atlases for CHA reporting are also used to develop RHA strategic plans. Over the years, numerous regions have told MCHP that resource allocation plans have been informed by evidence from our reports (e.g., the need to increase resources or support in some areas, while reducing them in others).

The establishment and early work of the NTK Team also resulted in organizational effects in all three partners (academic, provincial government, and RHAs). Several RHAs revised job descriptions and responsibilities to allocate more time and energy to finding and using evidence to inform decisions. At least one RHA actually created a new full-time position for this type of work. RHA representatives on the NTK Team are extremely valuable members of advisory groups for other deliverables, as they already have an established appreciation of the repository's data and its possible uses. The team also increased the effectiveness and efficiency of the CHA network group, which has many representatives in common with the NTK Team. Each atlas has resulted in a round

**Table 2** Frequently used health indicators in MCHP deliverables

Mortality	Quality of Primary Care
Total Mortality	Antidepressant Prescription Follow-Up
Premature Mortality	Asthma Care: Controller Medication Use
Causes of Mortality	Diabetes Care: Eye Examinations
Life Expectancy	Post-AMI Care: Beta-Blocker Prescribing
Potential Years of Life Lost (PYLL)	Benzodiazepine Prescribing for Community-Dwelling Seniors
Suicide	Benzodiazepine Prescribing for Residents of Personal Care Homes (PCH)
Illnesses, Diseases and Chronic Conditions	Immunizations and Prescription Drug Use
Diabetes	Influenza Immunization (Vaccination)**
Lower Limb Amputations Among Diabetics	Pneumococcal Immunization (Vaccination)**
Hypertension	Complete Immunization (Vaccination) Schedule (Ages 1, 2, 7, and 11)***
Total Respiratory Morbidity (TRM)	Number of Different Types of Drugs Dispensed per User
Asthma	Pharmaceutical Use
Arthritis	Cost of prescription drug use
Osteoporosis	Antibiotic Prescriptions
Multiple Sclerosis	Antidepressant Prescriptions
Stroke	Antipsychotic Prescriptions
Congestive Heart Failure (CHF)	Opioid Prescriptions
Coronary Heart Disease (CHD)/Ischemic Heart Disease (IHD)	Benzodiazepine Prescriptions
Acute Myocardial Infarction (AMI)	
Inflammatory Bowel Disease (IBD)	Preventive Care and Screening
Dialysis Initiation	Complete Physicals
Obesity	Breast Cancer Screening (Mammography)**
Cancer	Cervical Cancer Screening (Papanicolaou) (PAP) test**
Mental Illness	Long Term Care and Home Care
Substance Abuse	Supply of PCH Beds
Depression	Personal Care Home (PCH) Admissions
Mood and Anxiety Disorders	Personal Care Home (PCH) Residents
Personality Disorders	Median Waiting Times for PCH Admission from Hospital
Schizophrenia	Median Waiting Times for PCH Admission from the Community
Dementia	Level of Care on Admission to PCH
Suicide Attempt	Median Length of Stay by Level of Care on Admission to PCH
	Location: Where Residents Went for Personal Care Home (PCH)
	Catchment: Where Patients Came From Prior to Admission to Personal Care Home (PCH)
	Home Care
	Days of Home Care Service Received
Physician Services	Maternal and Child Health
Use of physicians/physician visits	Prenatal and Family Risk Factors (Family First Data)
Reasons for physician visits	Level of Maternal Education
Ambulatory Visits	Teen Pregnancy
Ambulatory Consultations	Maternal Depression/Anxiety
Majority of Care	Social Isolation
Continuity of Care	Relationship Distress
Location of Visits of General and Family Practitioners	Sexually Transmitted Infections
Location of Visits to Specialists	Prenatal Care
	Prenatal Alcohol Use
	Prenatal Smoking
	Breastfeeding Initiation
	Infant and Child Mortality
	Size for Gestational Age
	Newborn Hospital Readmission
	Children Not Ready for School (in One or More Early Development Instrument (EDI) Domains)
	Attention-Deficit Hyperactivity Disorder (ADHD)
	Asthma Prevalence
High Profile Surgical and Diagnostic Services	Education
Cardiac Catheterization	Early Development Instrument (EDI)
Percutaneous Coronary Interventions (PCI)	Number of school changes
Coronary Artery Bypass Surgery	High School completion
Total Hip Replacement	Grade Repetition
Total Knee Replacement	Grade 3 Reading
Cataract Surgery	Grade 3 Numeracy
Caesarean Section	Grade 7 Mathematics
Cholecystectomy	Grade 8 Reading and Writing
Hysterectomy	Grade 9 Achievement Index
Dental Extractions Among Young Children*	Grade 12 Language Arts Standards Tests
Computed Tomography (CT) Scans	Grade 12 Math Standards Tests
Magnetic Resonance Imaging (MRI) Scans	

\*also considered as child health indicators

\*\*also considered as quality of primary care indicators

\*\*\*also considered as child health and quality of primary care indicators

Note: further indicator information can be found in the MCHP Glossary / Concept Dictionary online:

[http://umanitoba.ca/faculties/health\\_sciences/medicine/units/community\\_health\\_sciences/departamental\\_units/mchp/resources/concept\\_dictionary.html](http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departamental_units/mchp/resources/concept_dictionary.html)

of site visits. Almost every region has invited MCHP scientists to workshops in their RHAs to explore local results in depth and to discuss implications for policy and planning. Feedback

from these regional workshops suggests that the impacts are significant and long-lasting.

Several NTK Team members are also public health officers who train medical students and

**Table 3** Impact – the list of deliverables highlighted in this chapter

#	Deliverable	Authors	Year published
(a)	The 2013 RHA Indicators Atlas	Fransoo R et al.	2013
(b)	Manitoba RHA Indicators Atlas 2009	Fransoo R et al.	2009
(c)	The Manitoba RHA Indicators Atlas: Population-Based Comparison of Health and Health Care Use	Martens PJ et al.	2003
(d)	Sex Differences in Health Status, Health Care Use, and Quality of Care: A Population-Based Analysis for Manitoba's Regional Health Authorities	Fransoo R et al.	2004
(e)	Patterns of Regional Mental Illness Disorder Diagnoses and Service Use in Manitoba: A Population-Based Study	Martens PJ et al.	2004
(f)	Profile of Metis Health Status and Healthcare Utilization in Manitoba: A Population-Based Study	Martens PJ, Bartlett J et al.	2010
(g)	The Health and Health Care Use of Registered First Nations People Living in Manitoba: A Population-Based Study	Martens PJ et al.	2002
(h)	The Epidemiology and Outcomes of Critical Illness in Manitoba	Garland A et al.	2012
(i)	Population Aging and the Continuum of Older Adult Care in Manitoba	Doupe M et al.	2011
(j)	An Initial Analysis of Emergency Departments and Urgent Care in Winnipeg	Doupe M et al.	2008
(k)	Using Administrative Data to Develop Indicators of Quality Care in Personal Care Homes	Doupe M et al.	2006
(l)	Assessing the Performance of Rural and Northern Hospitals in Manitoba: A First Look	Stewart D et al.	2000
(m)	Perinatal Services and Outcomes in Manitoba	Heaman M et al.	2012
(n)	Next Steps in the Provincial Evaluation of the Baby First Program: Measuring Early Impacts on Outcomes Associated with Child Maltreatment	Brownell M et al.	2007
(o)	Assessing the Health of Children in Manitoba: A Population-Based Study	Brownell M et al.	2001
(p)	How Do Educational Outcomes Vary With Socioeconomic Status? Key Findings from the Manitoba Child Health Atlas 2004	Brownell M et al.	2004

Note: All deliverables are available on our website: <http://mchp-appserv.cpe.umanitoba.ca/deliverablesList.html>

residents in their communities. In these regions, trainees may develop reports on the health of the communities in which they are working; they are frequently referred to the RHA atlas reports as a key source of information. Other NTK Team members have used atlases at their regional board of directors meetings, tackling one or two chapters of the report at each of a series of meetings. This provides valuable education for board members and the opportunity for discussion with senior management.

The two most recent RHA atlases are also listed 2nd and 16th on the list of the top 20 downloaded deliverables from MCHP's website over a 5-year period (from April 1, 2009, to March 31, 2014) (see Table 4).

### Other NTK Team Deliverables

The sex differences report (see Table 3-d) may have also played some role in the Winnipeg

Regional Health Authority's (WRHA) deliberations regarding heart health services in the mid-late 2000s. There had been some movement toward creating a women's heart health center, based on other evidence (not coming from MCHP) demonstrating that female heart attack patients were not receiving the same level of service as their male counterparts. The MCHP report showed that this apparent sex bias was not actually real. Within every 5-year age group, female and male heart attack patients received the same level of care. The difference in intervention rates was driven solely by the fact that female patients are known to experience heart attacks at a much older age (8–10 years older) than males. Males were not being treated more aggressively than females, but rather, younger patients received more treatments than older patients, and the younger patients were more likely to be male.



**Table 4** The top 20 downloaded deliverables April 1, 2009, to March 31, 2014

Rank	Deliverable	Year published/ available online	Page views per year
1	Perinatal Services and Outcomes in Manitoba	November 2012	104,735
2	The 2013 RHA Indicators Atlas	October 2013	54,460
3	Social Housing in Manitoba: Part I and Part II	June 2013	46,841
4	Projecting Personal Care Home Bed Equivalent Needs in Manitoba Through 2036	October 2012	27,572
5	Profile of Metis Health Status and Healthcare Utilization in Manitoba: A Population-Based Study	June 2010	15,941
6	Health Inequities in Manitoba: Is the Socioeconomic Gap in Health Widening or Narrowing Over Time?	September 2010	11,195
7	Pharmaceutical Use in Manitoba: Opportunities to Optimize Use	December 2010	9,331
8	The Additional Cost of Chronic Disease in Manitoba	April 2010	6,432
9	Manitoba Child Health Atlas Update	November 2008	6,400
10	What Works? A First Look at Evaluating Manitoba's Regional Health Programs and Policies at the Population Level	March 2008	6,118
11	Effects of Manitoba Pharmacare Formulary Policy on Utilization of Prescription Medications	December 2009	6,107
12	Defining and Validating Chronic Diseases: An Administrative Data Approach	July 2006	6,031
13	Patterns of Regional Mental Illness Disorder Diagnoses and Service Use in Manitoba: A Population-Based Study	September 2004	5,334
14	Assessing The Health Of Children In Manitoba: A Population-Based Study	February 2001	5,213
15	Who is in our hospitals and why	September 2013	5,103
16	Manitoba RHA Indicators Atlas 2009	September 2009	4,975
17	The Health and Health Care Use of Registered First Nations People Living in Manitoba: A Population-Based Study	March 2002	4,906
18	How are Manitoba's Children Doing?	October 2012	4,832
19	Composite Measures/Indices of Health and Health System Performance	August 2009	4,756
20	Population Aging and the Continuum of Older Adult Care in Manitoba	February 2011	3,068

Note: PDF copies of all deliverables became available on the MCHP website in 1999  
Averaged page views per year, over the 5-year period

The mental illness report (see Table 3-e) was important for documenting and spreading the word about the high prevalence of mental illness in Manitoba and the high use of healthcare services by people with mental illness. This topic was identified as a high priority by the rural and northern RHAs and by the Deputy Minister of Health and assistant deputy ministers. Between 1997 and 2002, more than one in four Manitobans had at least one mental illness diagnosis and used nearly half of the days people spent in hospitals. Most of the services used were not for mental illness, but across the entire spectrum of physical illness as well. This added important evidence to the understanding of the

comorbidity of physical and mental illness. The timeliness and prominence of the report also resulted in its principal investigator, Dr. Patricia Martens, being invited to join the first Scientific Advisory Board for the Mental Health Commission of Canada.

The Mental Health Commission of Canada has used MCHP research in launching its national research project to find sustainable solutions for homeless people with mental health issues. MCHP was included as a key partner in the Winnipeg demonstration project: [http://www.mentalhealthcommission.ca/sites/default/files/At%252520Home%252520Report%252520Winnipeg%252520ENG\\_0.pdf](http://www.mentalhealthcommission.ca/sites/default/files/At%252520Home%252520Report%252520Winnipeg%252520ENG_0.pdf).

The mental illness report (see Table 3-e) also revealed that close to 83 % of nursing-home residents have at least one mental illness diagnosis, yet the most frequent users of psychiatrists are people 35–55 years old. The report indicated that planners may want to ensure that facility staff are trained to provide care to address mental health as well as physical health needs and that people in personal care homes are referred for treatment. This finding may have contributed to the decision by the provincial health Minister at the time, to invest more than \$40 million to implement a comprehensive strategy to improve the quality of care in Manitoba's personal care homes. The funding was pledged to hire 250 registered nurses, registered psychiatric nurses, and licensed practical nurses, 100 personal healthcare aides, and 50 allied healthcare professionals to increase the direct hours of care, strengthen the work environment for staff, and provide dementia education to staff and families: <http://news.gov.mb.ca/news/index.html?archive=&item=2707>.

### Manitoba's Indigenous Population

The Métis community makes up roughly 6 % of Manitoba's population. The Metis Health deliverable (see Table 3-f) explored the Metis community's health status and healthcare use, as well as many social indicators of health. Overall, Métis people living in Northern Manitoba were found to be less healthy compared to those living in the southeast region (South Eastman) (see Fig. 1). This deliverable drew the attention of the Manitoba Metis Federation (MMF), who were concerned with identifying regions and health areas needing improvement in order to better the health and well-being of the Métis community. The MMF worked alongside MCHP to produce this report as one element in the regional planning profiles and to provide a springboard for other studies. This was the first attempt in Canada to do a population-level Metis health assessment.

The Health of First Nations deliverable (see Table 3-g) with the approval and collaborative support of the Health Information and Research

Committee of the Assembly of Manitoba Chiefs studied the health of Manitoba's Registered First Nations people, identifying factors that contribute to differences in health. The study focused on the First Nations population as a group, as well as by Tribal Council and by on-reserve versus off-reserve populations. Comparisons were made to the Manitoba population across various health-related indicators. Compared to all other Manitobans, a Registered First Nations person's life expectancy was 8 years shorter, dying at a young age was more than doubled, the chance of developing diabetes was more than quadrupled, and the chance of having an amputation as a result of diabetes increased 16-fold. Hospitalization rates were doubled for Registered First Nations persons compared to all other Manitobans, and they are three times higher for hospitalizations due to injury. Overall, health status rates varied across tribal councils. However, premature mortality rates were lowest in the north and highest in the south. This finding was surprising due to the "reversed" association with geography; in many previous MCHP studies and other reports, the health of residents of Northern Manitoba was usually shown to be worse than those in the south. However, this report showed the opposite to be true: First Nations residents of the north were healthier than their counterparts in the south.

These findings have been extensively used by the Assembly of Manitoba Chiefs (AMC) health councils for planning.

### Hospitals, Emergency Departments, ICUs, and Long-Term Care

The epidemiology and outcomes of critical illness in Manitoba report (see Table 3-h) allowed linkage of the extensive clinical database created by the Department of Critical Care Medicine to the repository. This combination of data sources is unique, allowing a first-ever population-based exploration of the use of intensive care units (ICUs) and fostered the development of an ongoing research group. In this report, the entire population of Manitoba and all hospitals were assessed from 1999/2000 to 2007/2008. About

0.6 % of Manitoba adults are admitted to an ICU each year, which means that about 8 % of those in hospitals are assessed as needing ICU care. Over a 9-year period, ICU beds in Winnipeg were full less than 5 % of the time. Outside of Winnipeg, ICU beds were full less than 1 % of the time. The average age for ICU patients was 64 years and admission rates peaked at those 80 years of age. Overall, about two-thirds of adult ICU care was for patients 60 years and older and the annual number of ICU admissions have dropped slightly; however, the length of stay in ICU's has increased over time. Repeated need for ICU care was surprisingly common (15 %) and previous ICU patients were almost four times more likely to be admitted again to an ICU in the year after discharge. Finally, the most common reason for ICU admission was cardiovascular conditions, followed by sepsis, lung disorders, accidents or traumas, and poisonings. This exploratory deliverable was the first of its kind to link clinical data on ICU patients into a population-based repository; thus it created a globally unique and flexible research tool. This tool is being leveraged for use in research projects and graduate student theses. The results on ICU bed utilization confirmed that the number of ICU beds in the Winnipeg RHA was within the recommended range. The report has resulted in four published manuscripts (Garland et al. 2013, 2014a, b; Olafson et al. 2014), with one more underway. It has also fostered several related research projects which have received peer-reviewed funding and provided additional publications.

The population in Manitoba, as it is in other parts of Canada and the developed world, is rapidly aging. The population aging deliverable (see Table 3-i) looked at the use of home care, supportive housing, and personal care homes (also known as nursing homes) in Winnipeg MB from several perspectives. First, past rates in nursing-home use were used to create two scenarios which showed that nursing-home use will increase by 30–50 % by 2031, emphasizing the importance of developing strategies to continually reduce rates of nursing-home use. This work also revealed the clinical profile of current day nursing-home residents, showing the potential for supportive housing to offset nursing-

home use. While about 50 % of newly admitted nursing-home residents required weight-bearing help to complete activities of daily living (ADLs), about a quarter of new residents had at most moderate challenges across several clinical domains (e.g., ADLs, behavior, continence, cognitive performance). Furthermore, about 12 % of newly admitted nursing-home residents had the same clinical profile as supportive housing clients (i.e., minor ADL and/or cognitive challenges, with few needs in other clinical areas), suggesting the potential of supportive housing to offset nursing-home use, now and into the future. Collectively, these findings emphasized the need to develop appropriate transitional strategies across the older adult continuum of care, ensuring that people have access to the right care at the right time. Subsequently the Manitoba government announced two initiatives which may have been informed by this work:

- Advancing Continuing Care – A Blueprint to Support System Change  
<http://news.gov.mb.ca/news/?item=31246>
- Manitoba's Framework for Alzheimer's Disease and Other Dementias  
<http://news.gov.mb.ca/news/index.html?item=31385>

The analysis of emergency department's (see Table 3-j) has had several impacts. Manitoba Health approved funding for the Eastman RHA (see Fig. 1) to hire 2.1 equivalent full-time staff to support mental health services. This is due to the reports' finding that 54 % of frequent emergency department (ED) users (seven or more ED visits per year) have been diagnosed with two or more mental illnesses. The funding was approved for the placement of Registered Psychiatric Nurses in EDs. Manitoba Health designated a total of \$165,302 for the 2008/09 and 2009/10 budget years: <http://news.gov.mb.ca/news/index.html?archive=&item=4458>. The Canadian Health Services Research Foundation (CHSRF) included some of the primary findings of this deliverable in their publication on emergency room overcrowding: <http://www.cfhi-fcass.ca/sf-docs/default-source/mythbusters/Myth-Emergency-Rm-Overcrowding-EN.pdf?sfvrsn=0>.

The CHSRF also wrote about MCHP's ability to transform data into quality care and transfer information down the chain of command to those that could make the appropriate changes and improvements. Their report highlighted the approach the principal investigator Dr. Malcolm Doupe took in explaining the deliverable "Using Administrative Data to Develop Indicators of Quality Care in Personal Care Homes" (see Table 3-k) to the Brandon RHA personal care homes' managers and policy-makers. Results were seen immediately in the quality of care: a pneumonia care map was introduced; the region's "personal care forum" became more productive, setting goals and action plans and updating each other on their progress; and a program for better managing medications of new residents was introduced: [http://www.cfhi-fcass.ca/sf-docs/default-source/building-the-case-for-quality/TRANSFORMING\\_DATA\\_ENG\\_1.pdf?sfvrsn=0](http://www.cfhi-fcass.ca/sf-docs/default-source/building-the-case-for-quality/TRANSFORMING_DATA_ENG_1.pdf?sfvrsn=0).

The performance of rural and northern hospitals deliverable (see Table 3-l) showed that rural Manitobans do not use nearby hospitals. Across 68 rural hospitals, occupancy rates were below 60 % and some hospitals and health centers were keeping admitted patients for too long (low scores on discharge efficiency). In 2002 the Manitoba Government announced a pilot project with the Southeast Manitoba RHA to serve more surgery patients at two local hospitals in an effort to make better use of rural facilities and provide patient care closer to home: <http://www.gov.mb.ca/chc/press/top/2002/07/2002-07-09-01.html>.

## Maternal and Child Health

The Perinatal Services and Outcomes deliverable (see Table 3-m) has been the number one deliverable downloaded from the MCHP website (see Table 4). The WRHA Women's Health Program used the report to validate their initiatives and reiterate the importance of the prenatal period in promoting optimal early childhood development. Inadequate prenatal care is being addressed through the "Partners in Inner-city Integrated Prenatal Care (PIIPC)" initiative, stimulated in part by the high rates of inadequate care found in the

Winnipeg Community Areas of Point Douglas, Downtown, and Inkster (see Fig. 2). The deliverable included new information on rates of postpartum depression/anxiety in Manitoba, revealing that women who experienced anxiety or depression during their pregnancy were eight times more likely to experience it postpartum. The WRHA reaffirmed the Women's Health Program's efforts to ensure that information and resources are continuously available in the postpartum period to foster mental health. Staff in the Population Health and Health Equity and Public Health Program, administered by Manitoba Health, noted that the perinatal deliverable influenced their thinking about potential positive impacts of public health engagement early with families in the prenatal period; findings from the deliverable have been used to inform development of the provincial public health nursing standards. The WRHA is actively interested in reducing health inequities. They have been particularly interested in breastfeeding initiation. The perinatal deliverable highlighted variations in initiation rates across the city (e.g., over 90 % in an affluent neighborhood and approximately 65 % in a less affluent one). These variations were significant in motivating the WRHA to begin tracking breastfeeding initiation and duration rates across Winnipeg.

The Baby First deliverable (see Table 3-n) evaluated how well the Manitoba Baby First screening program (established in 1999, now called "Families First") works with regards to identifying children at risk. About 75 % of babies had a Baby First screening form filled out; the screen was reasonably successful in picking out children who eventually ended up in foster care. The strongest predictors of a child ending up in care were having a file with local child protection services, being on income assistance, having a mother who did not finish high school, and living in a one-parent family with no social support. Because the age of the mother at the birth of her first child was also found to be highly predictive (and was not currently being asked on the screening form), Healthy Child Manitoba responded to preliminary drafts of the report by adding this item to the screening form (see Fig. 3). In addition,

**Fig. 3** The revised 2007 families first screening form (Reproduced with permission from the Manitoba Government)

**Families First SCREENING Form** 2007

**NUMERICAL INFORMATION ONLY**  
Please do not write any names or addresses on this form. See detailed instructions on reverse

**MOTHER:** Age (years):   When was pregnancy confirmed (weeks)?   Screened prenatally?  Yes  No

**BABY:** Day   Month   Year     Birth Date:     2 0 0 7

PHIN:           MHSC:           Gender:  Male  Female

Residence Postal Code:       RHA   Community Area Code:   PHIN:

**FATHER:** Age (years):   Education:  Grade 12 and up  Less than Grade 12

Aboriginal child?  Yes  No  
Aboriginal group:  North American Indian  Metis  Inuit  Other Aboriginal

**A. CHILDREN WITH KNOWN DISABILITY** (Fill in 'yes' if risk factor is present, 'no' if it is not. If unknown, leave blank.)

1. Congenital anomaly or acquired disability. Include: Major (probability of permanent disability) e.g., Down's syndrome, cerebral palsy, FASD Moderate (correction may be possible) e.g., cleft palate, loss of limb  Yes  No

**B. DEVELOPMENTAL RISK FACTORS**

2. Low birth weight (less than 2500 grams at birth).  Yes  No  
3. High birth weight (greater than 4000 grams at birth).  Yes  No  
4. Prematurity - an infant born at less than 37 weeks gestation.  Yes  No

Complications of pregnancy

5. Infections that can be transmitted in utero and may damage the fetus (e.g., rubella).  Yes  No  
6. Alcohol use by mother during pregnancy. If "yes", complete section D.  Yes  No  
7. Drug use by mother during pregnancy.  Yes  No

Complications of labour and delivery

8. Difficult vaginal birth (forceps or vacuum) or emergency caesarean  Yes  No  
9. Infant trauma or illness (e.g., convulsions, respiratory distress syndrome)  Yes  No  
10. Family history of a disability not detectable at birth that could affect development (e.g., deafness, mentally disabled/challenged)  Yes  No  
11. Multiple births (e.g., twins, triplets)  Yes  No  
12. Maternal smoking during pregnancy  Yes  No

**C. FAMILY RISK FACTORS**

13. Mother's age at birth of first child is less than 18 years.  Yes  No  
14. Mother's highest level of education completed is less than grade 12.  Yes  No  
15. On social assistance/income support or financial difficulties.  Yes  No  
16. Single parent family.  Yes  No  
17. No prenatal care before sixth month.  Yes  No

Mental illness or disability in mother and/or father:

18. Depression (including postpartum) Mother  Yes  No Biological father of babe  Yes  No  
19. Anxiety Disorder Mother  Yes  No Biological father of babe  Yes  No  
20. Schizophrenia or bipolar affective disorder Mother  Yes  No Biological father of babe  Yes  No  
21. Mentally disabled/challenged parent Mother  Yes  No Biological father of babe  Yes  No  
22. Antisocial behaviour Mother  Yes  No Biological father of babe  Yes  No  
23. Current substance abuse by mother or father Mother  Yes  No Biological father of babe  Yes  No  
24. Prolonged postpartum maternal separation (5 days or more with little or no contact).  Yes  No  
25. Assessed lack of bonding (e.g., minimal eye contact, touching)  Yes  No  
26. Social isolation (lack of social support and/or isolation related to culture, language or geography).  Yes  No  
27. Relationship distress.  Yes  No  
28. Current or history of violence between parenting partners.  Yes  No  
29. Harsh and/or inappropriate discipline practices (including other children).  Yes  No  
30. Existing file with local child protective services.  Yes  No  
31. Mother's own history of child abuse/neglect.  Yes  No  
32. Father/parenting partner's own history of child abuse/neglect.  Yes  No

**D. ALCOHOL USE DURING PREGNANCY** (complete if answered "yes" to item B6) (See reverse for detailed instructions)

In this section, check the option that is most descriptive of alcohol use before mother knew she was pregnant.

Frequency How often did mother consume alcohol?  Less than once a month  1-4 days/month  2-3 days/week  > 3 days/week  
Usual Amount How much alcohol would she consume in one sitting?  1 to 2 drinks or less  3 or 4 drinks  5 or more drinks  
Binge Did she ever drink more than five drinks in one sitting?  Yes  No  
How often did binge drinking occur?  Less than once a month  1-4 days/month  2-3 days/week  > 3 days/week

Once she discovered her pregnancy, did how much or how often she consumed alcohol change? Select one response.  No  Yes, reduced use  Yes, increased use  Yes, stopped altogether

Screen Completed By: Name:  (please print) Health Unit #     Day   Month   Year     TOTAL SCORE

Reproduced with permission from the Manitoba Government, 2015

child maltreatment and assault injury rates in children up to 3 years of age declined after the Baby First home visiting program was initiated.

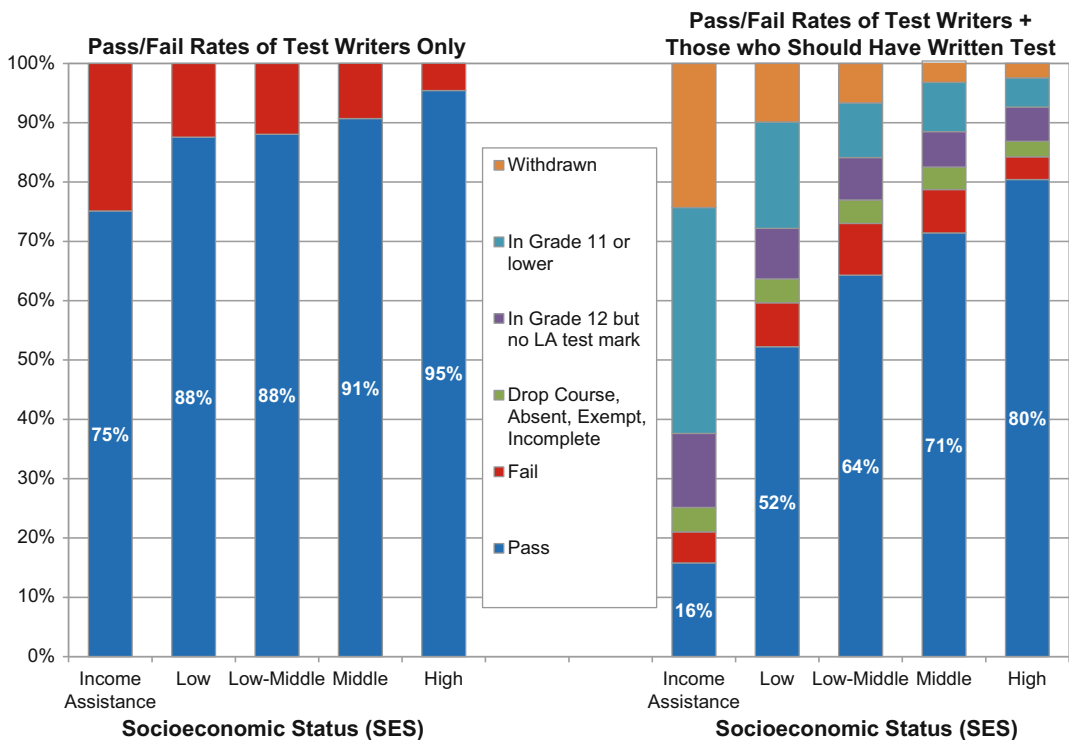
Poor health during childhood raises the risk of poor adult health. The Child Health Atlas

(see Table 3-o) found that infant mortality was double for the lowest-income areas compared to the highest-income areas, and the leading cause of death for children was injury due to motor vehicle crashes. Children living outside of Winnipeg are

twice as likely to die from injuries and almost two-and-a-half times as likely to be hospitalized for injuries. Because of these findings, Manitoba Health announced a new public initiative aimed at preventing childhood injuries in the home: <http://news.gov.mb.ca/news/index.html?item=25659&posted=2002-02-26>.

The children’s educational outcomes and socioeconomic status deliverable (see Table 3-p), which stemmed from the second Child Health Atlas, revealed some very surprising findings. This deliverable looked at performance on Grade 12 standard tests by socioeconomic status (SES) (see Fig. 4). The left side of Fig. 4 shows, for youths who wrote the test, that students from the poorest families (those receiving provincial income assistance) had a passing rate of 75 %, whereas students residing in the city’s highest-income neighborhoods

had a passing rate of 95 %. The right side of the graph not only includes those students who wrote the test, but more importantly, also includes those students born in the same year who are still residing in Winnipeg and who *should* have written the test had they progressed through the school system as expected. This population-based analysis shows a much steeper gradient, with the passing rates for youth in families on provincial income assistance dropping to 16 %. The two figures differ in that the one on the right includes those who have been held behind a grade or more or who have withdrawn from school. Such surprising findings demonstrate the need for better educational programs and initiatives for students from low-income families. This report, along with the Child Health Atlas, led to the development of two initiatives:



Note: A version of this figure has also been published in Roos, NP *et al.*, 2010, *Milbank Quarterly*, 88(3):382-403 and in Brownell, M *et al.*, *How Do Educational Outcomes Vary With Socioeconomic Status?* June 2004, *Manitoba Centre for Health Policy*

**Fig. 4** Grade 12 language arts (LA) test performance by Winnipeg socioeconomic status, 2001/02. Youths born in Manitoba in 1984

- The “Community School Investigators (CSI) program”  
<http://www.bgcwinnipeg.ca/system/resources/W1siZiIsIjIwMTQvMDEvMTYvMTgyMDQvMzUvNDE3L0NTSV9SZXBvcnRfMjAxMi5wZGYiXV0vCSI%20Report%202012.pdf> (p. 6)
- The Community Schools Partnership Initiative (CSPI)  
<http://www.edu.gov.mb.ca/cspi/>

Two additional child health atlases have been produced at MCHP since the 2001 and 2004 atlases: *The Child Health Atlas Update* (2008) (#9 in Table 4) which provided much needed information on child health for the annual Community Health Assessments and *How Are Manitoba's Children Doing?* (2012) (#18 in Table 4) which was a companion report to the legislated 5-year Healthy Child Manitoba report.

---

## Knowledge Translation (KT)

Situating MCHP within the University of Max Rady College of Medicine with ongoing, renewable core funding from the provincial government has allowed academic freedom, intellectual curiosity, and a high degree of research skill to combine with grounded work relevant to the questions of top-level decision-makers. The university also supports the work of MCHP through tenured or tenure-track faculty who work in the centre. Government input continues to be integral to the process of deciding the five deliverables funded by Manitoba Health annually. This model has been called “integrated knowledge translation (KT)” (Canadian Institute of Health Research (CIHR) 2014; Graham et al. 2007, 2009) and reflects the fact that users of the research are involved at the outset. If those individuals looking for answers have helped frame relevant questions with experienced researchers who know the limitations of the data, the scope of the literature, and what has already been done in the area, the findings are more likely to draw attention and result in action. Not only does the research have its feet on the ground, but it begins to walk (so to speak) because of the people involved. The findings are

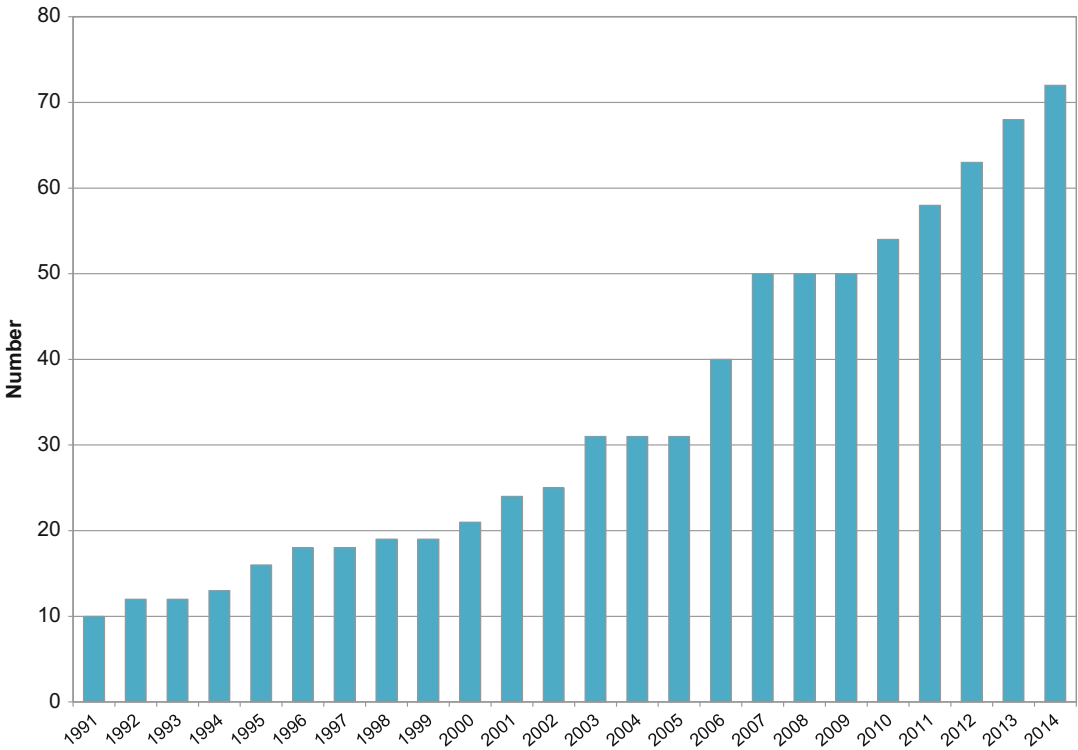
disseminated through the natural interest of the decision-makers involved in the programs or policies for which they are relevant. Although research evidence is not the only influence on policy (often other pressures, such as economic or political realities, override the evidence), if policy-makers and planners understand the research, there is a good chance it will be important in the decision-making process (Martens 2011).

Some people have expressed concerns about having policy- and decision-makers involved in the process from start to finish. What if they bias the results? What if they ask the wrong questions? What if they don't like the results? Such questions echoed our own fears in the early years. Through a combination of research funded from deliverables and our external grant-funded research from peer-reviewed granting agencies such as CIHR, Research Manitoba, and others, MCHP has learned that the best questions come from an exchange of ideas, both among researchers and between researchers and research users (Martens 2011).

## The Repository

MCHP continues to show leadership in the management of administrative data as it becomes the custodian of ever increasing numbers of de-identified (“anonymized”) but linkable datasets. Currently, the repository consists of more than 70 routinely collected administrative and clinical databases that are updated on an annual basis (see Fig. 5). The repository yields incredible opportunities to advance the understanding of complex relationships between population health and the use of health and social services (Martens 2011).

Documenting the repository and the concepts used in research projects are other forms of knowledge translation (KT). MCHP continuously dedicates resources to expand and improve documentation-related KT. Other researchers and policy analysts can read about, and even request, the statistical coding for various concepts that were derived using administrative data – such as how MCHP defines “continuity of care,” an “episode of care,” “comorbidity,” or “high school



**Fig. 5** The total count of databases in the repository by year

completion.” Accessibility in this respect continues to grow, evidenced by the fact that our concept dictionary and glossary receive more than 1.5 million hits a year (excluding bots and Web crawlers). This is a remarkably high frequency for a small academic unit (Martens 2011).

### Other KT Activities

MCHP has established a highly successful set of annual workshops attended by top-level planners, policy-makers, healthcare CEOs, VPs of planning, board members for RHAs, and front-line workers. These activities are based upon an interactive model of roundtable discussions concentrating on one or two MCHP reports. Attendees are encouraged to look for the stories in the data. Key to these workshop days is the presence of MCHP scientists to explain how to read the reports. In the book *Arabian Nights* by Tahir Shah, the author talks about his father explaining

the importance of stories to him as a child. “‘Stories are a way of melting the ice,’ [his father] said gently, ‘turning it into water. They are like repackaging something – changing its form – so that the design of the sponge can accept it’” (Shah 2007: 298). This is an apt metaphor for telling research stories. Sometimes providing a written report may not be enough. In these workshops, MCHP turns written reports into stories by explaining how to read the graphs, how to look for connections, or how to relate data to real-life settings. Repackaging the research allows it to be understood and incorporated into the audience’s way of thinking (Martens 2011).

### Impact of Large Integrated Data Repositories

Creation of a large integrated repository of data across multiple government domains has facilitated groundbreaking innovative research. Record



linkage has merged information from different departments while, at the same time, extensive longitudinal and familial data have allowed new types of studies and facilitated interdisciplinary work. The opportunities presented are unique advantages of large repositories.

## Looking Ahead

As seen in the discussion of deliverables, the very large numbers of cases that accumulate when such data are routinely gathered facilitate complicated multivariate analyses and allow studying low-prevalence conditions or events. Because these data are typically collected over long periods of time, pre- and post-observations can be organized around different life events at the individual level, and also before and after key program implementation – with a time frame extending for over 40 years in the case of the MCHP repository. Merging data across different ministerial departments can bring together individual information from several subject areas to create predictors useful in a variety of contexts (i.e., population-based research on ethnicity, developing risk assessment tools) and permit examining important connections affecting the lives of individuals and patients. Data documenting the use or lack of contact with the healthcare system and residential mobility data can be put together for any interval from 1 day to many years. A real but relatively unexplored advantage of the MCHP repository would be to follow those born in the 1970s, where the ability to track family structure events and health outcomes over the first decades of life is outstanding. This line of inquiry provides the possibility of life stage analysis: does a diagnosis of attention deficit disorder which first occurs at age 4–8 have a different impact on educational outcomes than a diagnosis which first occurs at age 9–12? How important is a chronic disease diagnosis, one which continues over time, compared with the same diagnosis occurring during only one age period?

There is great interest in improving both observational and interventional studies. In addition,

the population-based repository has great potential for “natural experiments” where administrative data may be used to consider the impact of policy and program changes. And research designs can be improved by building on the types of data available in Manitoba to construct control groups using propensity scores, sibling comparisons, and fine-grained ecological information. To date, such efforts are basically unexplored but have great potential for the future.

## Summing Up

Research platforms lend themselves to forming an “ecosystem,” “an intertwined set of products and services that work together” (El Akkad and Marlow 2012). The MCHP ecosystem involves relations with people, including key decision-makers, software (for data cleaning, record linkage, and analysis), the extensive documentation accessible through our concept dictionary and glossary, predictors and outcome measures derived from multiple files, and a methodological/statistical tool kit. New data in the Manitoba repository has expanded the type and number of studies being carried out. These capabilities foster useful interactions with a diversity of investigators; helping to avoid an overreliance on a single funding source and bringing in valuable new perspectives.

The approaches forwarded here seem generally relevant to “big data” where more attention needs to be paid to questions of design and analysis. The significant effort required to clean and prepare the databases should not be underestimated; Cukier and Mayer-Schoenberger have both noted the messiness of big data and highlighted the potential benefits of interagency collaboration in improving public services (Cukier and Mayer-Schoenberger 2013). The uses of population-based data are being more widely recognized. Information-rich environments should continue to facilitate opportunities for the next generation of researchers. That’s the *real* impact of MCHP’s academic and research history: building a culture where evidence informs policy in a way that works.

## References

- Canadian Institute of Health Research (CIHR). More about knowledge translation at CIHR. 2014. <http://www.cihr-irsc.gc.ca/e/39033.html>. Accessed 31 Oct 2014.
- Cukier K, Mayer-Schoenberger V. The rise of big data. *Foreign Aff.* 2013;92:28–40.
- El Akkad O, Marlow I. Apple at the summit: the trouble with being no. 1. 2012. <http://www.theglobeandmail.com/technology/apple-at-the-summit-the-trouble-with-being-no-1/article4546745/?page=all>
- Garland A, Olafson K, Ramsey CD, et al. Epidemiology of critically ill patients in intensive care units: a population-based observational study. *Crit Care.* 2013;17:R212.
- Garland A, Olafson K, Ramsey CD et al. A population-based observational study of ICU-related outcomes: with emphasis on post-hospital outcomes. *Ann Am Thorac Soc.* 2014a;12:202–208.
- Garland A, Olafson K, Ramsey CD et al. Distinct determinants of long-term and short-term survival in critical illness. *Intens Care Med.* 2014b;40:1097–105.
- Graham ID, Tetroe J, Gagnon M. Lost in translation: just lost or beginning to find our way? *Ann Emerg Med.* 2009;54:313–4.
- Graham ID, Tetroe J, KT Theories Research Group. Some theoretical underpinnings of knowledge translation. *Acad Emerg Med.* 2007;14:936–41.
- Martens PJ. Straw into gold: lessons learned (and still being learned) at the Manitoba Centre for Health Policy. *Healthcare Policy.* 2011;6:44–54.
- Olafson K, Ramsey C, Yogendran M et al. Surge capacity: analysis of census fluctuations to estimate the number of intensive care unit beds needed. *Health Serv Res.* 2014;50:237–252.
- Shah, T. In *Arabian Nights: A Caravan of Moroccan Dreams*. 1 edition. New York: Bantam Books; 2007.



# Health Services Information: Key Concepts and Considerations in Building Episodes of Care from Administrative Data

9

Erik Hellsten and Katie Jane Sheehan

## Contents

<b>Introduction</b> .....	192
<b>Health-Care Data and Defining the Unit of Analysis: Historical Perspective</b> .....	194
<b>The Episode of Care: A Unifying Concept</b> .....	195
<b>Episodes as an Analytical Tool: Advantages</b> .....	197
Flexibility .....	197
Comprehensiveness .....	197
Clinical Meaningfulness .....	197
<b>Episodes as an Analytical Tool: Challenges</b> .....	198
Data Requirements .....	198
Complexity .....	198
Time and Resources Required .....	198
Methodological Challenges .....	198
<b>Constructing an Episode of Care: Key Components</b> .....	199
Data Sources Required .....	199
Individual-Level Record Linkage .....	199
Information on Type of Service .....	199
Diagnosis Information .....	199
The Date/Time of the Service Delivered .....	199
Core Elements of the Episode .....	199
Defining the Index Event and/or Starting Point .....	200
Defining the Endpoint .....	200
Selecting the Scope of Services Included .....	202
Outcome Measures .....	203
<b>Constructing an Episode of Care: A Hip Fracture Example</b> .....	204
Research Question .....	204

---

E. Hellsten (✉)  
Health Quality Ontario, Toronto, ON, Canada  
e-mail: [erik.hellsten@hqontario.ca](mailto:erik.hellsten@hqontario.ca)

K. J. Sheehan  
School of Population and Public Health, The University of  
British Columbia, Vancouver, BC, Canada

Data Source: Canadian Institute for Health Information Discharge Abstract Database .....	204
Defining the Index Event .....	204
Defining the Endpoint .....	205
Scope of the Services Included .....	205
Data Model .....	207
Use of the Data .....	208
<b>Constructing an Episode of Care: A Cardiac Example</b> .....	208
Research Question .....	208
Data Sources .....	208
Capturing Events by Linking Data Sources .....	208
Linkage of Cardiac Registry, Hospital Separations, and Death Files .....	209
Use of the Data .....	209
<b>Expanding on and Applying Episodes of Care: Further Considerations</b> .....	211
Building Episode-Based Case Mix Classification Systems .....	211
Risk Adjustment and Severity Classification .....	211
Attributing Episodes to Providers .....	212
Policy Applications .....	212
<b>References</b> .....	214

## Abstract

Health-care utilization data are traditionally presented in discrete, itemized formats that offer a fragmented view of the total picture of services delivered to treat an individual patient's health condition. In response, health services researchers have struggled for 150 years to define more meaningful units of analysis from the outputs of health-care services that are suitable for investigation. Beginning with Florence Nightingale in 1863, the basis for and application of an alternate conceptual approach – the *episode of care* construct – for organizing health-care events into a clinically meaningful unit of analysis has evolved. In recent decades this approach has been operationalized to support a variety of health services research and policy applications. To construct an episode, researchers must define three key elements including the index event, the scope of services included, and the endpoint. How these elements are defined is dependent on the objective of the episode construction and the data that are available. Here, the history of the episode of care concept, the core elements of an episode, and the researcher's key considerations and decision points in determining appropriate

parameters for these elements are described. Episode-based case mix classification systems, risk adjustment, and attribution rules are also examined. Lastly, two examples of episode of care construction and policy applications are discussed.

## Introduction

Health services researchers routinely face the task of organizing and making sense out of data on health service utilization in order to tell the story behind it. Crucially, the types of health-care data that researchers typically work with are more often than not reported and presented in ways that obscure or fragment the underlying medical narrative they represent. Health services researchers often rely on data points collected for administrative purposes, representing discrete units of service such as physician claims for individual services provided, discharge abstracts from hospitalizations, or records for drug prescriptions filled. While these individual observations are undoubtedly important both as individual health-care events and in aggregated form – for example, a researcher may be interested in the total annual number of hospitalizations for heart failure in a

particular hospital and how this sum compares to that in previous years – presenting health-care utilization data in this discrete, itemized fashion typically captures only fragments of the total picture of services delivered to treat a patient’s health condition.

The challenges of organizing health-care data into a coherent narrative stem in part from the unique nature of the health-care “product”: unlike most other commodities, health care is often delivered through a series of separate but related encounters, rather than through a single stand-alone service (Feldstein 1966; Hornbrook et al. 1985). A patient presenting with a health condition may receive health-care services that span multiple different health-care providers over several points in time. The interrelated nature of this variety of providers and services in providing care for a health condition for an individual patient is typically not readily apparent in standard itemized or index-based presentations of health-care data.

Figure 1 provides an illustrative example of a series of individual health-care service data points, which on closer inspection are revealed to be a single patient’s journey through treatment with a total knee replacement for osteoarthritis of the knee. Beginning with a consultation with a primary care physician for chronic knee pain that has failed to respond to conservative treatment, the patient is referred for a radiograph several days later and booked for a consultation with an orthopedic surgeon in their office 4 weeks following. During this consultation, the patient and surgeon decide on a total knee replacement surgery, which is scheduled at a local hospital approximately 2 months after the consultation. Several days

before the patient enters hospital, they are assessed at a preoperative clinic to prepare for the surgery. The patient is then admitted to hospital, receives a total knee replacement on the day of admission, and is discharged home 3 days later without incident. Following their discharge home, the patient receives three weekly visits from a physiotherapist contracting with a local home care agency to assist with their rehabilitation. Three weeks later, the patient has a follow-up visit with the surgeon in their office to assess their recovery. Satisfied with the patient’s progress, the surgeon decides no further follow-up is needed; the patient’s care journey can now be considered to be at an end.

This complex series of encounters typifies a routine, simplified pathway for a patient receiving a successful total knee replacement. In some instances, the same patient’s journey might well be further complicated by additional health-care events, such as the appearance of in-hospital or postoperative complications, the need for readmission to hospital or revision surgery, and other potential sequelae.

For the health services researcher, the hypothetical knee replacement example likely produces over a dozen data points in the form of a series of individual encounters recorded between several health service providers and provider organizations over a span of several months. In many cases, this encounter data will also be housed across several discrete – and frequently disconnected – datasets: primary care physician and specialist billings, inpatient hospital discharge abstracts, home care agency records, and so on. The health service researcher faces the challenge of stitching these discrete observations

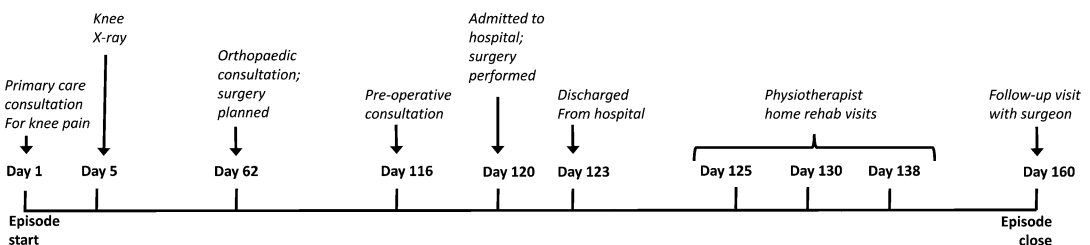


Fig. 1 Example episode of care for osteoarthritis of the knee. Illustrative example; timeline not to scale

together to form a meaningful and comprehensive picture of a patient's knee replacement journey through the health-care system.

This chapter explores the conceptual basis for and application of an alternate conceptual approach for organizing health-care events in a clinically meaningful unit of analysis known as the *episode of care*. Nearly half a century ago, health economist Jerry Solon published the seminal paper "Delineating episodes of medical care," which put forward the following definition of the concept:

An episode of medical care is a block of one or more medical services received by an individual during a period of relatively continuous contact with one or more providers of service, in relation to a particular medical problem or situation. (Solon et al. 1967)

This intuitive definition, while later nuanced and expanded upon by other researchers, still provides the basic foundation for the application of this concept today.

This chapter begins with a brief history of the evolution of the episode of care construct, from its conceptual origins to its operationalization in health services research applications to its use in modern policy applications. The core elements of an episode are described and the researcher's key considerations and decision points in determining appropriate parameters for these elements in order to define a meaningful episode of care. Case mix classification systems, risk adjustment, and severity classifications are also examined. Lastly, examples of recent research and policy applications using episodes of care are discussed.

---

### **Health-Care Data and Defining the Unit of Analysis: Historical Perspective**

For over 150 years, health services researchers have struggled to define meaningful units of analysis suitable for investigation from the outputs of health-care services. In the nineteenth century, Florence Nightingale produced what may have been the first outcome-oriented classification system for hospital care, labeling patients leaving

hospital as either "relieved," "unrelieved," or "dead" (Nightingale 1863). Following Nightingale, the Boston surgeon Ernest Amory Codman published his lecture "The Product of a Hospital" in 1914 (Codman 1914), which set out an early framework for classifying the outputs of hospitals such as counts of patients treated, beds, bed days, and student hours.

Throughout most of the twentieth century – and indeed, still largely today – health-care utilization data continued to be aggregated on the basis of Codman-esque sum totals or indices of individual services and outputs. Coinciding with the emergence of health services research in its modern form in the late 1950s and early 1960s, researchers began to note the inadequacy of routinely collected health-care data for the purposes of understanding the nature of health-care utilization. Whether presented as physician visits, bed days, or entire hospital stays, these isolated encounters were often insufficient in themselves for understanding the nature of a patient's encounters with health-care providers for treatment and the course of medical services delivered. In their 1960 paper "Delineating Patterns of Medical Care," Jerry Solon and colleagues noted the need to learn about the "patterns" of health-care utilization, "rather than merely documenting isolated incidents of use." Solon et al. proposed an approach for "consolidating detailed information on use of medical resources into meaningful, integrated forms" and translating "the vast spectrum of utilization into representative patterns" (Solon et al. 1960). Rather than presenting a "procession of chaotic data" from isolated medical encounters, this consolidation would enable the systematic organization of health-care data to better inform research and policy on the utilization of medical services (Solon et al. 1960).

In their classic 1961 paper "The Ecology of Medical Care," Kerr White and colleagues similarly encapsulated this issue (White et al. 1961). They identified the patient as the primary unit of observation, rather than the disease, visit, or admission (White et al. 1961). By following a patient's progression through all their encounters with health-care providers for treatment of a particular health condition, the course of medical

services delivered may be captured. White et al. suggested “the natural history of the patient’s medical care” may be the most relevant primary unit of observation and proposed some approaches for disaggregating the data found in traditional health-care indexes into more meaningful forms, such as employing time windows of weeks or months rather than years, and better understanding the decision-making process that unfolds between patients and medical care practitioners over the course of a particular illness. The paper is perhaps also the first to describe the “episode of ill health or injury” as its unit of observation (White et al. 1961).

In his 1966 article “Research on the Demand for Health Services,” Paul Feldstein extended Codman’s original work to define the “product” of health care, noting that in order to define a meaningful unit of output for analysis, researchers required “a better understanding of how the various components of care are used in its production” (Feldstein 1966). Feldstein emphasized the importance of comprehensively accounting for the entire combination of service inputs – such as hospital care and physician visits – used to treat a particular illness and considering differences in the relative contributions of these services in the production of treatment products between groups of providers and over time. He noted the limitations of conventional aggregate indices conventionally applied to quantify national medical production in terms of outputs such as numbers of visits or bed days.

---

### **The Episode of Care: A Unifying Concept**

Following this foundational papers’ assessment of the gaps in contemporary methods for analyzing health-care utilization data, 1967 saw the publication of three seminal health services research papers that each put forward a different perspective on establishing an operational definition for White et al.’s “natural history of the patient’s medical care.” In their series “The Development of Standards for the Audit and Planning of Medical Care,” Isidore Falk and colleagues took a

clinical practice perspective to the issue, defining a unit of analysis suitable for the development of “standards for the content of good clinical performance” in particular diseases, against which providers’ medical practices could be evaluated “from preventive to postclinical after-care” (Falk et al. 1967). Within these units, which they presciently termed “pathways” in a subsequent paper in the same series (Schonfeld et al. 1968), the authors consulted expert physicians to arrive at quantitative judgments on what constituted appropriate medical utilization, such as the average time required for a first diagnostic visit or the average hospital length of stay for various diseases.

Published the same year, “Changes in the Costs of Treatment of Selected Illnesses, 1951–1965” by Anne Scitovsky (1967) extended earlier work developing an alternate approach to address the inadequacies of the Bureau of Labor Statistics’ medical care price index – which was based on the prices of individual medical items and offered a limited and fragmented view of changes in medical spending – to introduce a “cost-per-episode-of-illness” approach that enabled the construction of a medical care price index based on the average costs of treatment of selected illnesses rather than the costs of discrete items. By demarcating patient episodes of illness within a claims dataset that included all relevant services delivered between an initial diagnosis or presentation for a health issue and either a service-defined endpoint (e.g., the last chemotherapy treatment following breast cancer treatment) or a prescribed follow-up time period that varied by disease, Scitovsky was able to compare changes in service utilization and cost for particular diseases between two time periods. The episode unit enabled Scitovsky to both comprehensively capture the full range of services delivered to treat a specified disease and examine changes in the provision of care, such as a reduction in the rate of home visits and the shift of forearm fracture repairs from office-based general practice to hospital-based specialty care.

While White et al. provided a clinical practice construct of the episode of care and anticipated the use of clinical pathways, and Scitovsky made

operational use of the concept for analyzing and comparing costs and utilization (an application that continues to see widespread use today), it was Jerry Solon and colleagues who provided the first comprehensive definition of this new concept in “Delineating Episodes of Medical Care” (Solon et al. 1967). The authors described three essential features found in any medical care episode: a beginning point, a course of services directed toward an objective, and a point of termination or suspension of the service. Episodes could be constructed around a variety of issues, including a general health-related complaint, a set of defined symptoms, a diagnosed disease, or the achievement of a particular health objective (such as preventive care) where no active morbidities are presented.

Solon et al. touched on range of important (and still relevant) methodological issues such as the definition of clinically meaningful time intervals for different medical conditions between service encounters to mark the end of a previous episode and the beginning of a new one. They discussed the conceptual challenges posed by chronic conditions that require ongoing medical management without a definite closure and expounded on the relationships between health services contained within a single episode, such as a chain of related physician visits. They identified potential interactions between multiple related episodes within the same individual, such as periodic exacerbations, remissions or acute sequelae linked to an underlying chronic condition, concurrent episodes for comorbid conditions, or iatrogenic events resulting from the treatment delivered for an initial health problem. They suggested that concurrent conditions in a patient might be treated as either part of a single episode or multiple distinct episodes, depending on whether the physician chooses to focus on one illness at a time or treat several within the same encounter (Solon et al. 1967).

Solon et al. distinguished between *episodes of care*, which are defined based on reported health services, and *episodes of illness*, which may occur without the provision of health services. While the episode of illness is an important concept for understanding the etiology of sickness and disease

apart from medical care, practically speaking, researchers typically face significant challenges in gathering precise data on episodes of illness that occur without corresponding provision of health services as these typically must be identified based on patient recollection. In their broadest definition, the episode of care may overlap with the episode of illness by including diagnostic follow-up after the point where medical care ceases, in order to understand the effect on a patient’s trajectory of illness (Solon et al. 1967).

Solon et al. also sketched out some potential applications of the episode concept in their 1967 paper, including using episodes as an organizing structure for clinicians planning a patient’s care and as a frame of reference for the development of standards of care for different medical conditions. They further applied the concept in their 1969 study “Episodes of Medical Care: Nursing Students’ Use of Medical Services,” analyzing and comparing the details of several years of health services received by nursing students and comparing episode-based utilization measures such as the volume and distribution of visits, diagnostic tests, and admissions within each episode (Solon et al. 1969).

After Solon’s codification of the essential elements of the episode of care, further refinements to and applications of the concept followed. In 1977, Moscovice first implemented episodes of care using computerized routines, constructing disease-specific algorithms to define episodes for several tracer conditions based on patient visit information (Moscovice 1977). The algorithms identified an initial encounter with the recorded incidence of a specified diagnosis code (the index event) and then tracked subsequent encounters by the same patient with reported codes for the same diagnosis or specified related comorbidities. For each condition, based on physician input, a maximum time interval was defined between service encounters to assign services to either part of an existing episode or as the start of a new episode. Services and resources expended for each health condition were similarly defined based on information contained in medical directives and through clinician input. Moscovice compared measures of utilization between providers and



treatment sites, including volumes of visits, laboratory procedures, prescription patterns, and total relative charges. Moscovice's computerized approach – using clinician input to define meaningful condition-specific parameters – is much the same as that used by modern episode grouping software algorithms today.

In 1985, Mark Hornbrook and colleagues published perhaps the most comprehensive paper on the subject, the widely cited “Health Care Episodes: Definition, Measurement and Use” (Hornbrook et al. 1985). Expanding upon Solon et al.'s original definitions, their paper distinguished between episodes of illness (a symptom, sign, or health complaint experienced by the patient), episodes of disease (the morbidity or pathology as viewed from the provider's perspective), and episodes of care (“a series of temporally contiguous health-care services related to treatment of a given spell of illness or provided in response to a specific request by the patient or other relevant entity”). They further differentiated episodes of care from health maintenance episodes, which are health-care services delivered with the goal of enhancing wellness, preventing disease, cosmetic, or contraceptive purposes, rather than toward the resolution of an existing pathology. Finally, Hornbrook et al. suggested that episodes of care may be delivered for the treatment of more than one episode of disease or illness concurrently.

Subsequent research on the episode concept has largely expanded on these earlier efforts and made incremental refinements in areas such as methods for risk adjustment and complexity stratification within episodes, methods for estimating episode costs at the system and provider levels, rules for attributing episodes to health-care providers, and the development of episode-based case mix classification systems which establish rules for comprehensively assigning all reported health-care services to mutually exclusive episodes. With these methodological advancements have come an impressively diverse array of applications of the concept, operationalizing episodes for use in a variety of research purposes, utilization review, provider profiling, and provider payment model design.

## **Episodes as an Analytical Tool: Advantages**

As a unit of observation, the episode of care offers several advantages for the health services researcher over other commonly used methods:

### **Flexibility**

Episodes do not have preset boundaries based on historical – and often arbitrary – observation units used in health-care administrative claims data such as hospitalizations or physician visits. The flexibility of the episode model allows for parameters such as the index event, endpoint, and types of services included to be customized based on the objectives of the study and the nature of the health conditions examined.

### **Comprehensiveness**

Episodes support the inclusion of all relevant health-care services for a particular condition or procedure, which may be delivered across multiple care settings, numerous individual providers, and overextended time frames. This broad, inclusive framework enables the researcher to present an integrated, comprehensive picture of the health-care services delivered to treat a specific issue, with the ability to cross historical silos existing between health-care providers, care settings, and subsystems. This also makes them an attractive analytical vehicle for policies aimed at promoting integration and coordination of care between providers and over time, such as payment models and performance reporting initiatives.

### **Clinical Meaningfulness**

Because the episode design parameters can be customized to the nature of a particular disease or procedure, they support the design of a more clinically meaningful unit of analysis than traditional service counts or indices. Episodes allow for the analysis and comparison of specific health

issues or treatments between providers, settings, or points in time, using a format that is both more clinically homogenous and more reflective of the underlying clinical reality of the health problem studied. The parameters of episodes used to develop analyses can also be set with the explicit input of clinicians who have understanding and expertise in the particular condition or intervention of focus, strengthening the credibility of the analysis.

---

## **Episodes as an Analytical Tool: Challenges**

There are also a number of important challenges associated with designing, implementing, and interpreting episode-based approaches:

### **Data Requirements**

Developing episodes requires the availability of several essential data elements, most notably the use of unique identifiers spanning multiple health-care encounters involving the same patient. Unique health service identifiers are still not commonly available in some health-care jurisdictions or datasets; in such cases, probabilistic matching algorithms may be used as a potential substitute to link service encounters involving the same patient. The researcher will typically have to merge multiple health-care datasets using unique identifiers in order to develop a comprehensive episode-based analysis and subsequently develop algorithms for defining episodes based on diagnoses, services, and calendar dates.

### **Complexity**

The episode of care is a multidimensional concept that can present challenges for defining appropriately, particularly if the health issue under study is in itself complex or heterogeneous. In order to develop meaningful parameters for the episode, the researcher is advised to either seek clinician input or draw from previously published literature

outlining such parameters. Finally, researchers may encounter difficulties in communicating around episode-based analysis to others who may not be familiar with the concept.

### **Time and Resources Required**

The increased complexity of the episode approach over traditional silo-based forms of analyses leads to increased time and resources required for tasks such as defining episodes, preparing datasets, and troubleshooting analyses. Many episode-based analyses also require substantial computing power to run.

### **Methodological Challenges**

Episode-based approaches also bring more complex methodological issues that need to be addressed, such as developing methods that measure variables across time, providers, and settings; attributing services within overlapping episodes; and adjusting for patient case mix and dealing with outliers. Some of these issues and their implications, as well as potential solutions, are described in the following sections.

Notwithstanding these challenges, when an episode-based analytic approach is well aligned with the intended research questions or analysis objectives and applied carefully and thoughtfully, it can be a powerful tool for both research and policy applications.

Although the episode of care concept is far from new, it has experienced a surge in popularity in recent years due to its growing use in high-profile applications such as provider payment policies and profiling efforts. Traditionally, while regarded as conceptually attractive, episode-based approaches were often difficult to implement in practice outside of research efforts. The increasing availability of linked health-care services datasets suitable for constructing episodes, ready-made episode grouping software, and advances in computing power has enabled episode-based approaches to become a viable option in a growing range of applications.

## Constructing an Episode of Care: Key Components

Health services researchers seeking to construct analyses using episode-based approaches should familiarize themselves with several basic requirements in terms of the nature of the data required and the essential elements for defining the episode. The growing body of literature around this approach also provides insight into a variety of methodological challenges and considerations that are frequently encountered in developing episodes of care.

### Data Sources Required

From an operational perspective, a researcher seeking to construct an episode of care requires data on individual health service encounters that contains several core elements necessary for defining the key parameters for an episode of care.

### Individual-Level Record Linkage

The temporal nature of the episode and its organization of related health-care events around an individual's health issue requires that data for analysis contain an identifier at the individual level that can be linked across records and over time. Typically, data elements from health-care datasets (such as hospital discharges, physician billings, and home care services) are merged and linked using either a unique patient identifier, probabilistic matching algorithms that match on some combination of variables (e.g., age, place of residence, and time of the encounter), or a combination of the two approaches. Health-care datasets with individual-level record linkage are made available through government sources such as the Centers for Medicare and Medicaid Services or research institutions such as the Institute for Clinical Evaluative Sciences in Ontario.

### Information on Type of Service

The type of service delivered (e.g., the type of physician procedure, hospital inpatient admission,

a drug prescription filled). Sometimes, a single record may contain multiple instances of this, such as a hospital inpatient stay with multiple procedures performed.

### Diagnosis Information

The patient's diagnosis. This may take the form of either a Principal or Most Responsible Diagnosis (diagnosis responsible for the majority of care provided) or a Primary (preadmission), Secondary (comorbidity), or Complication (postadmission) diagnosis.

### The Date/Time of the Service Delivered

This element is crucial in order to be able to assign encounters around a particular period of time or to arrange encounters in a medically logical order (e.g., initial diagnosis followed by disease staging followed by surgery followed by follow-up assessments). For hospitalization episodes, this may include an admission date, discharge date, and sometimes the date of procedures performed within the hospitalization.

### Core Elements of the Episode

As first described by Solon et al., every episode of care has a set of three core elements that must be defined in order to set the parameters for analysis: **the index event and/or starting point for the episode**, **the episode endpoint**, and **the scope of services included** (Solon et al. 1967). In parallel with the definition of these elements, the researcher must select the **outcome measures** of interest to be examined using the episode construct.

In defining these core elements, the researcher is advised to consider the research or policy applications of the analysis and to solicit clinician input on these definitions. One of the most attractive features of the episode of care approach is its resonance as a meaningful measurement unit for clinicians.

## Defining the Index Event and/or Starting Point

An episode of care requires an index encounter or event that triggers the start of the episode. This index event may be a specific health-care service (such as a knee replacement procedure), a particular diagnosis or health condition (such as a diabetes diagnosis, assigned by any provider in any setting), or some combination of these two, such as an admission to hospital for treatment of a congestive heart failure exacerbation. In most cases, the index event will also mark the start of the episode. Exceptions to this rule include examples where the episode definition employs a “look-back” period from the index event, as in the case where the incidence of a surgical hospitalization triggers an episode window preceding the admission with a defined period of presurgical care.

An index event may also take the form of a point in time rather than a particular health-care encounter. Some episode methodologies use this approach for defining episodes for chronic diseases such as diabetes or chronic obstructive pulmonary disease. These are expected to be lifelong conditions, but for ease of analysis, they may be annualized into year-long episodes. A point in time might also be used as the index event and starting point in the case of an incomplete episode, where the analyst’s dataset begins at a particular date in time and there may be encounters related to a particular episode falling before the start of the dataset, rendering it impossible to establish a definite starting point.

Some episode approaches may also employ index events that “shift” an existing episode category into a different category when they occur. For example, an ICU admission by a patient registered with a COPD episode might shift the patient’s episode into a higher severity level; a patient with an ongoing coronary artery disease episode that experiences a heart attack might be shifted into an acute myocardial infarction episode or a coronary artery bypass graft episode if they seek surgical treatment. Similarly, the occurrence of a complication such as a pressure ulcer during hospitalization for treatment of hip fracture

might trigger a separate, concurrent episode for the complication.

### Examples

Scitovsky (1967), Moscovice (1977), and others used an index event for their episodes defined by the first recorded instance of prespecified diagnoses (Moscovice 1977; Scitovsky 1967).

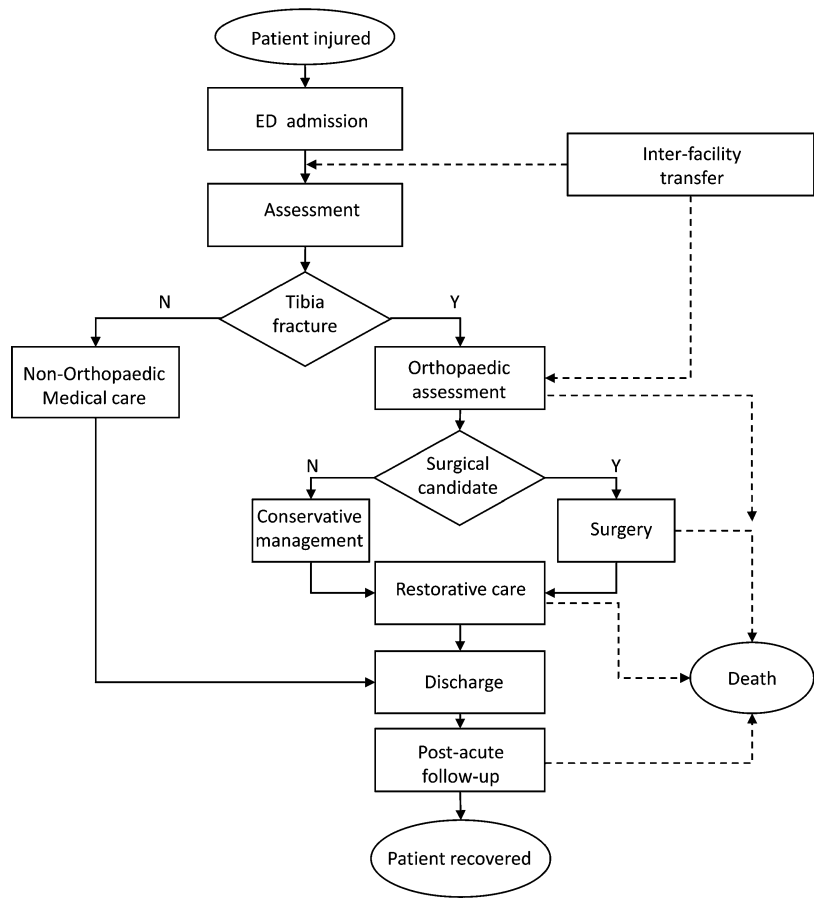
The American Board of Medical Specialties Research and Education Foundation defined the index event for the episode of care for colonoscopy as the provision of a colonoscopy procedure, but defined the episode to also include services provided in the 3 days preceding the colonoscopy (High Value Health Care Project 2011).

## Defining the Endpoint

Each episode has an event, time window (either a fixed time window from the episode index event or a window of time where any related services are absent), or other decision rule triggering the conclusion of the episode. Researchers may select a clinically logical event for concluding an episode such as a specific health-care event. This is more common in cases of elective and trauma procedures where a defined sequence of health-care events is expected to take place. For example, in Fig. 2, the patient arrives at the emergency department and is assessed for the presence of a tibia fracture (physician exam, diagnostic testing). The fracture is confirmed or refuted. If confirmed the patient is referred for orthopedic assessment and identified as a surgical or nonsurgical candidate. The patient receives surgery or conservative management, followed by restorative care before discharge. The patient then receives follow-up in the community. Health-care events which may end the episode include a patient’s death, discharge from hospital, or a follow-up appointment after surgery.

However, a clinically logical event is not always available. The original definition of the episode of care put forward by Scitovsky (1967) and Solon et al. (1967)’s suggested that episodes for a particular health issue concluded with the discontinuation of services for that health issue

**Fig. 2** Episode of care for patient arriving to an emergency department with a suspected fracture of the tibia



(Scitovsky 1967; Solon et al. 1967). Often described as a “clean period,” this generally takes the form of a specified window of time where no services related to the episode are provided. For example, in the case of chronic bronchitis, this might be 45 days without any services related to bronchitis treatment such as x-rays or relevant medication. Theoretically, using these definition episodes for a particular condition can have any duration, so long as relevant services continue to be provided for treatment of the condition. As with the duration of a fixed time window, the duration of a clean period should be condition or procedure specific and defined based on clinical input. Typically, episodes for acute conditions such as appendicitis – where a defined, time-limited course of treatment can be expected – will have shorter clean periods than episodes for chronic diseases or acute conditions

with chronic sequelae like stroke, where follow-on care can sometimes last for years. It should be noted that with endpoints based on clean periods, the same considerations apply in terms of “open” episodes: active episodes where a dataset or claims history is censored before the full duration of the clean period elapses are considered “open” at that point.

Alternatively, an endpoint can be a fixed point in time, such as 30 days following a hospital admission or discharge. These sorts of calendar-based episode endpoints are commonly used for outcome measures that seek to compare “apples to apples” across providers that might have different discharge practices. The current public reporting principles adopted by the Centers for Medicare and Medicaid Services to report on hospital mortality, readmission, and other outcomes stipulate the use of a standardized time period to facilitate

comparison. A point in time approach may also be adopted in the case of chronic disease episodes based around an annualized analysis period or where a dataset is censored at a particular date and truncates “open” episodes. Using migraine episodes, Schulman et al. (1999) put forward a novel approach to empirically defining the length of an episode of care (Schulman et al. 1999). The study used administrative claims data to determine the point in time following the index event where elevated weekly charges returned to their original pre-episode levels.

Finally, the start of a new episode may trigger the close of an existing one. For example, a patient suffering from osteoarthritis of the knees who receives a total knee replacement may have an ongoing osteoarthritis management episode replaced with a total knee replacement procedural episode. Following the surgery, should their osteoarthritis be completely addressed, the patient would not be expected to continue the original disease episode.

### Examples of Endpoints

Moscovice drew on published medical directives and clinician expert opinion concerning “reasonable periods of follow-up” to a time period for each condition where the absence of services related to the condition would mark the beginning of a new episode (Moscovice 1977). Scitovsky used a similar condition-specific approach to defining episode duration (Scitovsky 1967).

Health Quality Ontario used input from clinical expert panels, informed by analysis of linked administrative data on utilization, to define the typical duration of services provided in episodes of hip fracture care (Health Quality Ontario 2013).

Symmetry’s Episode Treatment Groups use the approach of “annualizing” the episode of care for chronic diseases with indefinite durations (Optum 2015).

## Selecting the Scope of Services Included

Episodes of care can be as comprehensive or as specific in their inclusion of services as a

researcher desires and as is feasible given available data. The scope of services included requires a decision on the part of the researcher: a more holistic episode approach might capture all services provided during the episode window, regardless of whether they appear to be directly related to a condition. This approach is being employed by the Centers for Medicare and Medicaid Services’ Bundled Payments for Care Improvement initiative (Centers for Medicare & Medicaid Services 2014). A more limited episode may include only those services directly related to a particular condition. For example, in defining services to be included in episodes of diabetes care, the Netherlands’ bundled payment initiative has included only community-based professional services, excluding drugs and hospitalizations (Struijs et al. 2012a).

Ultimately, the scope of services included in the episode depends on the objectives of the analysis and its intended applications and the nature of the data available. Payment applications, for example, may suggest the utility of a single episode payment that covers multiple different types of services over a fixed period of time, in order to prevent any risk of “double counting” payment (Struijs et al. 2012b). A truly comprehensive episode might even include services beyond those delivered by health-care providers: for an episode of care around complex patients with functional needs, it may be ideal to also include social care services delivered – to the extent that they are captured in databases.

If the researcher elects to use a more clinically focused approach or a categorically based approach to service inclusion, clinical input is imperative. Input from clinical panels is required to identify the services that are related to the episode of care and the types of services that would likely not be related.

### Examples

Moscovice used published medical directives and clinical input to define lists of medical services that could “realistically be used in the treatment of a particular problem or related comorbidity.” In the case of otitis media, this list of services included lab tests such as throat cultures that

might be used to rule out plausible related comorbidities. Based on these lists, Moscovice defined a set of “patterns of care” based on the most common combinations of services delivered to treat each episode. For otitis media, 20.6 % of episodes analyzed consisted of a single visit, while 13.8 % consisted of an initial visit, administration of an antibiotic, throat culture, and then a follow-up visit (Moscovice 1977).

Solon et al. examined nursing students’ utilization of health-care services within episodes of care through separating encounters into “universal” visits – those services, such as vaccinations, provided to all students – and “individual” visits specific to treating the nursing student’s episode (Solon et al. 1969).

## Outcome Measures

Ultimately, the episode of care is intended to serve as a clinically relevant unit of analysis for measuring particular aspects of care or outcomes delivered. In the broadest sense, any outcome measured at a standard time frame (e.g., 30-day mortality) might be considered an application of the episode-based approach. However, most episode-based studies have focused largely on process- or utilization-related measures. Following Falk et al.’s concept of the episode or pathway as a unit of analysis for auditing quality of care (Falk et al. 1967), Lohr and Brook (1980) compared providers’ use of appropriate therapy for respiratory infection, while Nutting et al. (1981) used episodes of care to compare health systems’ performance in terms of preventative services, timely diagnoses, continuity of care, and other factors.

By far the most common use of episodes since their earliest uses has been for examination and comparison of health-care costs and utilization: studies by Scitovsky (1967) and Solon et al. (1969) examined measures of total episode costs and number of visits by different health professionals, respectively. A popular use of episode-based cost measures involves the comparison of different physicians or physicians’ practices in terms of the total downstream health-care costs of their patients – a practice known as profiling

(Cave 1995). More recent studies have used episodes for similar cost and utilization profiling approaches with hospitals as the central unit of analysis (Birkmeyer et al. 2010), as well as exploring regional comparisons (Reschovsky et al. 2014). Regardless of the unit of analysis for comparison, the episode construct enables an “apples to apples” mechanisms that allows for comparison of the total treatment “product” between different providers or regions.

The vast majority of episode-based costing analyses have largely been conducted in the United States, where the predominant use of itemized claims data for reimbursing health-care services naturally lends itself to the aggregation of such claims into episodes of care. In countries such as Canada or some European nations that make greater use of global budgets for funding health-care services, constructing episode of care investigations of health-care costs requires the development of methodological approaches that serve as surrogates for “pricing.” In Ontario, such approaches have been developed using a combination of case mix cost estimation methodologies for globally budgeted hospital sectors and claims schedules for physicians and other fee-for-service providers (Sutherland et al. 2012).

---

## Examples

Sutherland et al. compared the total costs (including hospital, physician and inpatient, and community-based rehabilitation) of hip and knee replacement episodes between regions in Ontario, correlating higher costs with the use of less efficient care settings (Sutherland et al. 2012).

After defining the most common combinations of services (or “patterns of care”) used for each type of episode, Moscovice evaluated the proportion of episodes delivered according to these patterns and compared the results between different care providers and settings (Moscovice 1977).

Scitovsky used episode-based measures of total health-care costs per treated condition to assess differences in costs (and the changes in service mix driving these differences) for episodes of care over time (Scitovsky 1967).

Lohr and Brook (1980) used an episode-based analysis to compare quality of care for respiratory conditions before and after the publication of guidelines on the use of injectable antibiotics, defined as the percentage of episodes that included appropriate use of antibiotic therapy (Lohr and Brook 1980).

---

## Constructing an Episode of Care: A Hip Fracture Example

### Research Question

As described earlier, the episode of care can be as comprehensive or as specific in its inclusion of services as is desired and feasible. Here, an example is presented of a more focused episode construction to address the question of the effect of timing of hip fracture surgery on patient outcomes. Many argue that patients presenting to hospital with hip fracture should receive surgery as early as possible; however, the literature detailing the benefits of accelerated access to the procedure is inconclusive. Furthermore, little is known as to causes of delay: some patients wait to be medically stabilized, while others are delayed due to administrative factors such as hospital type, transfers, and date and time of admission.

The literature identifies the following pathways on the basis of treatment patients receive during acute hospitalization with hip fracture: surgical treatment (Menzies et al. 2010), nonsurgical treatment (Jain et al. 2003), or palliative care (Meier 2011). Most patients undergo surgical treatment during either their initial hospitalization or after transfers from hospitals where patients are initially admitted. While in the hospital, some patients are medically stabilized before surgery. Patients remain in the hospital after surgery until they are fit to be discharged home or to an alternative level of care. Some patients receive nonsurgical management of their hip fracture as their risk of complications and death is too high. These patients are medically stabilized and discharged home or to an alternative level of care. Palliative care is offered to patients at the end stage of a terminal illness.

## Data Source: Canadian Institute for Health Information Discharge Abstract Database

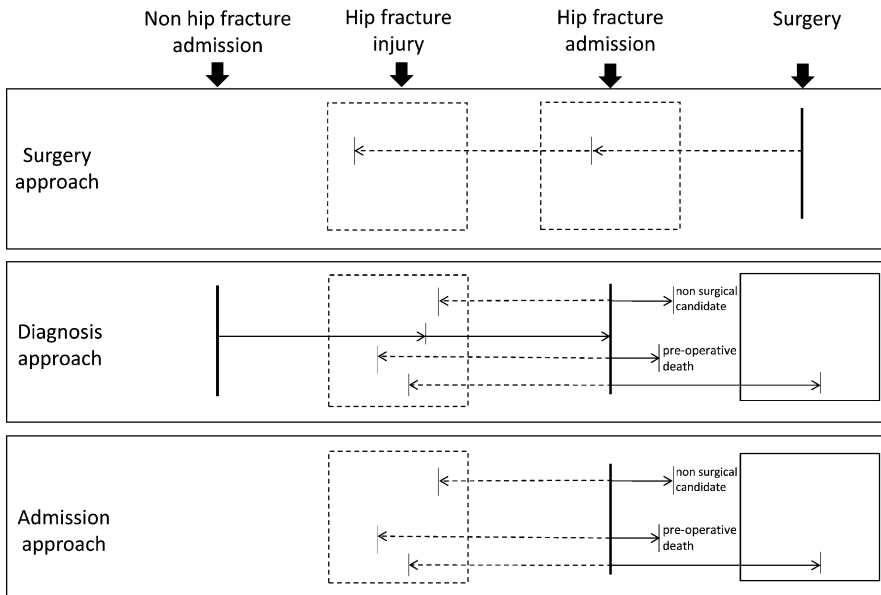
The Canadian Institute for Health Information (CIHI) is an independent, not-for-profit organization that provides information on Canada's health system and the health of Canadians (Canadian Institute for Health Information 2015). CIHI facilitates collection of standardized administrative, clinical, and demographic data from acute hospitalizations through the Discharge Abstract Database (DAD). The data (2003–2012) are presented as a series of flat comma-delimited files with multiple abstracts for some patients. To prepare data for analysis, researchers develop a relational database to facilitate combining abstracts into episodes of care. In the following sections, a conceptual framework for constructing an episode of hip fracture care and the approach for operationalizing it using the CIHI abstracts is described.

Here a method for constructing an episode of care to study the effects of timing of hip fracture surgery using acute care discharge abstracts is described, and therefore, the episode is confined to patients admitted to the hospital and outcomes occurring in-hospital. Data relating to emergency department wait times or post-acute care utilization was not provided.

### Defining the Index Event

The ideal index event is injury time. This event enables researchers to capture all hip fracture patients, includes events preceding hospital admission such as prehospital death, and captures the time from injury to admission which contributes to delays (Sheehan et al. 2015). However, injury time is not available through administrative databases and therefore alternative index events must be considered. When identifying the index event for the episode from administrative data, researchers may select the hip fracture surgery procedure, the hip fracture diagnosis, or admission with a diagnosis of hip fracture (Fig. 3). A procedure approach captures outcomes which occur postoperatively implying that time at risk





**Fig. 3** Approaches to defining the index event for a hip fracture episode of care. *Thick vertical lines* indicate the index event for constructing each care episode. *Dashed*

*boxes and arrows* represent events and their timings ascertained retrospectively. *Solid box and arrows* represent events and their timing ascertained prospectively

begins at the time of surgery. A diagnosis approach includes patients who incur a hip fracture in acute hospital following admission for another diagnosis. Here an admission approach is adopted as it allows researchers to capture outcomes which occur before surgery, including pre-operative death, while excluding patients who incur a hip fracture in the hospital after admission for another diagnosis (Sheehan et al. 2015).

### Defining the Endpoint

In this example, a clinically logical event defines the endpoint: death, discharge home, or discharge to an alternative level of care. A fixed point in time is also considered an endpoint as the dataset is censored at March 2012.

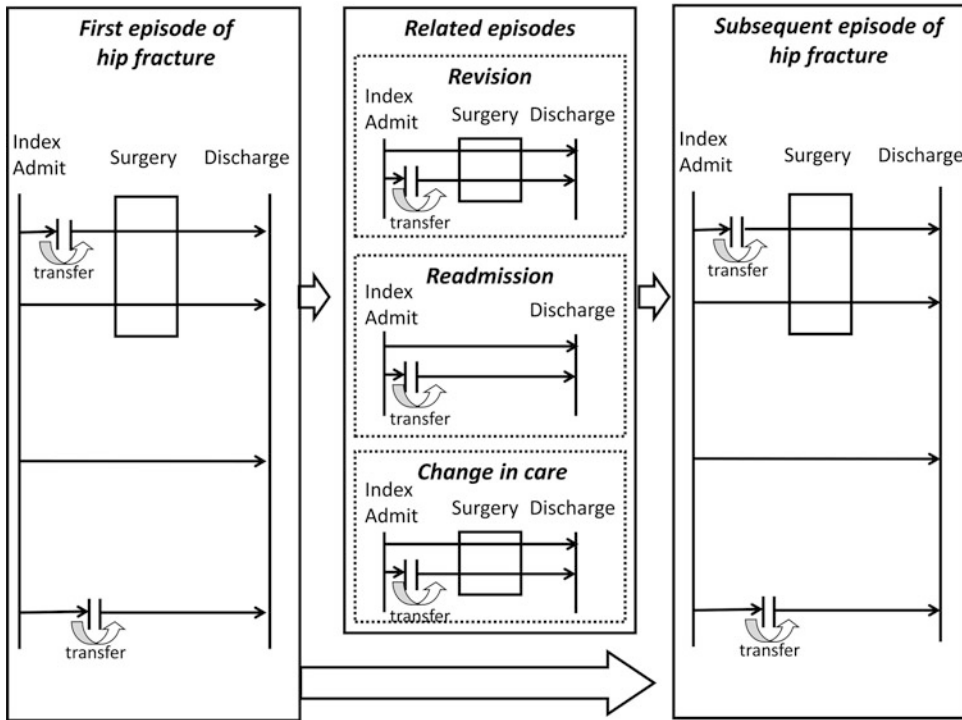
### Scope of the Services Included

In this example services included are specific to the effect of surgical timing on outcomes of acute hip fracture care. First, researchers define how time to surgery is measured. Where the index event is surgery, time to surgery is measured

from admission to surgery within a single discharge abstract. Where the index event is diagnosis, time to surgery is measured from diagnosis (preadmission or postadmission) to surgery. Where the index event is admission, time to surgery is measured from the earliest admission time to surgery time, preoperative death, or discharge without surgery. This approach is inclusive of transfers which occurred between admission and discharge, a potential administrative factor for delay (Fransoo et al. 2012).

Transfers from one acute care facility to another present in the data as a single patient with multiple records for hip fracture. Here, contiguous abstracts linked by transfers are combined in one episode; the earliest admission date and the latest discharge date are designated as the beginning and the end of episode (Fig. 4). To determine whether multiple records for a given patient reflect transfer before definitive care, the following rules are applied:

1. Less than 6 h between discharge on one abstract and admission on another abstract (12 h if at least one institution codes the transfer)



**Fig. 4** Conceptual framework for constructing hip fracture acute episodes of care. A patient is admitted to acute hospital for their first episode of hip fracture care. They may be transferred from one acute care facility to another before definitive care – surgery or conservative management. Once acute care is completed, they are discharged

home or to an alternate level of care. On completion of the first episode of care, a patient may return to acute care for a related episode – revision surgery, readmission, or for a change in care. Alternatively a patient may return to acute care with an entirely new subsequent hip fracture

2. Admission before 6:00 (12:00 if at least one institution codes the transfer), when discharge and admission occur on 1 day but discharge time is unknown
3. Discharge after 18:00 (12:00 if at least one institution codes the transfer), when discharge and admission occur on 1 day but admission time is unknown

surgery (Fig. 4). Finally, patients may present to the hospital with an entirely new subsequent hip fracture (Fig. 4). Following consultation with orthopedic surgeons, the following rules are created for patients with multiple discharge abstracts to identify related episodes as revision, readmission, change in care, or subsequent:

After discharge from acute care, some patients return to the hospital for an episode related to their hip fracture. They may return to acute care with a complication that requires revision surgery such as a failed fixation/prosthesis. Alternatively they may return to acute hospital for treatment of medical complications related to their hip fracture. Patients discharged without surgery may also return for surgery to alleviate pain or if they are no longer considered unfit for

- **Revision:** surgical admission within 90 days of discharge after initial surgical episode
- **Readmission:** nonsurgical admission within 90 days of discharge after initial surgical/nonsurgical episode
- **Change in care:** surgical admission within 30 days of admission for initial nonsurgical episode
- **Subsequent:** hip fracture admission more than 90 days after the initial episode

After the application of the rules, some adjacent abstracts remain unassigned because their admission and discharge dates are in reverse order. Only abstracts with the earlier admission date for constructing care episodes are used.

**Data Model**

For patients with a single discharge abstract, the abstract represents the first episode of hip fracture care. Multiple abstracts for a given patient could represent the first episode of hip fracture care, revision surgery, readmission, change in care, or a subsequent episode of hip fracture care. As such, the data fields from multiple discharge abstracts are used to construct new fields or update information in the same field but from a different abstract. A data model is developed to relate multiple abstracts of hip fracture care for a given patient, which explicitly defines how data fields relate to each other (Table 1). In particular, the data model establishes relationships among tables containing discharge abstracts of the first episode of hip fracture care, revision, readmission, change in care, and subsequent hip fracture episodes.

This involves creating a series of data tables and establishing relationships between them:

- *Episode of hip fracture care* table contains discharge abstracts of the first and subsequent episodes of hip fracture care, uniquely identified by patient id and hip fracture number. The episode may combine information from abstracts linked by transfers.
- *Revision surgery* table contains discharge abstracts of surgical hospitalization following first or subsequent episodes of hip fracture care.
- *Readmission* table contains discharge abstracts of nonsurgical hospitalization following first or subsequent episodes of hip fracture care whether surgical or medical.
- *Change in care* table contains discharge abstracts of surgical hospitalization following first or subsequent nonsurgical episodes of hip fracture care.
- Other tables contain demographic and comorbidity data.

Normalization is used to organize the CIHI discharge abstracts. First, repeating data fields with similar data in individual tables are eliminated, a separate table for each set of related data is created, and each set of related data is classified with two primary keys: patient id and hip fracture number. This normalization helps avoid multiple

**Table 1** Algorithm for identifying and classifying episodes of hip fracture care

Step 1	Remove duplicates from CIHI records
Step 2	For patients with single record, convert their records into episodes of initial hospitalization
Step 3	For patients with multiple records, combine records linked by transfers into care episodes:
	(a) Designate the earliest unlinked record as the start of a new episode
	(b) Combine contiguous records into an episode of care if transfer is identified
	(c) If records remain, go to 3a
Step 4	For each patient, classify the episode with earliest admission as initial hospitalization
Step 5	Classify episodes of surgical hospitalization with admission within 90 days of discharge from initial surgical hospitalization as revision
Step 6	Classify episodes with admission within 90 days of discharge from initial nonsurgical hospitalization as readmission
Step 7	Classify episodes of surgical hospitalization with admission within 30 days of admission from initial nonsurgical hospitalization as change in care
Step 8	For each patient, classify the episode with earliest admission beyond 90 days of discharge from initial surgical hospitalization as initial hospitalization with a new fracture
Step 9	Mark episodes with admission for open, pathological, and post-admit fracture
Step 10	Mark records not assigned to any episode as unassigned

fields storing similar data in one table. Second, separate tables for groups of data fields that apply to multiple abstracts are created, and these tables are related with a foreign key. This normalization maintains records that only depend on a table's primary key.

## Use of the Data

The dataset was created for estimating the frequency of preoperative deaths, postoperative complications, and in-hospital deaths following complications among patients exposed to various times before surgery. More specifically, the dataset creation enabled capturing events and durations associated with hip fracture care delivery. By operationalizing patient pathways in terms of data available from the CIHI, preoperative transfers, surgery, postoperative transfers, and outcomes of admission (preoperative death, postoperative complications, and death), as well as events following discharge (readmissions, revisions, subsequent hip fractures), were captured. From this dataset the durations of hospital stay, preoperative stay, and postoperative stay were estimated.

Patient and administrative factors for delay including demographic, clinical, and injury data fields and hospital type, date, and time of admission were also captured. These data facilitate the assessment of potential causes of delay. Combining discharge abstracts of all patients, whether they have surgical or nonsurgical treatment or die before surgery, facilitates assessment of the total harm from delays by considering deaths in those who did not make it to surgery.

---

## Constructing an Episode of Care: A Cardiac Example

### Research Question

A patient identified as in need of coronary artery bypass graft (CABG) while a hospital inpatient or as an outpatient is registered on a wait list for the elective procedure. A patient may encounter

different clinically logical events to define the endpoint: death, change in surgical candidacy, or the procedure itself. Sobolev and Kuramoto studied outcomes of surgical cardiac care according to time to surgery (Sobolev and Kuramoto 2007).

## Data Sources

Data on patients registered to undergo CABG are obtained from the British Columbia Cardiac Registry (BCCR) (Volk et al. 1997). This prospective database contains dates of registration on the list, procedure, and withdrawal from the list, along with disease severity and other risk factors, for all patients who are registered to undergo CABG in any of the four tertiary care hospitals that provide cardiac care to adult residents of British Columbia. Additional information on access to CABG is obtained from the BC Linked Health Database Hospital Separations File (Chamberlayne et al. 1998) and deaths from the provincial Death File (Sobolev et al. 2006).

## Capturing Events by Linking Data Sources

The care episode begins with a cardiac surgeons' assessment and includes hospital inpatients and outpatients registered on a wait list for elective CABG. A series of events take place preoperatively outside the hospital; preoperatively, perioperatively, and postoperatively in the hospital; and postoperatively outside the hospital. The care episode ends with death, change in surgical candidacy, or the procedure itself.

For patients registered on a wait list for elective CABG, a preoperative assessment, which may include additional tests, may occur prior to admission or in the hospital. Their surgical candidacy is then confirmed or refuted by an anesthesiologist. Once a patient is identified as a surgical candidate, their access to the procedure is determined through scheduling of operating room time. Patients are selected from hospital admissions and from the wait list on the basis of urgency, resource availability, and plan for discharge from

the hospital. The allocated time may change if emergent cases arise, if cancellations occur prior to the scheduled time, or if a patient's status changes during their wait. The patient is assessed again preoperatively, receives their surgery, is monitored postoperatively in the postanesthesia care unit, and is transferred to the ward or intensive care unit. The patient's postoperative recovery is managed in the hospital until they are suitable for discharge home or to an alternate level of care. On discharge the patient is followed up in the community until their recovery is complete or death occurs.

Patient-level records in administrative health databases may have multiple records for one patient. Patient records may be organized in two different formats – the “person-level” format or the “person-episode” format. The person-level format contains a single record per patient. In the current example, this approach would enable researchers to capture the time from inpatient registration on a wait list for elective CABG to the procedure, discharge, or transfer to an alternate level of care from a single hospitalization record. The person-episode format contains multiple records per patient. In the current example, this approach would enable researchers to capture the time from inpatient *or outpatient* registration on a wait list for elective CABG to the procedure, discharge, or transfer to an alternate level of care from multiple administrative records. As the present study aim is to determine the impact of waits on outcomes in cardiac care, all events contributing to the wait and potential outcomes of waiting should be captured. In order to achieve this, the person-episode approach is adopted whereby multiple data sources are linked.

The series of events during the care episode and patient characteristics are captured with administrative data entry. A data model which chronologically relates events captured by data elements is created. Events of interest include registration and removal from the wait list, hospital admission and discharge, scheduled surgery and unplanned emergency surgery, and preoperative, in-hospital, or follow-up death. Each event has an associated time stamp which allows researchers to sequence the events and to

determine the interval (wait time) between events. Once sequenced the person-episode is created which includes a de-identified patient number and an event number. This combination uniquely determines the patient-episode related to a specific event.

### **Linkage of Cardiac Registry, Hospital Separations, and Death Files**

A patients' Provincial Health Number is used to link BCCR records with the BC Linked Health Database Hospital Separations File and to the Death File. Events including hospital admission, comorbidities, surgery, hospital separation, and discharge type (home, alternate level of care, or death) are retrieved from the BC Linked Health Database Hospital Separations File. Deaths which do not occur in the hospital are captured by the Death File. Adopting a person-episode approach, the BCCR records are linked to the BC Linked Health Database Hospital Separations Files and the Death Files to create an analytical dataset. An analytical data dictionary is created to describe the variables created to represent events and patient characteristics (Table 2).

### **Use of the Data**

The dataset was created for estimating outcomes of registration for elective (nonemergency) procedures in surgical cardiac care. These outcomes included preoperative death, postoperative death, change in urgency status, and unplanned emergency surgery among patients exposed to various times before CABG. More specifically, the dataset creation enables capturing events and durations associated with registration on a wait list for CABG.

By operationalizing patient pathways in terms of the data available from the cardiac registry, hospital separations and death file preoperative events (delay to surgery, change in urgency status, unplanned emergency surgery, death) and postoperative death were captured. Furthermore, the durations of time spent on the wait list for elective

**Table 2** Analytical dataset data dictionary for records of patients awaiting elective coronary artery bypass grafting

Variable	Description	Source	Code
BCCR_ID	Patient identifier	BCCR	<Text>
AGECAT	Age decade	BCCR	1 – 20–29 years
			2 – 30–39 years
			...
			8–90 years
SEXF	Sex	BCCR	0 – man
			1 – woman
ANATOM	Coronary anatomy	BCCR	1 – left main disease
			2 – 2- or 3-vessel disease, with PLAD
			3 – 3-vessel disease, with no PLAD
			4 – 1-vessel disease, with PLAD
			5 – 1- or 2-vessel disease, no PLAD
			U – otherwise and unknown
UR_BR	Urgency at booking	BCCR	0 – emergency
			1 – urgent
			2 – semiurgent
			3 – nonurgent
			U – unknown
CM_CH	Comorbidities from Charlson index	Hospital separations	0, 1, 2, 3, or 4 (,4)
CM_BK	Major comorbidities	Hospital separations	1 – CHF or diabetes or COPD or rheumatism or cancer
			0 – other
INST_BK	Location at registration	BCCR	Hospital 1, 2, 3, or 4
WL_ST	Wait-list registration date	BCCR	mm/dd/yyyy
WL_EN	Wait-list removal date	BCCR	mm/dd/yyyy
WL_RM	Reason for removal	BCCR	0 – underwent surgery
			1 – death
			2 – medical treatment
			3 – at patient request
			4 – transfer to other hospital
			5 – otherwise removed from list
			6 – no surgical report
			7 – still on wait list
			8 – other surgery
			9 – death recorded in BCCR, not in Deaths File
DTHDATE	Death date	Death file	mm/dd/yyyy < . > – no date recorded
EXIT_CODE	Type of hospital discharge	Hospital separations	D – discharged alive
			S – left against medical advice
			X – died in the hospital
			N/A – not applicable
ADDDATE	Hospital admission date	Hospital separations	mm/dd/yyyy
			< . > – no date recorded
SEPDATE	Hospital separation date	Hospital separations	mm/dd/yyyy
			< . > – no date recorded

With kind permission from Springer Science + Business Media: Analysis of Waiting-Time Data in Health Services Research, Waiting-time data used in this book, volume 1, 2008, 21–22, Boris Sobolev and Lisa Kuramoto, Table 2.1

surgery were estimated by urgency status. These data enabled researchers to answer questions such as:

- What is the variation in time spent waiting for elective surgery?
- What is the effect of delays in scheduling an operation?
- Do longer delays contribute to preoperative mortality among patients with less urgent need for surgery?
- What is the survival benefit of cardiac surgery?
- What is the risk of death associated with delayed surgical treatment?

Combining data of all patients registered on the CABG wait list, whether they went on to receive surgery or not, facilitates assessment of the total harm from delays by considering change in urgency status and deaths in those who did not make it to surgery.

---

## Expanding on and Applying Episodes of Care: Further Considerations

### Building Episode-Based Case Mix Classification Systems

While most of the studies conducting episode-based analyses reviewed in this chapter focus on a limited set of conditions, episode grouping software such as the Symmetry Episode Treatment Groups (ETGs) (Optum 2015), Thomson Reuters Medical Episode Groups (MEGs) (MaCurdy et al. 2009), and the Centers for Medicare and Medicaid Services' episode grouping algorithms (Centers for Medicare and Medicaid Services 2015b) seek to assign all patient health-care encounters to mutually exclusive episodes based on their diagnosis and procedure combinations. Such systems are developed with the objective of establishing a comprehensive episode-based case mix classification system, analogous to the long-established diagnosis-related groups (DRGs) and other similar classification systems that categorize hospital inpatient stays into one of several hundred preestablished case mix groups that share

similar clinical and resource utilization characteristics (Fetter et al. 1980).

Developing a case mix classification system is a significant endeavor. Rather than development being limited to a few particular types of episodes of interest, case mix systems operate under principles of being mutually exclusive and comprehensively exhaustive: thus, an effective episode grouping system (also known as a “grouper”) would feature logic to assign every health-care service claim or encounter record to a particular type of episode, selected from a limited list of episode categories.

From the researcher's perspective, the decision on the appropriate approach here depends on the objectives of the analysis: if the objective is to develop an episode-based payment system that provides payments for all health-care services through an “episode bundle,” a full case mix system will be required to ensure all patients are assigned to a particular category. If the idea is to simply focus on analyzing a few different types of episodes, a full case mix system will not be required, although an existing public domain or commercial episode grouping product could be applied to define any number of episodes based on preexisting grouping algorithms. If an existing episode grouping solution is applied, the researcher is advised to acquire a thorough understanding of the underlying clinical logic of the software.

### Risk Adjustment and Severity Classification

A key enhancement made in the 1990s over the basic episode concept of episode grouping and classification systems was the development of episode-based risk adjustment models. Wingert et al. (1995) first noted the need to incorporate severity adjustment into episode-based analyses, beyond that offered by a diagnosis-based classification system (Wingert et al. 1995).

Some episode grouping methodologies such as the ETGs employ a hierarchy of subcategories within each type of episode to differentiate between episodes of different severity levels.

These subcategories may be defined with a variety of proxy data points, including patient characteristics such as comorbidities or the type of health-care services received. For example, a diabetes episode restricted to ambulatory services may be assigned to a lower severity level than a diabetes episode that includes a hospitalization for complications of diabetes. The use of different severity categories within episode groups allows for the expected cost (or sometimes, price) of the episode to differ by severity level, in order to compare “apples to apples” in performance profiling applications or ensuring fair reimbursement levels in funding applications.

Even with the use of severity levels within episode groups, there may still be challenges with episode heterogeneity: MaCurdy et al. conducted an extensive series of simulation analyses using proprietary episode groupers and found substantial residual variation in unexplained costs within each severity grouping (MaCurdy et al. 2009). Certain types of health-care utilization that may potentially be included in the scope of the episode have been found to contribute substantial portions of this unexplained cost variation: Vertrees and other researchers with 3M Inc. examined a variety of different sets of parameters for defining post-acute episode windows and found that by excluding readmissions from the episode, the performance of existing case mix systems in terms of predicting total episode costs was vastly improved (Vertrees et al. 2013). In addition to methods for risk adjustment within episode groups, some commercial groupers such as the ETG and MEG methodologies also enable the user to calculate an aggregate risk score for an individual based on their total episode history in a given time period. In such applications, a total risk score is calculated based on the sum of individual risk scores assigned to each type of episode experienced by an individual.

### Attributing Episodes to Providers

Episodes of care may be used in applications that involve assigning an episode to a particular provider entity: for example, comparing the relative

cost performance of physicians or determining what providers would be eligible to receive a share of a bundled payment. In such applications, business rules must be defined for the attribution of the episode to one or more providers. A variety of approaches to this task are possible and have been explored in the literature. Using a retrospective approach to assigning episodes to providers based on historical fee-for-service claims data, Hussey et al. (2009) examined the impacts of alternate rules for assigning episodes of care to physicians and facilities, with options including attribution to a single physician or facility with the highest total charges in retrospective claims, assignment to a group of physicians or facilities that met a minimum threshold of 25 % of total charges, and assignment to the physician with the highest proportion of evaluation and management claim charges, using the rationale that this physician was likely to be the “most responsible” for managing the patient’s care. They concluded that the performance of alternate rules depended significantly on the trajectory of the condition studied: for example, a largely hospital-based episode such as myocardial infarction was more easily assigned to a single facility and physician than a largely ambulatory-based episode such as diabetes, where facilities played a relatively minor role and a larger number of providers were involved in providing care to individual patients (Hussey et al. 2009).

### Policy Applications

Up until the 1990s, the use of episode of care methods was mainly confined to research-oriented applications and focused on a small set of conditions or procedures. In parallel, in the 1980s the US health policy landscape was transformed with the development and wide-scale use of the DRGs acute inpatient case mix classification system (Fetter et al. 1980). This was first developed for the purposes of utilization review and then subsequently, and most importantly, applied for the purposes of Medicare hospital payment.

In the 1990s, the first commercial episode-based case mix classification systems emerged and began to be employed by insurers and health maintenance



organizations for comparing efficiency across groups of providers (Wingert et al. 1995). These early efforts evolved into well-developed commercial platforms such as the ETGs (Optum 2015) and the MEGs (MaCurdy et al. 2009). The ETGs and MEGs both use a flexible time window used to delineate different episodes. Episode-based classification software enabled commercial insurers to assign all their claims and encounter data to distinct episodes, advancing the practical use of episodes of care for policy applications such as payment and physician profiling.

In the past decade, episode-based payment and performance measurement approaches have gathered huge momentum in the United States, in large part due to the Medicare Payment Advisory Committee (MedPAC)'s endorsement of bundled payment approaches as a transformative alternative to the predominantly fee-for-service payment systems employed in the United States. In their influential 2008 report, *Reforming the Medicare Delivery System*, MedPAC put forward a strong case backed by extensive analysis for a nationwide shift toward bundled payments for episodes of care defined by an acute hospitalization and a fixed window of post-acute care services (Medicare Payment Advisory Committee 2008). Such a payment approach, MedPAC argued, would have the promise of overcoming several important limitations of Medicare's fee-for-service payment approaches. Payments for episodes of care shared across groups of providers would offer strong financial incentives for groups of physicians, hospitals, and post-acute care providers to work together, coordinate services, and redesign patient pathways to improve efficiency across the episode. Bundled payments would also target observed unwarranted regional variations in the provision of post-acute care services for similar types of patients, where some areas made much more use of more costly and intensive settings such as inpatient rehabilitation beds and skilled nursing facilities than others. Finally, bundled payments would drive improved quality of care by ensuring that providers would be forced to absorb the costs of unplanned readmissions and complications occurring following discharge from acute care, as opposed to the "double

payment" providers effectively received for such incidents under the fee-for-service payment system.

Building on the success of some earlier bundled payment pilot programs that employed limited episodes of care focused on hospital and physician services within single acute care stays, in 2011 the Centers for Medicare and Medicaid Services announced the "Bundled Payments for Care Improvement" (BPCI) initiative, a landmark demonstration project that allowed providers to volunteer for participation in a suite of bundled payment options, including episodes indexed by an acute inpatient hospitalization for a set of eligible conditions and extending into either 30, 60, or 90 days of post-acute care and episodes limited to post-acute care only with similar 30, 60, or 90 day time window options. All Medicare Part A and Part B services are included in the episode. For each episode, a single payment is determined for the group of providers based on their historical service claims for similar episodes previously provided and adjusted for regional and national spending levels.

The majority of BPCI participants are enrolled in "retrospective" models, where providers continue to be paid on a fee-for-service basis followed by an episode-based reconciliation against the target total "price" for the episode by all providers participating in the demonstration project. Thus, groups of providers that are able to deliver episodes at a significantly lower cost than their target price are eligible for a share in the savings, whereas providers that exceed the target price may be eligible to return a share of the overspending to Medicare. As of July 2015, there were over 2000 provider entities that had contracted to participate in one of the BPCI models (Centers for Medicare and Medicaid Services 2015).

As these and other current major episode-driven policy initiatives in the United States, the Netherlands, Sweden, and elsewhere make abundantly clear, the episode of care is currently experiencing a renaissance in terms of its use as a foundational analytic construct to support payment system design, performance measurement initiatives, and a wide variety of health services research applications.

## References

- Birkmeyer JD, Gust C, Baser O, et al. Medicare payments for common inpatient procedures: implications for episode-based payment bundling. *Health Serv Res.* 2010;45(6 Pt 1):1783–95.
- Canadian Institute for Health Information. Canadian Institute for Health Information. 2015. [www.cihi.ca](http://www.cihi.ca). Accessed 27 Oct 2015.
- Cave DG. Profiling physician practice patterns using diagnostic episode clusters. *Med Care.* 1995;33(5):463–86.
- Centers for Medicare and Medicaid Services. Bundled Payments for Care Improvement (BPCI) initiative: general information. 2014. <https://innovation.cms.gov/initiatives/bundled-payments/>. Accessed 27 Oct 2015
- Centers for Medicare and Medicaid Services. Supplemental QRURs and episode-based payment measurement. 2015a. <https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeedbackProgram/Episode-Costs-and-Medicare-Episode-Grouper.html>. Accessed 28 Oct 2015.
- Centers for Medicare and Medicaid Services. Bundled payments for care improvement (BPCI) initiative fact sheet. 2015b. <https://www.cms.gov/Newsroom/MediaReleaseDatabase/Fact-sheets/2015-Fact-sheets-items/2015-08-13-2.html>. Accessed 28 Oct 2015.
- Chamberlayne R, Green B, Barer ML, et al. Creating a population-based linked health database: a new resource for health services research. *Can J Public Health.* 1998;89(4):270–3.
- Codman EA. The product of a hospital. *Arch Pathol Lab Med.* 1914;1990;114(11):1106–11.
- Falk IS, Schonfeld HK, Harris BR, et al. The development of standards for the audit and planning of medical care: I. Concepts, research design, and the content of primary physician's care. *Am J Public Health.* 1967;57: 1118–36.
- Feldstein PJ. Research on the demand for health services. *Health Serv Res.* 1966;44(3):128–65.
- Fetter RB, Shin Y, Freeman JL, et al. Case mix definition by diagnosis-related groups. *Med Care.* 1980;18 ((2 Suppl) iii):1–53.
- Fransoo R, Yogendran M, Olafson K, et al. Constructing episodes of inpatient care: data infrastructure for population-based research. *BMC Med Res Methodol.* 2012;12:133.
- Health Quality Ontario. Quality-based procedures: clinical handbook for hip fracture. Toronto: Health Quality Ontario; 2013. <http://www.hqontario.ca/evidence/publications-and-ohtac-recommendations/clinical-handbooks>. Accessed 27 Oct 2015.
- High Value Health Care Project. 2011. [http://www.rwjf.org/content/dam/farm/reports/program\\_results\\_reports/2011/rwjf71110](http://www.rwjf.org/content/dam/farm/reports/program_results_reports/2011/rwjf71110). Accessed 27 Oct 2016.
- Hornbrook MC, Hurtado AV, Johnson RE. Health care episodes: definition, measurement and use. *Med Care Rev.* 1985;42(2):163–218.
- Hussey P, Sorbero M, Mehrotra A, et al. Using episodes of care as a basis for performance measurement and payment: moving from concept to practice. *Health Aff (Project Hope).* 2009;28(5):1406–17.
- Jain R, Basinski A, Kreder HJ. Nonoperative treatment of hip fractures. *Int Orthop.* 2003;27(1):11–7.
- Lohr KN, Brook RH. Quality of care in episodes of respiratory illness among Medicaid patients in New Mexico. *Ann Intern Med.* 1980;92(1):99–106.
- MaCurdy T, Kerwin J, Theobald N. Need for risk adjustment in adapting episode grouping software to Medicare data. *Health Care Financ Rev.* 2009;30(4):33–46.
- Medicare Payment Advisory Committee. Report to the congress: reforming the delivery system; 2008. [http://www.medpac.gov/documents/reports/Jun08\\_EntireReport.pdf](http://www.medpac.gov/documents/reports/Jun08_EntireReport.pdf). Accessed 28 Oct 2015.
- Meier DE. Increased access to palliative care and hospice services: opportunities to improve value in health care. *Milbank Q.* 2011;89(3):343–80.
- Menzies IB, Mendelson DA, Kates SL, et al. Prevention and clinical management of hip fractures in patients with dementia. *Geriatr Orthop Surg Rehabil.* 2010;1 (2):63–72.
- Moscovice I. Selection of an appropriate unit of analysis for ambulatory care settings. *Med Care.* 1977;15 (12):1024–44.
- Nightingale F. Notes on hospitals. London: Longman, Gree, Longman, Roberts and Green; 1863.
- Nutting PA, Shorr GI, Burkhalter BR. Assessing the performance of medical care. *Med Care.* 1981;21(3): 281–96.
- Optum. Symmetry Episode Treatment Groups. 2015. <https://www.optum.com/providers/analytics/health-plan-analytics/symmetry/symmetry-episode-treatment-groups.html>. Accessed 27 Oct 2015.
- Reschovsky JD, Hadley J, O'Malley AJ, et al. Geographic variations in the cost of treating condition-specific episodes of care among Medicare patients. *Health Serv Res.* 2014;49(1):32–51.
- Schonfeld HK, Falk IS, Lavietes PH, et al. The development of standards for the audit and planning of medical care: pathways among primary physicians and specialists for diagnosis and treatment. *Med Care.* 1968; 6(2):101–14.
- Schulman KA, Yabroff KR, Kong J, et al. A claims data approach to defining an episode of care. *Health Serv Res.* 1999;34(2):603–21.
- Scitovsky A. Changes in the costs of treatment of selected illnesses, 1951–65. *Am Econ Rev.* 1967;57(5): 1182–95.
- Sheehan KJ, Sobolev B, Bohm E, Sutherland J, Kuramoto L, Guy P, Hellsten E, Jaglal S for the Canadian Collaborative on Hip Fractures. Constructing episode of care from acute hospital records for studying effects of timing of hip fracture surgery. *J Orthop Res.* 2016; 34(2):197–204.
- Sobolev BG, Kuramoto L. Analysis of waiting-time data in health services research. New York: Springer; 2007.
- Sobolev BG, Levy AR, Kuramoto L, et al. Do longer delays for coronary artery bypass surgery contribute

- to preoperative mortality in less urgent patients? *Med Care*. 2006;44(7):680–6.
- Solon J, Sheps CG, Lee SS. Delineating patterns of medical care. *Am J Public Health Nations Health*. 1960;50(8):1105–13.
- Solon JA, Feeney JJ, Jones SH, et al. Delineating episodes of medical care. *Am J Public Health Nations Health*. 1967;57(3):401–8.
- Solon JA, Rigg RD, Jones SH, et al. Episodes of care: nursing students' use of medical services. *Am J Public Health Nations Health*. 1969;59(6):936–46.
- Struijs JN, De Jong-Van Til JT, Lemmens LC, Drewes HW, De Bruin SR, Baan CA. Three years of bundled payment for diabetes care in the Netherlands. Impact on health care delivery process and the quality of care. RIVM Report. 2012a;260013001. [https://www.researchgate.net/profile/Jeroen\\_Struijs/publication/233407675\\_Three\\_years\\_of\\_bundled\\_payment\\_for\\_diabetes\\_care\\_in\\_the\\_Netherlands\\_Effect\\_on\\_health\\_care\\_delivery\\_process\\_and\\_the\\_quality\\_of\\_care/links/09e4150a50b96ad6cb000000.pdf](https://www.researchgate.net/profile/Jeroen_Struijs/publication/233407675_Three_years_of_bundled_payment_for_diabetes_care_in_the_Netherlands_Effect_on_health_care_delivery_process_and_the_quality_of_care/links/09e4150a50b96ad6cb000000.pdf). Accessed 27 Oct 2016.
- Struijs JN, Mohnen SM, Molema CC, De Jong-van Til JT, Baan CA. Effects of bundled payment on curative health care costs in the Netherlands: an analysis for diabetes care and vascular risk management based on nationwide claim data, 2007-2010. RIVM Report. 2012b;260013001. <http://rivm.openrepository.com/rivm/handle/10029/257206>. Accessed 27 Oct 2016.
- Sutherland JM, Hellsten E, Yu K. Bundles: an opportunity to align incentives for continuing care in Canada? *Health Policy (Amsterdam, Netherlands)*. 2012;107(2–3):209–17.
- Vertrees JC, Averill RF, Eisenhandler J, et al. Bundling post-acute care services into MS-DRG payments. *Medicare Medicaid Res Rev*. 2013;3(3):E1–E19.
- Volk T, Hahn L, Hayden R, et al. Reliability audit of a regional cardiac surgery registry. *J Thorac Cardiovasc Surg*. 1997;114(6):903–10.
- White KL, Williams TF, Greenberg BG. The ecology of medical care. *N Engl J Med*. 1961;265:885–92.
- Wingert TD, Kralewski JE, Lindquist TJ, et al. Constructing episodes of care from encounter and claims data: some methodological issues. *Inquiry*. 1995;32(4):430–43.



# Health Services Information: Lessons Learned from the Society of Thoracic Surgeons National Database

# 10

David M. Shahian and Jeffrey P. Jacobs

## Contents

<b>Introduction</b> .....	218
The Evolution of Healthcare Quality Measurement and Clinical Registries .....	218
<b>Database Structure</b> .....	220
STs Adult Cardiac Surgery Database (STs-ACSD) .....	220
STs Congenital Heart Surgery Database (STs-CHSD) .....	223
STs General Thoracic Surgery Database (STs-GTSD) .....	226
<b>Database Operations</b> .....	227
Data Sources .....	227
Vendors .....	228
STs Staff .....	228
Data Warehouse and Analytic Center .....	228
Data Quality and Audit .....	229
STs Quality Measurement Task Force (STs-QMTF) .....	229
STs Quality Initiatives Task Force (STs-QIT) .....	233
STs Public Reporting Task Force .....	234
STs Research Center .....	235
STs Task Force on Longitudinal Follow-Up and Linked Registries (STs-LFLR) ....	235
Device Surveillance .....	236
<b>Summary</b> .....	236
<b>References</b> .....	237

---

D. M. Shahian (✉)  
Department of Surgery and Center for Quality and Safety,  
Massachusetts General Hospital, Harvard Medical School,  
Boston, MA, USA  
e-mail: [dshahian@partners.org](mailto:dshahian@partners.org)

J. P. Jacobs  
Division of Cardiac Surgery, Department of Surgery, Johns  
Hopkins University School of Medicine, Baltimore, MA,  
USA

Johns Hopkins All Children's Heart Institute, Saint  
Petersburg/Tampa, FL, USA

**Abstract**

The Society of Thoracic Surgeons (STS) National Database was initiated in 1989 with the goal of providing accurate clinical data to support quality assessment and improvement activities in cardiothoracic surgery. Participation among adult cardiac surgery centers and pediatric cardiac surgery centers in the USA currently exceeds 90 %, and the STS National Database now also includes a general thoracic surgery registry with growing national penetration.

The specific functions of the STS National Database have also evolved, as reflected in its various task forces. Quality assessment remains the primary function, and the STS Quality Measurement Task Force is responsible for developing risk models and performance metrics (often composites) for frequently performed cardiothoracic procedures. Each quarter, participants in the STS Adult Cardiac Database are provided with detailed feedback reports of their practice characteristics, including demographics, risk factor profiles, operative data, and outcomes benchmarked against STS national data. Similar feedback reports are provided to participants in the STS Congenital Heart Surgery Database and the STS General Thoracic Database every 6 months. In addition, given its belief in accountability, STS established a Public Reporting Task Force to coordinate voluntary public reporting initiatives using the *Consumer Reports* or STS websites.

The ultimate goal of all database activities is to improve patient outcomes, and specific quality improvement projects are developed and led by the STS Quality Initiatives Task Force. The STS Task Force on Longitudinal Follow-Up and Linked Registries coordinates the linkage of STS clinical registry data with complementary information regarding long-term outcomes and costs from other data sources. Additional STS National Database Task Forces include International Relations, Appropriateness, Cost and Resource Use, Dashboards, and Informatics.

In summary, the STS National Database is a uniquely valuable resource that is largely responsible for the dramatic improvements in cardiothoracic surgical outcomes that have occurred over the past quarter century.

---

**Introduction****The Evolution of Healthcare Quality Measurement and Clinical Registries**

Valid and reliable assessment of healthcare performance requires high-quality data, appropriate analytical methodologies, modern computing power, and, most importantly, a conceptual framework. Regarding the latter, several healthcare leaders were prescient in their recognition of the need to collect, analyze, and publish the outcomes of medical and surgical care.

Florence Nightingale, best known as the founder of modern nursing, is less well recognized for her seminal contributions to public health research and provider profiling (Spiegelhalter 1999; Iezzoni 2003). Upon her return to England after service in the Crimean War, she published mortality rates of English hospitals, using approaches that were admittedly flawed by today's standards. However, this publication, which was roughly contemporaneous with the American Civil War, represents the first time that outcomes rates for a diverse group of hospitals were compared.

In the early 1900s, Ernest Amory Codman, a surgeon at Boston's Massachusetts General Hospital, was distraught by the lack of objective data regarding surgeon performance, as well as the lack of correlation between surgeon's results and their reputations or hospital leadership positions. Codman incurred the wrath of the Boston medical community when he unveiled a large cartoon depicting an ostrich-goose laying golden "humbug" eggs for the well-heeled residents of the Back Bay, who were woefully ignorant of the results actually produced by their doctors. He famously wrote (Spiegelhalter 1999; Codman 1914, 1995; Donabedian 1989; Mallon 2000; Neuhauser 1990; Passaro et al. 1999):

I am called eccentric for saying in public that hospitals, if they wish to be sure of improvement. . .

- Must find out what their results are.
- Must analyze their results to find their strong and weak points.
- Must compare their results with those of other hospitals.
- Must care for what cases they can care for well, and avoid attempting to care for cases which they are not qualified to care for well.
- Must welcome publicity not only for their successes, but for their errors, so that the public may give them their help when it is needed.
- Must promote members of the medical staff on the basis which gives due consideration to what they can and do accomplish for their patients.
- Such opinions will not be eccentric a few years hence

Codman started his own End Result Hospital built upon these principles, but it eventually closed. Although Codman was ridiculed and disdained by many colleagues at the time, his work led directly to the formation of the American College of Surgeons and the Joint Commission on Accreditation of Healthcare Organizations.

The third visionary leader in healthcare quality measurement was Professor Avedis Donabedian at the University of Michigan (Donabedian 1966, 1988). Donabedian was the first to propose that healthcare quality could be measured using structure (e.g., 24/7 intensivist availability, nursing ratios, adoption of computerized physician order entry), process (e.g., achieving an “open artery” within 90 min for patients suffering an acute MI, administering aspirin to acute MI patients), and outcomes (e.g., mortality, complications, readmissions, patient-reported outcomes). Donabedian stressed that “Outcomes, by and large, remain the ultimate validators of the effectiveness and quality of medical care” (Donabedian 1966), anticipating the current emphasis on outcomes measurement as the optimal way to assess quality in healthcare.

The science and technology necessary to actualize the conceptual framework of Nightingale, Codman, and Donabedian did not become widely available until the latter half of the twentieth century. The enactment of Medicare legislation in 1965 resulted in a huge new claims data source

which could be used for provider profiling, research, and policy development. The use of statistical techniques such as logistic regression and hierarchical regression expanded dramatically in the latter half of the twentieth century, facilitated in large part by the exponential growth of computing power and mass data storage capacity.

Another essential component for the development of robust quality assessment and improvement was the evolution of clinical data registries. Several seminal events in the mid- and late 1980s provided the proximate stimulus for the development of cardiac surgery databases, including the Society of Thoracic Surgeons (STS) National Database, which was the first large-scale clinical registry developed by a professional society. On March 12, 1986, the Health Care Financing Administration (HCFA, the predecessor of the Centers for Medicare and Medicaid Services or CMS) published a list of hospital mortality rates which were based on administrative claims data and minimally adjusted for patient risk. This was referred to by some as the Medicare “death list,” and it was widely criticized for its methodological shortcomings. However, despite these flaws, it was apparent to some farsighted leaders that this was the beginning of a new era in healthcare transparency. Among those who envisioned this future state were the leaders of STS. The most commonly performed procedure by members of that organization, coronary artery bypass grafting surgery (CABG), was a natural target for early efforts to assess performance. CABG was one of the most frequently performed and costly procedures in healthcare at that time and had well-defined outcomes including mortality, stroke, reoperation, kidney failure, and infections. Owing in part to a torrent of requests from STS members who believed that the HCFA “death list” had mischaracterized their programs as underperforming, STS leaders recognized the inadequacy of using minimally adjusted claims data to evaluate program performance. An ad hoc committee on risk factors was developed by STS (Kouchoukos et al. 1988) in order to define those patient factors that would be required to fairly adjust for inherent patient risk.

These were incorporated into what subsequently was to become the STS National Database (STS National Database 2014), which was made available to STS members in 1989 under the direction of Dr. Richard Clark (Clark 1989; Grover et al. 2014).

The 1986 HCFA release of mortality reports also stimulated the development of other cardiac surgery database and performance monitoring initiatives. In New York State, Dr. David Axelrod, the commissioner of health, was aware of a fivefold variation in unadjusted mortality rates for coronary artery bypass grafting surgery (CABG) among the 28 cardiac surgical programs in that state. However, he and the New York Cardiac Advisory Committee recognized that acting upon this data would be challenging, as low-performing hospitals would likely assert that their patients were “sicker,” just as they had when HCFA released its mortality reports. Accordingly, in collaboration with Dr. Edward Hannan, a clinical data registry for CABG was developed (the New York Cardiac Surgery Reporting System or CSRS) (Hannan et al. 2012). Using these data, expected results for each patient were estimated and aggregated to the program and surgeon levels. Comparing observed and expected results made it possible to generate risk-standardized mortality rates and ratios, and these were first released to the public in 1990. These results demonstrated that not only was there wide variation in unadjusted mortality rates but also in risk-adjusted mortality rates. Similarly, the Northern New England Cardiovascular Disease Study Group (O’Connor et al. 1991) found wide variation in the ratio of observed to expected mortality among CABG programs in that region.

In summary, the early development of clinical data registries by STS, as well as a few states and regions, was driven by a desire to produce valid, risk-adjusted results that would allow fair comparisons of performance among providers, accounting for the preoperative risk of their patients. Availability of such data would facilitate quality improvement by providers and might also impact consumer choice of providers, shifting market share to better performing groups, although the latter goal has yet to be achieved.

Over the next quarter century, the STS National Database has expanded from its initial focus on adult cardiac surgery, particularly CABG, to encompass all major cardiac surgical procedures in the adult, as well as congenital heart surgery and general thoracic surgery. By 2014, over 1080 programs participated in the STS Adult Cardiac Surgery Database (90–95 % of all US programs), 114 programs were contributors to the STS Congenital Database (95 % of all US programs), and 244 programs participated in the STS General Thoracic Database. Seven international sites also participated. Figures 1, 2, and 3 demonstrate the geographical distribution of participants in the three STS National Databases.

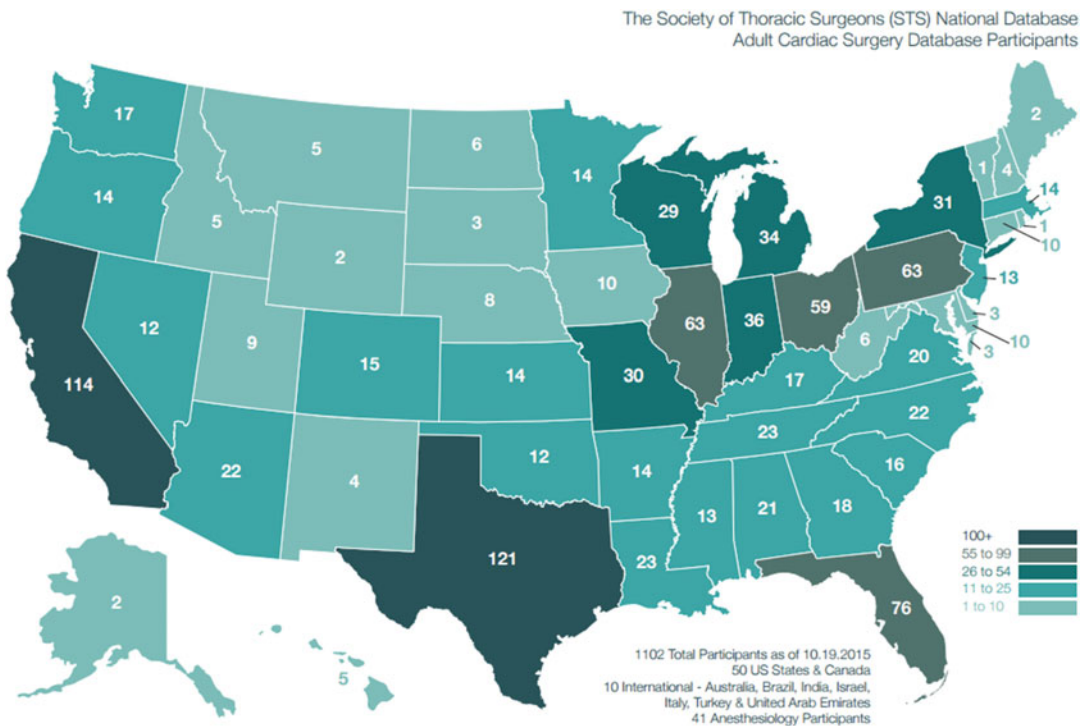
---

## Database Structure

The STS National Database is composed of three clinical specialty databases and 10 functionally oriented, crosscutting task forces (Table 1). Each of the three clinical specialty databases has its own unique features and, in some instances, challenges.

### STS Adult Cardiac Surgery Database (STS-ACSD)

The STS Adult Cardiac Surgery Database (STS-ACSD) is the oldest of the three specialty databases and has the largest number of participants (approximately 1080 in the USA). Based on studies by Jacobs and colleagues, center-level penetration (number of CMS sites with at least one matched STS participant divided by the total number of CMS CABG sites) increased from 83 % to 90 % between 2008 and 2012 (Jacobs et al. 2016). In 2012, 973 of 1,081 CMS CABG sites (90 %) were linked to an STS site. Patient-level penetration (number of CMS CABG hospitalizations done at STS sites divided by the total number of CMS CABG hospitalizations) increased from 89 % to 94 % from 2008 to 2012. In 2012, 71,634 of 76,072 CMS CABG hospitalizations (94 %) were at an STS site. Finally, completeness of case inclusion at STS sites (number of CMS CABG cases at STS sites



**Fig. 1** STS Adult Cardiac Surgery Database Map, accessed July 2, 2016, at [http://www.sts.org/sites/default/files/documents/adultcardiacMap\\_4.pdf](http://www.sts.org/sites/default/files/documents/adultcardiacMap_4.pdf). © The Society of

Thoracic Surgeons, 2016. All rights reserved (Reprinted with permission from STS)

linked to STS records, divided by the total number of CMS CABG cases at STS sites) increased from 97 % to 98 % from 2008 to 2012. In 2012, 69,213 of 70,932 CMS CABG hospitalizations at STS sites (98 %) were linked to an STS record. This suggests that at STS-participating sites that billed CMS for CABG procedures, virtually all these billed cases were captured in the STS National Database. These high degrees of national penetration and completeness, together with high accuracy verified in the ongoing external audits (see “[Data Quality and Audit](#)” section below), are of critical importance when STS advocates for the use of its measures, rather than those based on claims data, in various public reporting programs. Lack of high national penetration is, in fact, a commonly used rationale for the continued use of claims-based metrics in many areas; however, this justification for use of claims-based metrics is clearly not applicable to adult cardiac surgery.

The STS-ACSD now encompasses the entire spectrum of adult cardiac surgery. This includes CABG; surgery of the aortic, mitral, tricuspid, and pulmonary valves; surgery of the thoracic aorta; arrhythmia procedures; and less commonly performed procedures such as pulmonary thromboendarterectomy and removal of tumors of the heart and inferior vena cava. Data are collected regarding:

Patient demographics

Risk factors that may impact the outcomes of surgery

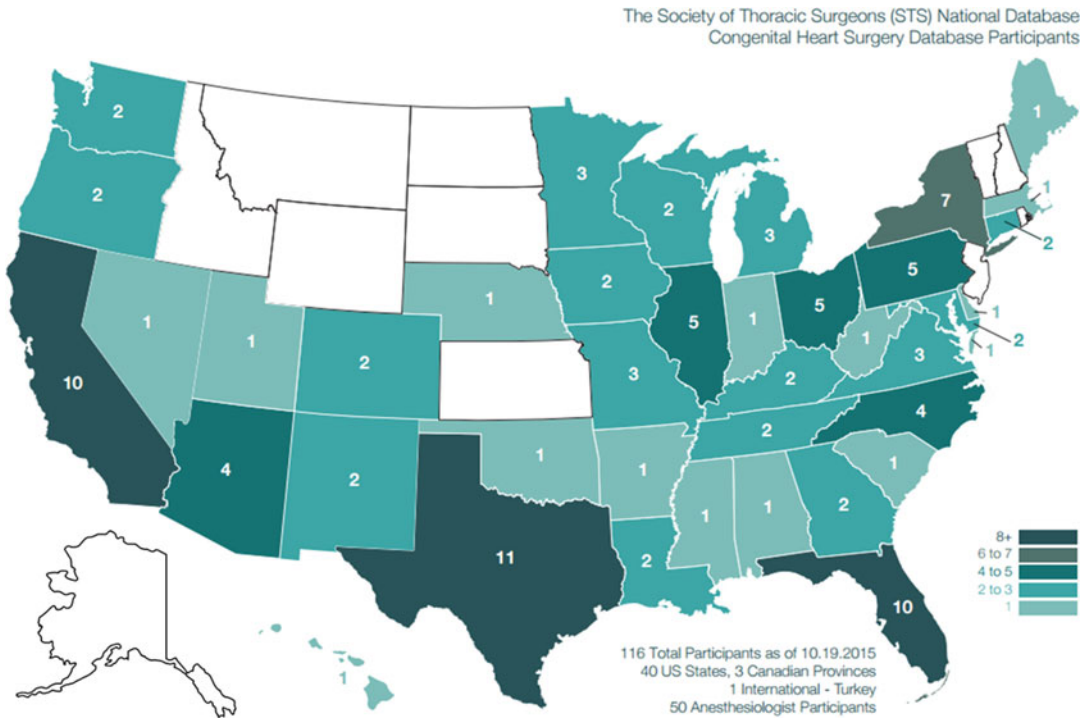
Details of the specific disease process that led to surgery (e.g., degree of coronary artery stenosis in each vessel, etiology and severity of valvular lesions, type of thoracic aortic pathology)

Technical details of the conduct of the procedure that was performed

Detailed clinical outcomes

Disposition of the patient (e.g., home, rehabilitation facility, or deceased)





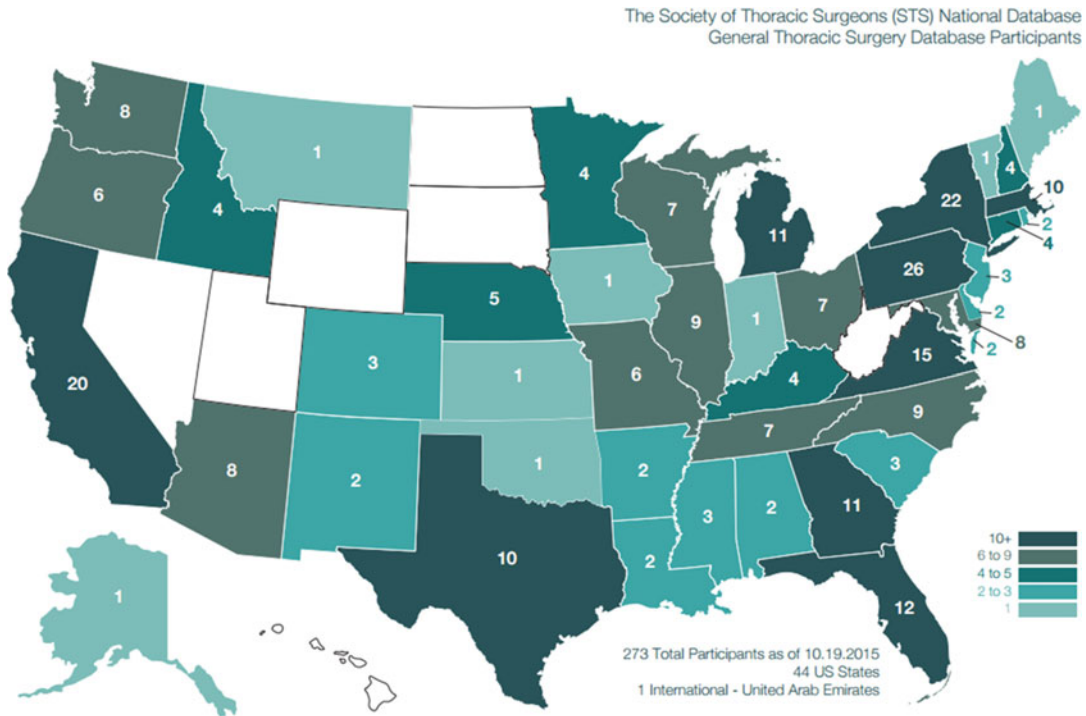
**Fig. 2** STS Congenital Heart Surgery Database Map, accessed July 2, 2016, at [http://www.sts.org/sites/default/files/documents/congenitalMap\\_4.pdf](http://www.sts.org/sites/default/files/documents/congenitalMap_4.pdf). © The Society of

Thoracic Surgeons, 2016. All rights reserved (Reprinted with permission from STS)

Data from the STS-ACSD are reported back to participants on a quarterly basis (STS National Database 2014). These data include the types of procedures performed, demographics and risk factors of the patients, details about the conduct of the surgical procedure, and outcomes. In each case, this information is benchmarked against aggregate data from all STS-participating programs nationally and also against aggregate data from programs that are similar in terms of teaching intensity and size. Finally, participants are given their last several years of data so that important trends may be recognized. Twice yearly, in addition to the routine harvest feedback reports, participants also receive reports of their performance on National Quality Forum (NQF)-endorsed STS metrics and on the various STS composite performance metrics for specific procedures (e.g., isolated CABG, isolated aortic valve replacement, aortic valve replacement plus CABG)

(Shahian et al. 2007a, 2012a, 2014; O'Brien et al. 2007). These performance reports provide numerical point estimates with credible intervals based on a Bayesian hierarchical model, and they also assign participants to a “star rating” category based on the true Bayesian probabilities (e.g., 99 % for isolated CABG) that the provider has worse than expected, as expected, or better than expected performance (see “[STS Quality Measurement Task Force \[STS-QMTF\]](#)” section below). These reports also include guidance as to which areas of performance are most in need of remediation and improvement.

In addition to these regular confidential feedback reports, STS-ACSD data are used for quality assessment, performance improvement initiatives, research, and public reporting and to satisfy regulatory and reimbursement imperatives. Many of these additional functions are discussed in subsequent sections.



**Fig. 3** STS General Thoracic Surgery Database Map, Thoracic Surgeons, 2016. All rights reserved (Reprinted with permission from STS) accessed July 2, 2016, at [http://www.sts.org/sites/default/files/documents/thoracicMap\\_5.pdf](http://www.sts.org/sites/default/files/documents/thoracicMap_5.pdf). © The Society of

**Table 1** STS National Database Task Forces

STS National Database Task Forces	
Specialty-specific task forces	Functional, crosscutting task forces
STS Adult Cardiac Surgery Database (STS-ACSD) Task Force	STS Quality Measurement Task Force (STS-QMTF)
STS Congenital Heart Surgery Database (STS-CHSD) Task Force	STS Quality Improvement Task Force (STS-QIT)
STS General Thoracic Surgery Database (STS-GTSD) Task Force	STS Access and Publications (A and P) Task Force (within the STS Research Center)
	STS International Relations Task Force
	STS Task Force on Longitudinal Follow-up and Linked Registries (STS-LFLR) (within the STS Research Center)
	STS Appropriateness Task Force
	STS Public Reporting Task Force
	STS Cost and Resource Use Task Force
	STS Dashboards Task Force
	STS Informatics Task Force

**STS Congenital Heart Surgery Database (STS-CHSD)**

The report of the 2010 STS Congenital Heart Surgery Practice and Manpower Survey,

undertaken by the STS Workforce on Congenital Heart Surgery, documented that 125 hospitals in the USA and 8 hospitals in Canada perform pediatric and congenital heart surgery (Jacobs et al. 2011a). In 2014, the STS Congenital Heart

Surgery Database (STS-CHSD) included 114 congenital heart surgery programs representing 119 of the 125 hospitals (95.2 % penetrance by hospital) in the USA and 3 of the 8 centers in Canada.

The analysis of outcomes of patients undergoing pediatric and congenital cardiac surgery presents several unique challenges in the domains of nomenclature and risk adjustment. Unlike adult cardiac surgery where the majority of operations involve CABG, aortic valve replacement, and mitral valve replacement or repair or a combination of these, congenital cardiac surgery involves a much wider variety of procedures.

One of the greatest challenges in the development and application of the STS-CHSD has involved standardization of nomenclature and definitions related to surgery for pediatric and congenital cardiac disease. During the 1990s, both the European Association for Cardio-Thoracic Surgery (EACTS) and STS created databases to assess the outcomes of congenital cardiac surgery. Beginning in 1998, these two organizations collaborated to create the International Congenital Heart Surgery Nomenclature and Database Project. By 2000, a common nomenclature and a common core minimal data set were adopted by EACTS and STS and published in the *Annals of Thoracic Surgery* (Mavroudis and Jacobs 2000; Franklin et al. 2008). In 2000, The International Nomenclature Committee for Pediatric and Congenital Heart Disease was established. This committee eventually evolved into the International Society for Nomenclature of Paediatric and Congenital Heart Disease (ISNPCHD). By 2005, members of the ISNPCHD crossmapped the nomenclature of the International Congenital Heart Surgery Nomenclature and Database Project of the EACTS and STS with the European Paediatric Cardiac Code (EPCC) of the Association for European Paediatric Cardiology (AEPCC) and therefore created the International Pediatric and Congenital Cardiac Code (IPCCC) (Franklin et al. 2008; Jacobs et al. 2008), which is available for free download from the Internet at <http://www.IPCCC.NET>. This common nomenclature, the IPCCC, and the common minimum database data set created by the International Congenital

Heart Surgery Nomenclature and Database Project are now utilized by the STS-CHSD, the EACTS Congenital Heart Surgery Database (EACTS-CHSD), and the Japan Congenital Cardiovascular Surgery Database (JCCVSD). As of January 1, 2014, the STS-CHSD contains data from 292,828 operations; the EACTS-CHSD contains data from over 157,772 operations; and the JCCVSD contains data from over 29,000 operations. Therefore, the combined data set of the STS-CHSD, the EACTS-CHSD, and the JCCVSD contains data from over 479,000 operations performed between 1998 and January 1, 2014, inclusive, all coded with the EACTS-STS-derived version of the IPCCC, and all coded with identical data specifications.

Similar to investigations of data sources used for adult cardiac surgery studies, several studies have examined the relative utility of clinical and administrative nomenclature for the evaluation of quality of care for patients undergoing treatment for pediatric and congenital cardiac disease. Given the far greater diversity of anatomic lesions and procedures compared with adult cardiac surgery, it is not surprising that the superiority of clinically rich data sources is even more apparent in congenital heart disease. Evidence from several investigations suggests inferior accuracy of coding of lesions in the congenitally malformed heart using administrative databases and the ninth revision of the International Classification of Diseases (ICD-9) (Cronk et al. 2003; Frohnert et al. 2005; Strickland et al. 2008; Pasquali et al. 2013; Jantzen et al. 2014). Analyses based on the codes available in ICD-9 are likely to have substantial misclassification of congenital cardiac disease. Furthermore, differences in case ascertainment between administrative and clinical registry data for children undergoing cardiac operations can translate into important differences in outcomes assessment.

Risk modeling is essential when assessing and comparing healthcare performance among programs and surgeons, as this adjusts for differences in the complexity and severity of patients they treat. Reliably accounting for the risk of adverse outcomes mitigates the possibility that providers caring for sicker patients will be unfairly

penalized, as their unadjusted results may be worse simply because of case mix (Shahian et al. 2013a). However, formal risk modeling is challenging for rare operations because sample sizes are small. Risk adjustment in congenital cardiac surgery is particularly challenged by this reality, as the specialty is defined by a very wide variety of operations, many of which are performed at a relatively low volume. Consequently, the STS-CHSD has implemented a methodology of risk adjustment based on complexity stratification. Complexity stratification provides an alternative methodology that can facilitate the analysis of outcomes of rare operations by dividing the data into relatively homogeneous groups (called strata). The data are then analyzed within each stratum.

Three major multi-institutional efforts have used complexity stratification to measure the complexity and potential risk of congenital cardiac surgical operations (Jacobs et al. 2009; O'Brien et al. 2009a):

1. **Risk Adjustment in Congenital Heart Surgery-1 methodology (RACHS-1 methodology)**
2. **Aristotle Basic Complexity Score (ABC Score)**
3. **STS-EACTS Congenital Heart Surgery Mortality Categories (STS-EACTS Mortality Categories) (STAT Mortality Categories)**

RACHS-1 and the ABC Score were developed at a time when limited multi-institutional clinical data were available and were therefore based in a large part on subjective probability (expert opinion). The STAT Mortality Categories are a tool for complexity stratification that was developed from an analysis of 77,294 operations entered into the EACTS-CHSD (33,360 operations) and the STS-CHSD (43,934 patients) between 2002 and 2007. Procedure-specific mortality rate estimates were calculated using a Bayesian model that adjusted for small denominators. Operations were sorted by increasing risk and grouped into five categories (the STAT Mortality Categories) that were designed to be optimal with respect to minimizing within-category variation and maximizing between-category variation. STS and

EACTS have transitioned from the primary use of Aristotle and RACHS-1 to the primary use of the STAT Mortality Categories for three major reasons:

1. STAT Score was developed primarily based on objective data, while RACHS-1 and Aristotle were developed primarily on expert opinion (subjective probability).
2. STAT Score allows for classification of more operations than RACHS-1 or Aristotle.
3. STAT Score has a higher c-statistic than RACHS-1 or Aristotle.

Data from the STS-CHSD are reported back to participants every 6 months in feedback reports. Similar to the STS-ACSD, the data in these feedback reports include the types of procedures performed, demographics and risk factors of the patients, details about the conduct of the surgical procedure, and outcomes. In each case, this information is benchmarked against aggregate data from all participants in the STS-CHSD. Participants are given their last 4 years of data so that important trends may be recognized. The feedback report also includes an assessment of programmatic performance using the empirically derived 2014 STS Congenital Heart Surgery Database Mortality Risk Model that incorporates both procedural stratification by STAT Mortality Category and patient factors. This 2014 STS-CHSD Mortality Risk Model includes the following covariates:

- STAT Mortality Category
- Age
- Previous cardiovascular operation(s)
- Any noncardiac abnormality
- Any chromosomal abnormality or syndrome
- Important preoperative factors (mechanical circulatory support, shock persisting at time of surgery, mechanical ventilation, and renal dysfunction)
- Any other preoperative factors
- Prematurity (for neonates only)
- Weight (for neonates only)
- Weight-for-age-and-sex Z-score (for infants only)

Centers for which the 95 % confidence interval for observed-to-expected mortality ratio does not include unity (does not overlap with the number one) are identified as one-star (low-performing) or three-star (high-performing) programs with respect to operative mortality. Star ratings are provided for the single category of ‘all ages and all STAT Categories.’ Public reporting of data from the STS-CHSD began in January 2015 using this star rating system, with reporting of both star ratings and the actual numerical mortality data on which the star rating is based. As of March 2016, 68 out of 113 (60.2 %) participants in STS-CHSD from the United States had agreed to publicly report their outcomes using this system.

Data quality in the STS-CHSD is evaluated through intrinsic data verification, including identification and correction of missing/out-of-range values and inconsistencies across fields and on-site audit. In 2014, approximately 10 % of participants (11 participants) will be randomly selected for audits of their center. The audit is designed to complement the internal quality controls. Its overall objective is to maximize the integrity of the data in the STS-CHSD by examining the accuracy, consistency, and completeness of the data. In 2013, the audit of the STS-CHSD included the following documentation of rates of completeness and accuracy for the specified fields of data:

- Primary diagnosis (completeness = 100 %, accuracy = 96.2 %)
- Primary procedure (completeness = 100 %, accuracy = 98.7 %)
- Mortality status at hospital discharge (completeness = 100 %, accuracy = 98.8 %)

Similar to the STS-ACSD, in addition to regular confidential feedback reports, STS-CHSD data are used for quality assessment, performance improvement initiatives, research, and public reporting (beginning in early 2015) and to satisfy regulatory and reimbursement imperatives. Many of these additional functions are discussed in subsequent sections.

## **STS General Thoracic Surgery Database (STS-GTSD)**

The STS General Thoracic Database (STS-GTSD) is the newest of the three specialty databases, and it faces a unique challenge. Unlike adult and congenital heart surgery, both of which are practiced almost exclusively by board-certified cardiothoracic (CT) surgeons, general thoracic surgery in the USA is more often performed by general surgeons or by surgical oncologists. These surgeons are allowed to submit data to the STS National Database, but they rarely take advantage of this opportunity. Therefore, there are essentially two populations of patients undergoing noncardiac chest surgery in the USA. In the first group are patients operated upon by board-certified CT surgeons, many of whom are involved in academic or referral centers and most of whom participate in the STS-GTSD. The second group of patients is operated upon by surgeons who are not board-certified thoracic surgeons, who rarely if ever participate in the STS National Database, and who do not receive regular feedback information on their performance from the STS-GTSD. This diverse population of surgeons performing general thoracic surgery is an important consideration when assessing the performance of an STS-GTSD program, as their benchmark population of providers is already preselected to be among the best thoracic surgeons in the nation. An average STS-GTSD participant program may well have performance that substantially exceeds that of procedures performed by non-board-certified surgeons. Potentially useful areas of performance comparison include adequacy of preoperative staging, functional evaluation, intraoperative lymph node sampling, and morbidity and mortality.

Despite this challenge, the STS-GTSD is growing, and in 2015, it enrolled patients from 273 participants. External audit revealed high accuracy (overall 95 %). Mortality and morbidity risk models for lung cancer and esophageal resection have been developed in collaboration with the STS Quality Measurement Task Force (QMTF) (Kozower et al. 2010; Shapiro et al. 2010;

Wright et al. 2009), and performance metrics using these risk models will be used to classify thoracic programs as one, two, or three stars, similar to the approach used in adult cardiac surgery. Because STS-GTSD participants represent a high-performing subset of all US surgeons performing general thoracic procedures, STS has also compared the unadjusted results of STS surgeons with those available from the Nationwide Inpatient Sample (NIS) for all surgeons performing chest operations nationally. This comparison has revealed that surgeons who are actively participating in the STS-GTSD have superior results, likely both because of their specialized training as well as the feedback reports they receive.

Similar to the efforts by the STS-CHSD to standardize nomenclature internationally (see “[STS Congenital Heart Surgery Database \[STS-CHSD\]](#)” section above), the STS-GTSD continues to update its data specifications and harmonize data definitions with the European Society of Thoracic Surgeons. This work will facilitate joint research and quality improvement initiatives, as well as international comparisons of care.

Members of the STS General Thoracic Surgery Database Task Force are exploring options for obtaining long-term outcomes for cancer resection, including linking the STS-GTSD with Medicare data (see “[STS Task Force on Longitudinal Follow-Up and Linked Registries \[LFLR\]](#)” section below). However, other data sources will also be required, including various cancer registries, as 40 % of lung cancer resections and 60 % of esophageal cancer resections are under the age of 65. (Medicare data only includes patients 65 or older and younger patients on dialysis.)

---

## Database Operations

### Data Sources

Although many investigators use claims data (e.g., Medicare) for performance evaluation and research, the distinguishing feature of the STS National Database and similar clinical registries

is the degree of granularity and specificity of its data elements (STS National Database 2014). Since the inception of the STS National Database, periodic (typically every 3 years, in a cycle that allows one of each of the three databases to be updated each year) data specification upgrades occur based on the evolution of scientific knowledge as well as feedback from database managers, end users, and participants. Every data element collected has an associated sequence number which is mapped to a detailed clinical data specification. This feature of clinical registries – their highly structured and clinical granular data – distinguishes them from alternative data sources such as claims data (not clinically rich) and electronic health record (EHR) data (unstructured, lacking specific definitions used by all institutions).

This unique advantage of clinical registries, including the STS National Database, also poses one of their greatest challenges – data collection burden. Rather than allowing anyone to enter the data that become part of a patient’s STS record, these data are either entered by a trained abstractor, or data entered by caregivers are carefully reviewed by the data abstractor. These data managers work with surgeons, physician assistants, nurse practitioners, and others to ensure that that data entered into the STS National Database adhere to the definitions established by STS and that they are supported by documentation in the patient’s medical record. These data managers have many resources available to them including:

- The detailed written specifications themselves.
- A teaching manual that expands upon the formal specifications and often includes clinical examples
- Advice of colleagues in regional collaboratives around the nation
- Biweekly telephone calls with STS National Database and Duke Clinical Research Institute leaders
- Email alerts
- Newsletters
- A 4-day annual national meeting (The Society of Thoracic Surgeons Advances in Quality and Outcomes [AQO] Conference: A Data

Managers Meeting) attended by nearly 500 data managers from around the country (at which data managers and surgeon leaders present educational sessions on challenging coding issues and new developments in data specifications)

Numerous studies have been conducted (Shahian et al. 2007b; Mack et al. 2005) showing that both the number and type of procedures performed and their results differ substantially with the use of detailed clinical data as opposed to claims data sources.

STS is working with EHR vendors to investigate how some STS variables might be automatically extracted from routinely collected EHR data. The most straightforward variables for this type of capture would include demographics, labs, and structured diagnostic testing such as percent coronary artery obstruction, ejection fraction, and valve areas. Other STS data elements which have complex data specifications would be more challenging to map from EHRs, and these complex elements might require the addition of specific fields to the EHR.

## Vendors

The Society of Thoracic Surgeons has contractual relationships with a number of vendors who provide the data entry software by which participant programs enter data into the STS National Database. Each vendor differs in the sophistication of the reports they produce, opportunity for customization, cost, and ability to link with other databases such as the American College of Cardiology (ACC) National Cardiovascular Data Registry (NCDR). However, each vendor must achieve basic certification by STS to ensure that their software is capable of producing accurate and consistent results.

## STS Staff

Numerous full-time staff at STS headquarters are devoted to database operations and serve multiple functions:

- Assist programs in joining the database.
- Develop and maintain appropriate contractual relationships with vendors, participants, and our warehouse and analytic center.
- Coordinate and staff the various STS National Database Task Forces and their respective conference calls and meetings.
- Develop and maintain budgets.
- Assure compliance with all relevant regulatory processes, including the Health Insurance Portability and Accountability Act of 1996 (HIPAA).
- Serve as the main resource for data managers.
- Arrange the annual STS Advances in Quality and Outcomes [AQO] Conference.
- Work with external organizational partners on issues such as public reporting.
- Coordinate the ongoing upgrades of all three clinical databases.

## Data Warehouse and Analytic Center

Since 1998, the Duke Clinical Research Institute (DCRI) has served as the data warehouse and analytic center for the STS National Database. DCRI receives data from participants, which then undergo extensive data quality and consistency checks. Each participant receives a comprehensive harvest feedback report generated by DCRI, as previously described. These feedback reports are distributed every 3 months to participants in the STS-ACSD and every 6 months to participants in the STS-CHSD and the STS-GTSD. These feedback reports include extensive educational and explanatory materials describing how each report and metric are calculated. DCRI also provides statistical support for most of the STS National Database Task Forces, particularly the Quality Measurement Task Force, and they are also involved in the Access and Publications Task Force, the STS Task Force on Longitudinal Follow-Up and Linked Registries (LFLR), and the STS Research Center. DCRI statisticians play an integral role in the design and implementation of all STS risk models and performance measures.

## Data Quality and Audit

Regardless of the granularity and specificity of the data elements in any registry, they are only useful if data are actually inputted in strict conformity with their specifications. A firm belief in the accuracy of data submitted by all programs nationally, and the metrics derived from them, provides the foundation of trust necessary to implement STS programs such as voluntary public reporting.

Data quality checks exist at several stages of the STS data entry process. First, there are internal consistency and out-of-range audits that take place at the time of data entry. For example, an age of 150 years would be rejected because it falls out of the acceptable data input range. Second, submitted data are reviewed at DCRI, and excessive rates of missing data or other irregularities not captured during data submission are reported back to STS participant for revision. Third, STS participant sites receive a list of their demographics, risk factors, operative data, and outcomes compared to STS nationally and to hospitals of similar size and academic status. Substantial differences from these benchmarks would lead a program to evaluate the accuracy of its submissions.

Finally, STS has an extremely robust annual audit of all three of its databases, all conducted by a highly respected external organization. Ten percent of all STS National Database sites are randomly selected for audit annually. Each audit consists of 20 coronary bypass procedures and 10 valve procedures; approximately 82 data elements are abstracted from each medical record. Previously this process had required on-site visits by the external auditing agency, but a mechanism has been developed to access patient records electronically in a HIPAA-compliant fashion. In addition to validating STS submissions against the medical record (for accuracy of the data), STS submissions are also checked against hospital operative logs in order to ensure that all cases have been collected (for completeness of the data).

Each year, all three clinical databases comprising the STS National Database are audited. An extensive report is generated showing the agreement rate for all audited data elements and an

overall assessment of the accuracy at audit sites. In 2013, among nearly 100,000 individual data elements audited, the overall agreement rates in the STS-ACSD averaged nearly 97%. As described above, similar agreement rates are documented in the STS-CHSD and the STS-GTSD. In the STS-CHSD, an STS congenital heart surgeon volunteer leader also participates in each audit.

## STS Quality Measurement Task Force (STS-QMTF)

The STS Quality Measurement Task Force (STS-QMTF) is responsible for all risk model and performance measure development for the Society. These quality measurement activities are fully integrated into the STS National Database, a unique arrangement that has numerous advantages. First, the performance measures are based on readily available STS clinical data. Second, the performance measures are developed through direct collaboration between statistical consultants and surgeons who have both clinical expertise and knowledge of performance measurement and health policy. Third, the performance measures can be tested for reliability and validity by using them in confidential participant feedback reports prior to public reporting. Pilot testing is a difficult process for many measure developers, but it is an inherent capability provided by a clinical registry such as the STS National Database.

In addition to having the best available clinical data, the next most important factor in performance measure development is risk models. These are essential to adjust for inherent differences in patient risk, and they are crucial if performance measures are to have face validity with stakeholder groups, especially the providers (Shahian et al. 2013a). Risk model development typically begins by identifying the most relevant outcomes for a particular type of procedure and specialty. Initial exploratory analyses are performed to determine if an adequate number of cases and endpoints are available and over what period of time these need to be aggregated in order to assure adequate sample size for the outcome in question.



The selection and definition of relevant endpoints is critical to the development of risk models. In both quality assessment activities and clinical research to improve patient care, STS has defined its major outcomes endpoint, mortality, in a unique fashion. Typically, mortality after hospitalizations or procedures has used one of two definitions. In-hospital mortality is collected with high accuracy, but it misses early post-discharge deaths occurring at home or in extended care facilities. Collecting only in-hospital outcomes may also create a perverse incentive to discharge patients earlier than desirable so that potential adverse outcomes do not occur during the index hospitalization. Another approach is to measure adverse outcomes at 30 days, regardless of where the patient is located. This avoids providing an incentive for premature discharge, but it may encourage some providers to keep a severely ill patient alive through artificial support just long enough to meet the 30-day threshold. STS seeks to avoid the disadvantages of either of these approaches alone by combining them. The time period of mortality data collection for all three STS National Databases is based upon the STS definition of operative mortality (Overman et al. 2013), which is now used by all three STS National Databases: operative mortality is defined as (1) all deaths, regardless of cause, occurring during the hospitalization in which the operation was performed, even if after 30 days (including patients transferred to other acute care facilities), and (2) all deaths, regardless of cause, occurring after discharge from the hospital, but before the end of the 30th postoperative day.

As the next step in risk model development, bivariate analyses are performed to study the association between individual risk factors and the outcome. A comprehensive array of candidate risk factors is entered into multivariable risk models, and odds ratios (with 95 % CI) are determined for each. In some instances, certain variables are “forced” into the model regardless of statistical significance because they are regarded by clinical experts as critical for face validity. The output of these models is assessed using measures of calibration, discrimination, and reliability and using actual data from the STS National Database.

After endorsement by the Executive Committee of STS, all STS performance measures are published in their entirety in the peer-reviewed literature (Shahian et al. 2009a, b; O’Brien et al. 2009b), including all special considerations discussed during the measurement development process, the final covariates and their parameterization, and the associated intercepts and coefficients of the risk model equations.

Risk-adjusted outcomes based on national benchmark STS data are provided back to participants at each quarterly harvest. Risk models are fully updated every few years, but annually a calibration factor is introduced so that the observed-to-expected ratio for a given year equals one. Multiple STS risk models are publicly available as online calculators on the STS website (STS short-term risk calculator 2014; STS long-term risk calculator 2014), and these sites are visited thousands of times each month.

The appropriate interpretation of risk-adjusted results bears special mention, given both its centrality in performance measurement and the fact that it is often misunderstood by many who view these reports. There are two primary statistical methods by which outcomes results are adjusted for inherent risk (Shahian and Normand 2008). In direct standardization, the stratum-specific rates (e.g., age, sex, ethnicity) for each population of interest (e.g., a particular hospital’s stratum-specific rate of adverse events) are applied to a standard or reference population. This method is often used in epidemiology where there are a limited number of strata to be considered, and the rates for each stratum are available. However, for most provider profiling applications, the number of strata, corresponding to individual risk factors, is too large to standardize in this fashion. Accordingly, almost all healthcare profiling initiatives use another statistical method, indirect standardization, for risk adjustment. In this approach, the rates derived from a reference or standard population of hospitals, often in the form of a risk model with intercepts and coefficients, are applied to the particular case mix of the institutions being studied. The actual results for an individual program’s case mix are compared to what would have been expected had that program’s

population of patients been treated by an average provider from the reference population.

Both methods of standardization provide risk adjustment in a generic sense – they “level the playing field” – so that programs caring for sicker patients are not penalized. However, only direct standardization permits *direct comparison* of the risk-standardized results of one specific program with those of another. In indirect standardization, the results for any particular program are based solely on its specific mix of patients, and these results can only be compared with the overall results of all providers for a similar case mix (Shahian and Normand 2008). For example, a small community heart surgery program may have a lower risk-adjusted mortality rate than a tertiary/quaternary center. However, using indirect standardization, it cannot be assumed that if faced with the same case mix of the tertiary center, it would also have superior results.

The primary motivation for development of the STS National Database was the need to provide accurate performance assessment, and this remains the highest priority of the STS-QMTF. A variety of measures have been developed including structure, process, and outcomes (the Donabedian triad) (Donabedian 1966). Risk-adjusted mortality rates for CABG were the original outcome used to classify cardiac surgery performance, but even this archetypal measure can be inadequate. For example, consider three survivors of coronary artery bypass surgery (CABG), all of whom would be considered to have had identical quality procedures based on mortality alone. One patient receives all the appropriate bypass grafts and medications and sustains no complications. The second patient receives only vein grafts, which have limited longevity, and does not receive postoperative medications to prevent progression of coronary disease. The third patient experiences the new onset of dialysis-dependent renal failure which will markedly impact both longevity and quality of life. Despite having all survived surgery, the quality received by these three patients varied markedly.

The STS-QMTF has recognized the inadequacy of using CABG risk-adjusted mortality as the sole quality metric for cardiac surgery, and it

has addressed this in a number of ways. First, it has expanded its activities in risk modeling and performance metrics beyond CABG to include other major cardiothoracic procedures such as isolated aortic valve replacement, aortic valve replacement combined with CABG, mitral valve replacement, mitral valve repair, multiple valve procedures, and numerous procedures in general thoracic surgery and congenital cardiac surgery. This expansion of the procedures that are available for risk modeling and performance assessment provides a much more comprehensive assessment of quality than focusing solely on CABG, whose incidence and rate of adverse outcomes have both been declining over the past decade. Second, instead of collecting information only on mortality, the STS-QMTF has developed risk models for more of the individual surgical complications such as stroke, reoperation, prolonged ventilation, infection, renal failure, prolonged length of stay, and a composite of major morbidity and mortality.

Third, in addition to viewing these measures individually, STS has increasingly focused on composite measures using multivariate hierarchical approaches. The first STS composite measure, CABG, included the risk-adjusted mortality, the occurrence of any (any or none) of the five major complications of CABG surgery (stroke, renal failure, prolonged ventilation, reoperation, and infection), the use of at least one internal mammary artery graft, and the provision of all four (all or none) NQF-endorsed medications (preoperative beta blockade, discharge beta blockade, lipid-lowering agents such as statins and aspirin) (Shahian et al. 2007a; O'Brien et al. 2007). Similar composite measures have been developed for isolated aortic valve replacement (Shahian et al. 2012a) and for aortic valve replacement combined with CABG (Shahian et al. 2014), and a composite measure is currently under development for mitral valve surgery. These latter measures differ from the isolated CABG composite in that they consist solely of outcomes measures (mortality and morbidity) and do not include process measures. This reflects both a shift in healthcare performance measurement toward outcomes measures (rather than structure or process

measures) and the fact that evidence-based, widely accepted process measures suitable for performance measurement are not available for these other procedures.

STS envisions a portfolio of such procedure-specific composite measures and, ultimately, an overall composite of procedural performance encompassing information from all these individual composite metrics (a “composite of composites”). However, even this “composite of composites” will only be one component of an overall STS performance measurement system that will include multiple other domains. For example, just as important as the outcome of particular procedure is the question of whether that procedure was indicated in the first place. Accordingly, STS has mapped both the ACCF/AHA CABG guidelines (Hillis et al. 2011) and the multi-societal 2012 Appropriate Use Criteria (AUC) for Coronary Revascularization (Patel et al. 2012) to the relevant data elements in the STS-ACSD. This will ultimately allow STS participants to receive immediate documentation that their patient meets one of these CABG guidelines or AUC. Similar mapping is underway for valve procedures. STS has also begun to explore failure to rescue (mortality following the development of a complication of surgery) as an additional new quality metric (Pasquali et al. 2012a). Previous research suggests that the ability to salvage a patient from a serious complication is a distinguishing feature of high-quality programs and complements other metrics such as overall morbidity. Patient-reported outcomes are also increasingly recognized for their value in assessing quality. These include both patient-reported functional outcomes (e.g., return to work and overall functional capacity) as well as patient satisfaction (e.g., HCAHPS or Hospital Consumer Assessment of Healthcare Providers and Systems, CGCAHPS or Clinician and Group Consumer Assessment of Healthcare Providers). STS has also formed a Cost and Resource Task Force within the STS National Database. The objective of this task force is to link the STS National Database with cost data from hospital, commercial, federal, or state payer data. Such a linkage would provide accurate data regarding

variability in resource use among programs, as well as the development of risk models for cost, so that programs being evaluated for cost efficiency are not unfairly penalized when they care for particularly complex patients. STS ultimately envisages a comprehensive portfolio of performance measures which might include a composite of multiple procedural composite measures, appropriateness, failure to rescue, patient-centered outcomes, and risk-adjusted resource utilization.

Finally, the most appropriate level of attribution for performance measures is a focus of continuing discussion. STS has historically measured performance only at the participant level (typically a hospital) for a variety of reasons. There are sample size concerns at the individual surgeon level, and cardiac surgery is a “team sport” requiring many participants in addition to the surgeon (e.g., cardiologist, anesthesiologist, perfusionist, nurses, critical care specialists, respiratory therapists). However, notwithstanding these concerns, many commercial payers and governmental agencies are now publishing (or requiring) information about surgeon-level performance, much of which are based on inadequately adjusted administrative claims data and/or flawed analytics. Consequently, STS feels a responsibility to offer a valid, surgeon-level metric. An individual surgeon performance metric has now been developed by STS for adult cardiac surgery. It is a composite measure based on morbidity and mortality data for 5 of the most common performed procedures, aggregated over 3 years. This measure has very high reliability (0.81) because of the large number of endpoints being analyzed (Shahian et al. 2015).

Regardless of the particular performance measure, the general STS-QMTF approach to profiling performance results across providers is similar. Results are estimated in Bayesian hierarchical models, and providers are classified as having expected, better than expected, or worse than expected performance based on true Bayesian probabilities rather than frequentist confidence intervals (Shahian et al. 2007a; O’Brien et al. 2007). Unlike the latter, the Bayesian credible interval has an intuitive probability interpretation. For example, given a database participant’s

observed data, if the lower limit of the 98 % Bayesian credible interval is greater than the STS average value, then there is at least 99 % probability (98 % credible interval plus 1 % upper tail) that the participant's true performance (e.g., in avoiding mortality or morbidity or in using an internal mammary artery graft) exceeds the STS average value for their particular case mix. The Bayesian probability (and corresponding Bayesian credible interval) selected for a particular measure varies depending on factors such as event rates, variation of scores across programs, and sample sizes for typical providers. For procedures such as CABG which are frequently performed, STS has used 99 % Bayesian probabilities, which result in approximately 10–15 % of STS providers being labeled as low performing and 10–15 % classified as high performing, with the remainder being average. For less common procedures such as isolated valve replacement, STS-QMTF has used 95 % Bayesian probabilities (97.5 % credible intervals), which results in fewer outliers (Shahian et al. 2012a). Even with the lower probability requirement, the smaller number of observations means there is less data upon which to base an estimate of a provider's performance, and the percentage of outliers is typically lower than for CABG. If the probability criterion were even lower (e.g., 90 % Bayesian probability), then more participants would be classified as outliers, but our certainty would also be much lower, jeopardizing face validity with providers and other stakeholders.

Importantly, when estimated in this fashion, there is no requirement for any fixed number of high or low outliers. If, for example, all programs function at a high level and were statistically indistinguishable using these criteria, they would all be average (or, in STS parlance, two-star) programs. In contrast to payers and commercial report card developers, who often seem determined to demonstrate differences among providers, STS believes the ideal situation from a societal perspective would be for all programs to be functioning at a very high level and statistically indistinguishable (e.g., the very high safety record of the commercial aircraft industry). Then,

consumers could choose surgeons or hospitals based on other criteria, such as convenience, availability, or service.

In reporting their results, STS provides varying levels of granularity. These range from point estimates with credible intervals for statistically sophisticated users and star ratings corresponding to as expected, better than expected, or worse than expected for typical consumers (based on the work of Professor Judith Hibbard (Hibbard et al. 2001)). When a composite measure encompasses multiple procedures or performance domains, STS always provides the ability to drill down to the lowest level of the composite, its constituent elements.

### **STS Quality Initiatives Task Force (STS-QIT)**

The acquisition of healthcare data and their use in performance assessment are not goals in themselves. The primary objective of all these activities is to improve healthcare quality. Just as the Quality Measurement Task Force is an integral part of the STS National Database, the STS Quality Initiatives Task Force (STS-QIT) is similarly fully integrated. This facilitates the use of STS data as the basis for quality improvement projects and allows both baseline and subsequent performance to be measured, thus documenting the effectiveness of interventions. Another advantage of integrating the Quality Initiatives Task Force within the database is to facilitate the identification of gaps and variability in national performance and to focus quality initiatives in these areas.

At the national level, quality improvement initiatives have been conducted using the STS National Database to improve compliance with preoperative beta blockade and use of internal mammary artery bypass grafts for CABG, both of which are NQF-endorsed performance measures (Ferguson et al. 2003). A 2012 report by ElBardissi and colleagues (ElBardissi et al. 2012) suggests that the STS National Database and its quality measurement and improvement activities have dramatically improved cardiac surgery results over the past decade.

STS-QIT has begun to identify key opportunities for improvement within cardiothoracic surgery and has developed focused webinars and online libraries of best practice articles to address these issues. Specific recent webinars (STS Quality Improvement webinars 2014) include blood conservation and transfusion triggers, glucose management, and mediastinal staging prior to lung cancer surgery. The Quality Initiatives Task Force is also exploring the possibility of identifying consistently low-performing programs using STS data and then offering such programs the possibility of external review of their database integrity (to identify potential coding issues that might lead to false outlier classification) and clinical practice (to facilitate quality improvement).

A number of states and regions have also used STS data to improve quality. For example, in a collaborative effort with Blue Cross Blue Shield of Michigan, the Michigan Society of Thoracic and Cardiovascular Surgeons has brought together representatives from all cardiac surgery programs in the state (Prager et al. 2009). They review performance of all programs, identify gaps and variability in outcomes, and review each cardiac surgery death using a standardized phase of care mortality analysis (POCMA). They have also implemented a number of best practice initiatives. Similarly, the Virginia Cardiac Surgery Quality Initiative (Speir et al. 2009) has brought together surgeons from across the state. They have linked STS clinical data to cost data with a focus on reducing both complications and their associated costs.

### STS Public Reporting Task Force

Among healthcare professional societies, STS has taken the lead in public reporting by providing easily understandable cardiothoracic surgical outcomes data to the public (Shahian et al. 2011a, b). STS support of public reporting and transparency is based on several principles:

- Public reporting and accountability are our professional responsibilities.
- Patients and their families have a right to know the outcomes of cardiothoracic surgical procedures.

- Public reporting demonstrates commitment to quality improvement.
- Public reporting is one approach to improving quality.
- Public reporting promotes patient autonomy and facilitates shared decision-making.
- If professional medical and surgical societies do not publish accurate information about performance using the best available clinical data and risk adjustment, then the public will be forced to judge our performance based on unadjusted or inadequately adjusted administrative claims data.

The STS Public Reporting Task Force is responsible for the development and maintenance of the web-based platforms for public reporting of data from the STS National Database. STS has implemented voluntary public reporting through its STS Public Reporting Online Initiative [[www.sts.org/publicreporting](http://www.sts.org/publicreporting)] and through collaboration with *Consumers Union* [[www.consumerreports.org/health](http://www.consumerreports.org/health)]. In each case, these reports are based on the STS composite measures and star ratings (with drill-down capability) described above.

In September 2010, STS began publicly reporting outcomes of isolated CABG surgery based on its NQF-endorsed composite CABG metric. In January 2013, STS began publicly reporting outcomes of isolated aortic valve replacement (AVR) surgery based on its NQF-endorsed AVR composite score. In August 2014, STS began publicly reporting outcomes of combined AVR + CABG surgery, using an NQF-endorsed composite score with the same two domains (risk-adjusted morbidity and mortality) as the isolated AVR composite.

STS plans to expand its portfolio of publicly reported cardiothoracic surgical quality measures by at least one additional new operation every year. Future publicly reported metrics will include pediatric and congenital heart surgery risk-adjusted operative mortality based on the 2014 STS Congenital Heart Surgery Database Mortality Risk Model (planned for public reporting in January 2015), mitral valve replacement (MVR) and mitral valve repair, a multi-domain composite

for pulmonary lobectomy for cancer, and a multi-domain composite for esophagectomy. As of mid-2016, 50 % of adult cardiac surgery participants in the STS National Database and 60 % of congenital heart surgery participants had consented to voluntary public reporting.

### STS Research Center

The initial and still primary purpose of the STS National Database is quality assessment and quality improvement in cardiothoracic surgery. The STS National Database and its quality assessment activities, development of nationally recognized quality measures, and performance improvement initiatives are all built on the foundation of more than five million surgical records (STS National Database 2014; Shahian et al. 2013b). Because it is such a robust source of clinical data, the STS National Database also provides a platform for research to advance the state of the art of cardiothoracic surgery. This research activity is overseen by the STS Research Center (2014).

Launched in 2011, the STS Research Center is a nationally recognized leader in outcomes research. The center seeks to capitalize on the value of the STS National Database and other resources to provide scientific evidence and support cutting-edge research. Such research ultimately helps cardiothoracic surgeons, government, industry, and other interested parties to improve surgical care and outcomes.

All research that is confined to the STS National Database and to its standard period of data collection (the index operative hospitalization and 30 days postoperatively) is vetted through the STS Access and Publications (A and P) Task Force. Research that involves linking the STS National Database to other databases, or longitudinal follow-up beyond the standard period of data collection of the STS National Database, is vetted by the STS Task Force on Longitudinal Follow-Up and Linked Registries (STS-LFLR) (see “[STS-LFLR](#)” section below). Using this process, research activities based on data from the STS National Database have resulted in more than 300 peer-reviewed publications in the scientific

literature and have significantly advanced knowledge in cardiothoracic surgery.

### STS Task Force on Longitudinal Follow-Up and Linked Registries (STS-LFLR)

The STS Task Force on Longitudinal Follow-Up and Linked Registries (STS-LFLR) is responsible for oversight of research initiatives that involve longitudinal follow-up of patients and linking of the STS National Database to other sources of data. The transformation of the STS National Database to a platform for longitudinal follow-up will ultimately result in higher quality of care for all cardiothoracic surgical patients by facilitating capture of long-term clinical and nonclinical outcomes on a national level. Important strategies include the development of clinical longitudinal follow-up modules within the STS National Database itself and linking the STS National Database to other clinical registries, administrative databases, and national death registries:

1. Using probabilistic matching with shared indirect identifiers, the STS National Database can be linked to administrative claims databases (such as the CMS Medicare Database (Jacobs et al. 2010; Hammill et al. 2009) and the Pediatric Health Information System (PHIS) database (Pasquali et al. 2010, 2012)) and become a valuable source of information about long-term mortality, rates of re-hospitalization, morbidity, and cost (Shahian et al. 2012b; Weintraub et al. 2012; Pasquali et al. 2012b).
2. Using deterministic matching with shared unique direct identifiers, the STS National Database can be linked to national death registries like the Social Security Death Master File (SSDMF) and the National Death Index (NDI) in order to verify life status over time (Jacobs et al. 2011b).

Through either probabilistic matching or deterministic matching, the STS National Database can link to multiple other clinical registries, such as the ACC NCDR, and to claims data sources, in order to provide enhanced clinical follow-up and

opportunities for comparative effectiveness research. The NIH-funded ASCERT trial (American College of Cardiology Foundation–Society of Thoracic Surgeons Collaboration on the Comparative Effectiveness of Revascularization Strategies trial) exemplifies this approach. ASCERT linked STS and ACC clinical registry data with Medicare data to compare longer-term outcomes for surgical and percutaneous approaches to coronary revascularization (Weintraub et al. 2012). Similarly, the NIH-funded linkage of the STS-CHSD to the Pediatric Health Information System (PHIS) Database used linked clinical and administrative data to facilitate comparative effectiveness research in the domains of perioperative methylprednisolone and outcome in neonates undergoing heart surgery (Pasquali et al. 2012c) and antifibrinolytic medications in pediatric heart surgery (Pasquali et al. 2012d).

## Device Surveillance

Another role of the STS National Database is the longitudinal surveillance of implanted medical devices. The use of the STS National Database as a platform for device surveillance is best exemplified by the STS/ACC Transcatheter Valve Therapies (TVT) Registry (Carroll et al. 2013; Mack et al. 2013), which tracks patients who undergo Transcatheter Aortic Valve Replacement (TAVR). Since December 2011, the TVT Registry has collected data for all commercial TAVR procedures performed in the USA. As of mid-2016, it had 457 enrolled sites and had acquired 74,240 patient records (personal communication, Joan Michaels).

The TVT Registry was launched as a joint initiative of STS and ACC in collaboration with CMS, the US Food and Drug Administration (FDA), and the medical device industry. It serves as an objective, comprehensive, and scientifically rigorous resource to improve the quality of patient care, to monitor the safety and effectiveness of TVT devices through post-market surveillance, to provide an analytic resource for TVT research, and to enhance communication among key stakeholders.

## Summary

The STS National Database, comprised of three specialty-specific registries, is the premier clinical data registry for cardiothoracic surgery. In comparison with other available data sources, the STS National Database and similar clinical registries have the advantages of structured, granular data elements defined by clinical experts, standardized data specifications, high accuracy as confirmed by external audit, and the capability to provide more robust risk adjustment.

Clinical registries like the STS National Database are the best sources for measuring healthcare outcomes. In contrast to many claims data sources, the STS National Database provides “real-world” data from all age groups and payers. Furthermore, as described in this chapter, the ability to accurately measure clinical outcomes requires standardized clinical nomenclature, uniform standards for defining and collecting data elements, strategies to adjust for the complexity of patients, and techniques to verify the completeness and accuracy of data. All of these elements exist in clinical registries such as the STS National Database. Consequently, metrics derived from clinical registries are ideally suited for high-stakes applications such as public reporting, center of excellence designation, and reimbursement. STS performance measures based on the STS National Database have been used to develop more than 30 measures endorsed by the National Quality Forum.

Clinical registries can be linked to other data sources to obtain information about long-term outcomes and risk-adjusted cost and resource utilization, all increasingly important considerations in healthcare. Clinical registries are also used to satisfy regulatory and governmental requirements, as exemplified by Qualified Clinical Data Registries in the CMS Physician Quality Reporting System (PQRS) program, and the use of registries for post-market surveillance of new implantable devices, in collaboration with CMS and FDA, as exemplified by the Transcatheter Valve Therapies (TVT) Registry.

Clinical registries are the ideal platform for developing evidence for best practice guidelines and to document appropriateness of procedures.

They are also invaluable for comparative effectiveness research. Although randomized trials have been considered by many to be the gold standard of comparative effectiveness research, recent efforts have examined the possibility of using clinical registries as platforms for randomized trials (Frobert et al. 2013; Lauer and D'Agostino 2013). Performing randomized trials within clinical registries would potentially accomplish the dual objectives of decreasing the cost of these trials and increasing the generalizability of the results (as the included patients are more representative of “real-world” populations).

Clinical registries provide practitioners with accurate and timely feedback of their own outcomes and can benchmark these outcomes to regional, national, or even international aggregate data, thus facilitating quality improvement.

The STS National Database exemplifies that potential value of clinical registries for all of healthcare. High-quality data are collected once and then used for multiple purposes, with the ultimate goal of improving the care of all patients.

## References

- Carroll JD, Edwards FH, Marinac-Dabic D, et al. The STS-ACC transcatheter valve therapy national registry: a new partnership and infrastructure for the introduction and surveillance of medical devices and therapies. *J Am Coll Cardiol*. 2013;62(11):1026–34.
- Clark RE. It is time for a national cardiothoracic surgical data base. *Ann Thorac Surg*. 1989;48(6):755–6.
- Codman EA. The product of a hospital. *Surg Gynecol Obstet*. 1914;18:491–6.
- Codman EA. A study in hospital efficiency. As demonstrated by the case report of the first two years of a private hospital. Reprint edition (originally published privately 1914–1920) ed. Oakbrook Terrace: Joint Commission on Accreditation of Healthcare Organizations; 1995.
- Cronk CE, Malloy ME, Pelech AN, et al. Completeness of state administrative databases for surveillance of congenital heart disease. *Birth Defects Res A Clin Mol Teratol*. 2003;67(9):597–603.
- Donabedian A. Evaluating the quality of medical care. *Milbank Mem Fund Q*. 1966;44(3):166–206.
- Donabedian A. The quality of care. How can it be assessed? *JAMA*. 1988;260(12):1743–8.
- Donabedian A. The end results of health care: Ernest Codman's contribution to quality assessment and beyond. *Milbank Q*. 1989;67(2):233–56.
- ElBardissi AW, Aranki SF, Sheng S, O'Brien SM, Greenberg CC, Gammie JS. Trends in isolated coronary artery bypass grafting: an analysis of the Society of Thoracic Surgeons adult cardiac surgery database. *J Thorac Cardiovasc Surg*. 2012;143(2):273–81.
- Ferguson Jr TB, Peterson ED, Coombs LP, et al. Use of continuous quality improvement to increase use of process measures in patients undergoing coronary artery bypass graft surgery: a randomized controlled trial. *JAMA*. 2003;290(1):49–56.
- Franklin RC, Jacobs JP, Krogmann ON, et al. Nomenclature for congenital and paediatric cardiac disease: historical perspectives and The International Pediatric and Congenital Cardiac Code. *Cardiol Young*. 2008;18 Suppl 2:70–80.
- Frobert O, Lagerqvist B, Olivecrona GK, et al. Thrombus aspiration during ST-segment elevation myocardial infarction. *N Engl J Med*. 2013;369(17):1587–97.
- Frohnert BK, Lussky RC, Alms MA, Mendelsohn NJ, Symonik DM, Falken MC. Validity of hospital discharge data for identifying infants with cardiac defects. *J Perinatol*. 2005;25(11):737–42.
- Grover FL, Shahian DM, Clark RE, Edwards FH. The STS National Database. *Ann Thorac Surg*. 2014;97 Suppl 1: S48–54.
- Hammill BG, Hernandez AF, Peterson ED, Fonarow GC, Schulman KA, Curtis LH. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. *Am Heart J*. 2009;157(6): 995–1000.
- Hannan EL, Cozzens K, King III SB, Walford G, Shah NR. The New York State cardiac registries: history, contributions, limitations, and lessons for future efforts to assess and publicly report healthcare outcomes. *J Am Coll Cardiol*. 2012;59(25):2309–16.
- Hibbard JH, Peters E, Slovic P, Finucane ML, Tusler M. Making health care quality reports easier to use. *Jt Comm J Qual Improv*. 2001;27(11):591–604.
- Hillis LD, Smith PK, Anderson JL, et al. ACCF/AHA guideline for coronary artery bypass graft surgery: executive summary: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *Circulation*. 2011;124(23):2610–42.
- Iezzoni LI. Risk adjustment for measuring health care outcomes. 3rd ed. Chicago: Health Administration Press; 2003.
- Jacobs JP, Jacobs ML, Mavroudis C, et al. Nomenclature and databases for the surgical treatment of congenital cardiac disease—an updated primer and an analysis of opportunities for improvement. *Cardiol Young*. 2008;18 Suppl 2:38–62.
- Jacobs JP, Jacobs ML, Lacour-Gayet FG, et al. Stratification of complexity improves the utility and accuracy of outcomes analysis in a multi-institutional congenital heart surgery database: application of the risk adjustment in congenital heart surgery (RACHS-1) and Aristotle systems in the Society of



- Thoracic Surgeons (STS) Congenital Heart Surgery Database. *Pediatr Cardiol.* 2009;30(8):1117–30.
- Jacobs JP, Edwards FH, Shahian DM, et al. Successful linking of the Society of Thoracic Surgeons adult cardiac surgery database to Centers for Medicare and Medicaid Services Medicare data. *Ann Thorac Surg.* 2010;90(4):1150–6.
- Jacobs ML, Daniel M, Mavroudis C, et al. Report of the 2010 Society of Thoracic Surgeons congenital heart surgery practice and manpower survey. *Ann Thorac Surg.* 2011a;92(2):762–8.
- Jacobs JP, Edwards FH, Shahian DM, et al. Successful linking of the Society of Thoracic Surgeons database to social security data to examine survival after cardiac operations. *Ann Thorac Surg.* 2011b;92(1):32–7.
- Jacobs JP, Shahian DM, He X, et al. Penetration, completeness, and representativeness of the Society of Thoracic Surgeons adult cardiac surgery database. *Ann Thorac Surg.* 2016;101(1):33–41.
- Jantzen DW, He X, Jacobs JP, et al. The impact of differential case ascertainment in clinical registry versus administrative data on assessment of resource utilization in pediatric heart surgery. *World J Pediatr Congenit Heart Surg.* 2014;5(3):398–405.
- Kouchoukos NT, Ebert PA, Grover FL, Lindesmith GG. Report of the Ad Hoc Committee on risk factors for coronary artery bypass surgery. *Ann Thorac Surg.* 1988;45(3):348–9.
- Kozower BD, Sheng S, O'Brien SM, et al. STS database risk models: predictors of mortality and major morbidity for lung cancer resection. *Ann Thorac Surg.* 2010;90(3):875–81.
- Lauer MS, D'Agostino Sr RB. The randomized registry trial—the next disruptive technology in clinical research? *N Engl J Med.* 2013;369(17):1579–81.
- Mack MJ, Herbert M, Prince S, Dewey TM, Magee MJ, Edgerton JR. Does reporting of coronary artery bypass grafting from administrative databases accurately reflect actual clinical outcomes? *J Thorac Cardiovasc Surg.* 2005;129(6):1309–17.
- Mack MJ, Brennan JM, Brindis R, et al. Outcomes following transcatheter aortic valve replacement in the United States. *JAMA.* 2013;310(19):2069–77.
- Mallon WJ. Ernest Amory Codman: the end result of a life in medicine. Philadelphia: W.B.Saunders Company; 2000.
- Mavroudis C, Jacobs JP. Congenital heart surgery nomenclature and database project: overview and minimum dataset. *Ann Thorac Surg.* 2000;69(3, Suppl 1):S1–17.
- Neuhauser D. Ernest Amory Codman, M.D., and end results of medical care. *Int J Technol Assess Health Care.* 1990;6(2):307–25.
- O'Brien SM, Shahian DM, DeLong ER, et al. Quality measurement in adult cardiac surgery: part 2—Statistical considerations in composite measure scoring and provider rating. *Ann Thorac Surg.* 2007;83 Suppl 4:S13–26.
- O'Brien SM, Clarke DR, Jacobs JP, et al. An empirically based tool for analyzing mortality associated with congenital heart surgery. *J Thorac Cardiovasc Surg.* 2009a;138(5):1139–53.
- O'Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. *Ann Thorac Surg.* 2009b;88 Suppl 1:S23–42.
- O'Connor GT, Plume SK, Olmstead EM, et al. A regional prospective study of in-hospital mortality associated with coronary artery bypass grafting. The Northern New England Cardiovascular Disease Study Group. *JAMA.* 1991;266(6):803–9.
- Overman DM, Jacobs JP, Prager RL, et al. Report from the Society of Thoracic Surgeons National Database Workforce: clarifying the definition of operative mortality. *World J Pediatr Congenit Heart Surg.* 2013;4(1):10–2.
- Pasquali SK, Jacobs JP, Shook GJ, et al. Linking clinical registry data with administrative data using indirect identifiers: implementation and validation in the congenital heart surgery population. *Am Heart J.* 2010;160(6):1099–104.
- Pasquali SK, Li JS, Jacobs ML, Shah SS, Jacobs JP. Opportunities and challenges in linking information across databases in pediatric cardiovascular medicine. *Prog Pediatr Cardiol.* 2012a;33(1):21–4.
- Pasquali SK, He X, Jacobs JP, Jacobs ML, O'Brien SM, Gaynor JW. Evaluation of failure to rescue as a quality metric in pediatric heart surgery: an analysis of the STS Congenital Heart Surgery Database. *Ann Thorac Surg.* 2012b;94(2):573–9.
- Pasquali SK, Gaies MG, Jacobs JP, William GJ, Jacobs ML. Centre variation in cost and outcomes for congenital heart surgery. *Cardiol Young.* 2012c;22(6):796–9.
- Pasquali SK, Li JS, He X, et al. Perioperative methylprednisolone and outcome in neonates undergoing heart surgery. *Pediatrics.* 2012d;129(2):e385–91.
- Pasquali SK, Li JS, He X, et al. Comparative analysis of antifibrinolytic medications in pediatric heart surgery. *J Thorac Cardiovasc Surg.* 2012e;143(3):550–7.
- Pasquali SK, Peterson ED, Jacobs JP, et al. Differential case ascertainment in clinical registry versus administrative data and impact on outcomes assessment for pediatric cardiac operations. *Ann Thorac Surg.* 2013;95(1):197–203.
- Passaro Jr E, Organ CH, Ernest Jr A. Codman: the improper Bostonian. *Bull Am Coll Surg.* 1999;84(1):16–22.
- Patel MR, Dehmer GJ, Hirshfeld JW, et al. ACCF/SCAI/STS/AATS/AHA/ASNC/HFSA/SCCT 2012 appropriate use criteria for coronary revascularization focused update: a report of the American College of Cardiology Foundation Appropriate Use Criteria Task Force, Society for Cardiovascular Angiography and Interventions, Society of Thoracic Surgeons, American Association for Thoracic Surgery, American Heart Association, American Society of Nuclear Cardiology, and the Society of Cardiovascular Computed Tomography. *J Thorac Cardiovasc Surg.* 2012;143(4):780–803.

- Prager RL, Armenti FR, Bassett JS, et al. Cardiac surgeons and the quality movement: the Michigan experience. *Semin Thorac Cardiovasc Surg.* 2009;21(1):20–7.
- Shahian DM, Normand SL. Comparison of “risk-adjusted” hospital outcomes. *Circulation.* 2008;117(15):1955–63.
- Shahian DM, Edwards FH, Ferraris VA, et al. Quality measurement in adult cardiac surgery: part 1–Conceptual framework and measure selection. *Ann Thorac Surg.* 2007a;83 Suppl 4:S3–12.
- Shahian DM, Silverstein T, Lovett AF, Wolf RE, Normand SL. Comparison of clinical and administrative data sources for hospital coronary artery bypass graft surgery report cards. *Circulation.* 2007b;115(12):1518–27.
- Shahian DM, O’Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1–coronary artery bypass grafting surgery. *Ann Thorac Surg.* 2009a;88 Suppl 1:S2–22.
- Shahian DM, O’Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3–valve plus coronary artery bypass grafting surgery. *Ann Thorac Surg.* 2009b;88 Suppl 1:S43–62.
- Shahian DM, Edwards FH, Jacobs JP, et al. Public reporting of cardiac surgery performance: part 1–history, rationale, consequences. *Ann Thorac Surg.* 2011a;92 Suppl 3:S2–11.
- Shahian DM, Edwards FH, Jacobs JP, et al. Public reporting of cardiac surgery performance: part 2–implementation. *Ann Thorac Surg.* 2011b;92 Suppl 3:S12–23.
- Shahian DM, He X, Jacobs JP, et al. The Society of Thoracic Surgeons isolated aortic valve replacement (AVR) composite score: a report of the STS Quality Measurement Task Force. *Ann Thorac Surg.* 2012a;94(6):2166–71.
- Shahian DM, O’Brien SM, Sheng S, et al. Predictors of long-term survival following coronary artery bypass grafting surgery: results from The Society of Thoracic Surgeons Adult Cardiac Surgery Database (The ASCERT Study). *Circulation.* 2012b;125(12):1491–500.
- Shahian DM, He X, Jacobs JP, et al. Issues in quality measurement: target population, risk adjustment, and ratings. *Ann Thorac Surg.* 2013a;96(2):718–26.
- Shahian DM, Jacobs JP, Edwards FH, et al. The Society of Thoracic Surgeons National Database. *Heart.* 2013b;99(20):1494–501.
- Shahian DM, He X, Jacobs JP, et al. The STS AVR + CABG composite score: a report of the STS Quality Measurement Task Force. *Ann Thorac Surg.* 2014;97(5):1604–9.
- Shahian DM, He X, Jacobs JP, et al. The Society of Thoracic Surgeons composite measure of individual surgeon performance for adult cardiac surgery: a report of the Society of Thoracic Surgeons quality measurement task force. *Ann Thorac Surg.* 2015;100:1315–1325.
- Shapiro M, Swanson SJ, Wright CD, et al. Predictors of major morbidity and mortality after pneumonectomy utilizing the Society for Thoracic Surgeons General Thoracic Surgery Database. *Ann Thorac Surg.* 2010;90(3):927–34.
- Speir AM, Rich JB, Crosby I, Fonner Jr E. Regional collaboration as a model for fostering accountability and transforming health care. *Semin Thorac Cardiovasc Surg.* 2009;21(1):12–9.
- Spiegelhalter DJ. Surgical audit: statistical lessons from Nightingale and Codman. *J R Stat Soc (Series A).* 1999;162(Part 1):45–58.
- Strickland MJ, Riehle-Colarusso TJ, Jacobs JP, et al. The importance of nomenclature for congenital cardiac disease: implications for research and evaluation. *Cardiol Young.* 2008;18 Suppl 2:92–100.
- STS long-term risk calculator. <http://www.sts.org/quality-research-patient-safety/quality/ascert-long-term-survival-calculator>. Accessed 11 July 2014.
- STS National Database. <http://www.sts.org/sections/stsnationaldatabase/>. Accessed 26 July 2014.
- STS Quality Improvement webinars. <http://www.sts.org/education-meetings/sts-webinar-series>. Accessed 12 July 2014.
- STS Research Center. [http://www.sts.org/sites/default/files/documents/pdf/DirectorSTSResearchCenter\\_April2014.pdf](http://www.sts.org/sites/default/files/documents/pdf/DirectorSTSResearchCenter_April2014.pdf). Accessed 13 July 2014.
- STS short term risk calculator. <http://www.sts.org/quality-research-patient-safety/quality/risk-calculator-and-models>. Accessed 11 July 2014.
- Weintraub WS, Grau-Sepulveda MV, Weiss JM, et al. Comparative effectiveness of revascularization strategies. *N Engl J Med.* 2012;366(16):1467–76.
- Wright CD, Kucharczuk JC, O’Brien SM, Grab JD, Allen MS. Predictors of major morbidity and mortality after esophagectomy for esophageal cancer: a Society of Thoracic Surgeons General Thoracic Surgery Database risk adjustment model. *J Thorac Cardiovasc Surg.* 2009;137(3):587–95.



# Health Services Information: Patient Safety Research Using Administrative Data

# 11

Chunliu Zhan

## Contents

<b>Introduction</b> .....	242
<b>Administrative Data: Definition, Data Resources, and Potential Patient Safety Measures</b> .....	243
Medical Claims, Discharge, and Other Health Encounter Abstracts .....	244
Medical Records and Electronic Health Records .....	247
Reports and Surveillance of Patient Safety Events .....	249
Surveys of Healthcare Encounters and Healthcare Experiences .....	250
Other Data Sources and Data Linkage .....	251
<b>Patient Safety Research Using Administrative Data: General Framework, Methods, and Tools</b> .....	252
General Framework for Administrative Data-Based Patient Safety Research .....	252
Methodological Considerations .....	253
AHRQ Patient Safety Indicators: An Exemplary Tool for Administrative Data-Based Patient Safety Research .....	257
<b>Patient Safety Research Using Administrative Data: Potentials and Limitations</b> .....	259
Screen Patient Safety Events for In-depth Examination .....	260
Epidemiological Study .....	260
Public Reporting on Patient Safety Events .....	262
Advantages and Challenges in Administrative Data-Based Patient Safety Research .....	262
<b>References</b> .....	263

### Abstract

A wide variety of data is routinely collected by healthcare providers, insurers, professional

organizations, and government agencies for administrative purposes. Readily available, computer readable, and covering large populations, these data have become valuable resources for patient safety research. A large number of exemplary studies have been conducted that examined the nature and types of patient safety problems, offered valuable

C. Zhan (✉)  
Department of Health and Human Services, Agency for  
Healthcare Research and Quality, Rockville, MD, USA  
e-mail: [chunliu.zhan@ahrq.hhs.gov](mailto:chunliu.zhan@ahrq.hhs.gov)

insights into the impacts and risk factors, and, to some extent, provided benchmarks for tracking progress in patient safety efforts at local, state, or national levels. Various methods and tools have been developed to aid such research. The main disadvantage lies with the fact these administrative data are often collected without following any research design, protocol, or quality assurance procedure; therefore health services researchers using these data sources must make extra efforts in devising proper methodologies and must interpret their findings with extra caution. As more and more administrative data are collected and digitized and more tailored methodologies and tools are developed, health services researchers will be presented with ever-greater opportunity to extract valid information and knowledge on patient safety issues from administrative data.

---

## Introduction

A guiding principle for medical professionals is the Hippocratic oath: *First, Do No Harm*. But, inevitably, patient harms occur, and research is needed to understand why and how to prevent them. Since the Institute of Medicine (IOM) published its landmark report, *To Err Is Human: Building a Safer Healthcare System* (Kohn et al. 1999), in 1999, the importance of vigorous, systematic research on patient safety has been recognized worldwide, and patient safety research has become a prominent domain of health services research. Using a variety of definitions, taxonomies, methods, and databases, health services researchers have addressed a wide range of patient safety-related questions, producing a large body of literature.

To the general public, patient safety is self-defined. As a research topic, its definition is far from universally agreed. IOM defines patient safety as “the prevention of harm to patients”, and its emphasis is placed on “the system of care delivery that (1) prevents errors; (2) learns from the errors that do occur; and (3) is built on a culture of safety that involves health care professionals, organizations, and patients” (Kohn et al. 1999). The Agency for Healthcare Research

and Quality (AHRQ), the US federal agency charged with improving patient safety, defined patient safety as “freedom from accidental or preventable injuries produced by medical care.” The literature is littered with systems of definitions, taxonomies, categorizations, terms, and concepts associated with patient safety. The National Quality Forum’s list of “never events” or “serious reportable events” offers concrete examples of the types of issues patient safety research is concerned with:

- Surgical events: surgery or other invasive procedure performed on the wrong body part or the wrong patient, the wrong surgical or other invasive procedure performed on a patient, and unintended retention of a foreign object in a patient after surgery or other procedure
- Product or device events: such as patient death or serious injury associated with the use of contaminated drugs, devices, or biologics
- Patient protection events: discharge or release of a patient/resident of any age, who is unable to make decisions, to other than an authorized person and patient suicide, attempted suicide, or self-harm resulting in serious disability while being cared for in a healthcare facility
- Care management events: such as patient death or serious injury associated with a medication error (e.g., errors involving the wrong drug, wrong dose, wrong patient, wrong time, wrong rate, wrong preparation, or wrong route of administration), patient death or serious injury associated with unsafe administration of blood products, maternal death or serious injury associated with labor or delivery, and patient death or serious injury resulting from failure to follow up or communicate laboratory, pathology, or radiology test results
- Environmental events: patient or staff death or serious injury associated with a burn incurred in a healthcare setting and patient death or serious injury associated with the use of restraints or bedrails while being cared for in a healthcare setting.
- Radiologic events: death or serious injury of a patient or staff associated with introduction of a metallic object into the MRI area

- Criminal events: any instance of care ordered by or provided by someone impersonating a physician, nurse, pharmacist, or other licensed healthcare provider, abduction of a patient/resident of any age, sexual abuse/assault on a patient within or on the grounds of a healthcare setting, and death or significant injury of a patient or staff member resulting from a physical assault (i.e., battery) that occurs within or on the grounds of a healthcare setting

To focus on the subject at hand, that is, how to use administrative data to conduct patient safety research, this chapter will refer to all events with patient safety implications as *patient safety events* without distinction.

Patient safety research can be done in a number of ways, such as follow-up of cohorts of patients as they come into contact with healthcare systems and randomized trials to examine whether a certain intervention works to reduce patient safety events. However, such studies are rare, due to the fact that patient safety events are accidental in nature, in other words, rare; to gather a sufficient number of cases of patient safety events, a researcher must collect a substantially large study sample. Unsurprisingly, most studies on patient safety were conducted using administrative data, the type of data collected routinely and processed in large volume for administrative purposes.

Health services researchers have used administrative data to study a variety of patient safety issues, from the prevalence to risk factors and effectiveness of interventions to reduce patient safety events. The apparent advantage of administrative data is in its large volume and its computerization, which make the most tenuous and expensive part of research – data collection – relatively easy and cheap. Another advantage is that, because of little risk to interrupting patient care in the data collection process and little risk of patient privacy breach with patient identifiers stripped, data acquisition can be done without jumping through many hoops. The apparent disadvantage lies with the fact that these administrative data are collected without following any research design, protocol, or quality assurance procedure; therefore, researchers using these data must make

extra efforts in devising methodologies and must interpret their findings with extra caution.

Patient safety as a research domain is relatively new compared with other health services research domains, and the issues are diverse and constantly evolving. Administrative data is also fast expanding, with more and more data collected and accumulated as computer technologies progress and interest in mining big data increases. Consequently, patient safety research using administrative data does not follow any clearly defined agenda, methodologies, or processes, giving researchers great room for creativity and innovation and also greater room for error.

This chapter provides a review of the administrative data sources currently available for patient safety research, the common methodologies and tools employed, and the types of patient safety research that can be conducted using administrative data. By going through some well-developed concepts, tools, and examples, the chapter intends to offer health services researchers a road map on how to use administrative data to generate information and knowledge to advance their patient safety agenda.

---

### **Administrative Data: Definition, Data Resources, and Potential Patient Safety Measures**

Administrative data refer to data collected for administrative purposes. Such data are essential for running any kind of business, and the business of healthcare is no exception. Hospitals, outpatient clinics, nursing homes, home care providers, pharmacists, and all other healthcare providers collect and compile data on patients, medical conditions, treatments, and patient directives, create bills for patients and submit claims to insurers and other third-party payers for reimbursements, and compile business data for governance, internal audits, credentialing, and statistical reports. Health insurance companies deal with medical claims in addition to enrolling patients, generating enormous amount of data on a daily basis. Drug companies collect data on drug sales, establish drug registries for postmarket research, and

compile data on drug safety to meet regulatory needs. Professional societies, such as the American Medical Association and the American Hospital Association, also compile extensive data on their members for membership management, licensing, accreditation, and other administrative purposes. Many employers, especially large and traditional companies, offer extensive health benefits, and, for management purposes, compile extensive data to track their employee's use of health benefits and expenses. Last but not least, government agencies compile extensive data, including claims in order to pay the bills for patients covered by government programs, data from healthcare providers to monitor this important sector of the economy, and regular surveys to generate national statistics and track changes over time. Together, tremendous amounts of administrative data are produced and maintained by various entities, and these data hold great potential for research on a wide range of issues, including patient safety issues.

In general, any data source that records personal encounters or experiences with healthcare systems has the potential to contribute information and knowledge on patient safety. Many other data sources containing no patient care data can also be useful when merged with patient encounter data. Table 1 provides a brief summary of the types of administrative data sources that are available and that have been used by health services researchers to study patient safety.

It should be noted that, in health services research literature, claims data are often treated as synonymous to administrative data. It is because medical claims, which record individual patients' individual episodes of care for insurance claims, are the most voluminous data, the first extensively computerized data, and the first administrative data sources extensively used in health services research. However, similar data on individual healthcare encounters are also collected in many countries or programs under universal insurance coverage, and these data are sometimes called discharge abstracts. Following the basic definition of "administrative data," this review also includes other data sources that are collected for administrative purposes, but may be smaller in scale, less computerized, and less often used in health services research. The basic characteristics of these data sources and the potential patient safety measures that can be derived from these data sources are discussed in detail below.

## Medical Claims, Discharge, and Other Health Encounter Abstracts

### Data Sources

A healthcare provider must collect and compile data on each service rendered to each patient, for record keeping, patient tracking, billing, and other administrative purposes. At minimum, the data

**Table 1** Administrative data sources and potential patient safety measures

Data source	Potential patient safety measures
Medical care claims and abstracts	Screening algorithms based on ICD codes, interactive drug-drug pairs, contraindicative drug-event pairs, utilization-condition pairs indicative of inappropriate, over- and underuse of specific medications or procedures
Medical records, electronic medical records	Screening algorithms above, expanded to include more clinical data and text narratives
Reports of medical errors and adverse events, malpractice claims	Each report describes a patient safety event and contextual factors
Survey of healthcare encounters or experiences	Screening algorithms based on ICD codes, interactive drug-drug pairs, contraindicative drug-event pairs, utilization-condition pairs indicative of inappropriate, over- and underuse of specific medications or procedures
Other administrative databases such as census, provider databanks, geo-eco-political databases	Contain no patient safety measures but expand research into population, provider, and regional statistics in relation to patient safety events

include some patient demographics, medical conditions, diagnoses, treatments, discharge or disposition status, and charges and payments. As mentioned earlier, the most important use of such data is to make insurance claims; therefore this type of data is often called “claims data” and further categorized as inpatient claims, outpatient claims, pharmacy claims, and so on. In many countries other than the United States, health encounters are similarly recorded and compiled but not for insurance claims purposes, and this type of administrative records may be called discharge abstracts, for example. Regardless of terms, data on individual healthcare encounters are universal and are available in various capacity for research use.

Researchers rarely have the need to deal with individual hospitals, primary care institutions, nursing homes, outpatient surgical centers, or home care agencies to access such data. Government agencies, insurers, health systems, and many commercial companies compile the data and offer them to various end users. In the United States, the Centers for Medicare and Medicaid Services (CMS) has been a major source of such administrative data. Medicare, a national social insurance program, guarantees access to health insurance for about 50 million Americans aged 65 and older and younger people with disabilities. Medicaid, a social healthcare program jointly funded by the state and federal governments and managed by the states, provide coverage for families and individuals with low income and resources. Together, Medicare and Medicaid process millions of claims each day. CMS has made great efforts to make these claims available to researchers and to standardize the data release process. The latest incarnation of these efforts is called the CMS Data Navigator (CMS 2014), intended to be the one stop for all CMS data sources, through standard processes that include formulated requests, approval, pricing, and payment procedures to ensure proper use and security of the data. The CMS data suite covers enrollment, outpatient care, hospitalization, pharmacy, and services delivered by other types of providers, and the data can be linked to form a rather complete history of individual’s healthcare encounters.

Besides CMS, other federal agencies, state health departments, health plans, and private data institutions have also compiled claims data into research databases. One prominent example is AHRQ’s Healthcare Cost and Utilization Project (HCUP), a partnership of the federal government and states that compiles uniform hospital discharge records for research purposes (HCUP 2014). As of today, HCUP includes databases covering all hospital admissions from 47 states, emergency department visits from 31 states, and ambulatory surgery claims from 32 states. It has derived research databases with a sampling design to yield national estimates and developed various tools to reliably and effectively use these databases. On the private side, Truven Health Analytics MarketScan<sup>®</sup> databases contain complete claims for more than 199 million unique patients, and IMS Health compiles information from 100,000 suppliers from over 100 countries, with more than 45 billion healthcare transactions processed annually.

With the government paying for all health services provided by mostly private providers, Canada collects data on individual health encounters for almost the entire population. Some provinces have data on virtually all records of hospitalizations, pharmacy, physician visits, emergency department visits, and so on for every resident. Many efforts are made to make such data easy for researchers to access and use. For example, the Canadian Institute for Health Information maintains discharge abstract databases of administrative, clinical, and demographic information on hospital discharges received directly from acute care facilities or from their respective health authority or department of health. In the United Kingdom, hospital episode statistics comprises an administrative database of all inpatients in England, covering about 13 million episodes of care annually. Similar databases exist, in various forms, in almost all nations, most of which are available for research purposes.

Regardless of country, healthcare system, or purpose, administrative data of this sort record patient encounters with the healthcare system and capture with similar sets of data elements:

- Patient demographics such as age, sex, race/ethnicity, county of resident and zip code, and expected payer
- Admission status including admission date, admission source and type, and primary and secondary diagnoses
- Treatments such as procedures and medications
- Discharge status entailing discharge date, patient disposition, or death
- Charges and payments

In addition, some identifiers for patients and providers, usually encrypted, are included, allowing for linking individual patient's claims from multiple care settings.

### Potential Patient Safety Measures

A coding system for diagnosis and procedures is essential for recording patient encounters and for generating bills. The United States currently uses International Classification of Diseases, the ninth revision, Clinical Modification (ICD-9-CM), a coding system with three-digit numbers (i.e., 001–999) followed by a decimal point and up to two digits, supplemented by a group of E codes (E000–999) capturing external causes of injury (Iezzoni et al. 1994). Canada, Australia, New Zealand, and many European and Asian countries have been using ICD-10, an alphanumeric system each starting with a letter (i.e., A–Z), followed by two numeric digits, a decimal point, and a digit (Quan et al. 2008).

Some of the codes specifically identify a patient safety event, and some codes suggest that there may be an event of patient safety concern. For example, there are ICD-9-CM diagnosis codes for “foreign object accidentally left in body during a procedure”: 998.4. Some other codes may also suggest such occurrence, including:

998.7: postoperative foreign substance reaction  
 E8710: post-surgical foreign object left in body  
 E8711: postinfusion foreign object left in body  
 E8712: postperfusion foreign object left in body  
 E8713: postinjection foreign object left in body  
 E8714: postendoscopy foreign object left in body  
 E8715: postcatheter foreign object left in body

E8716: post heart catheter foreign object left in body

E8717: post catheter removal foreign object left in body

E8718: foreign object left in body during other specified procedure

E8719: foreign object left in body during non-specified procedure

The corresponding ICD-10 codes for foreign object accidentally left in body during a procedure may include:

T81.509A: unspecified complication of foreign body accidentally left in body following unspecified procedure, initial encounter

T81.519A: adhesions due to foreign body accidentally left in body following unspecified procedure, initial encounter

T81.529A: obstruction due to foreign body accidentally left in body following unspecified procedure, initial encounter

T81.539A: perforation due to foreign body accidentally left in body following unspecified procedure, initial encounter

The process of identifying the right codes and eligible patients to measure patient safety is a mix of science and art. It rarely is clear that one code specifically records a specific patient safety event. The art of the process includes not only selection of relevant codes but also exclusion of patients for whom the codes are not likely to be relevant. Another consideration is whether a recorded event occurred during the current hospitalization (i.e., hospital-acquired condition) or whether it was already present on admission (i.e., comorbid condition). If the code appears as the first, or primary diagnosis, in a claim or discharge abstract, then it can be considered to record an event that is present on admission. But as many as 25 secondary diagnosis codes are recorded in some claims data, and only recently, a code was introduced in Medicare claims to indicate whether a diagnosis is present on admission. A great deal of effort in administrative data-based patient safety research goes into the artistic process with the dual purpose to maximize specificity (i.e., an



event flagged by the codes is truly a patient safety event) and sensitivity (i.e., all patient safety events are flagged). This process is further illustrated in later sections, in conjunction with the discussion of the methods and tools used in administrative data-based patient safety research.

Algorithms can also be built based on coded data other than ICD codes. Claims for medications can be used to screen harmful drug-drug interactions and contraindicative drug-condition interactions. With data linked from multiple settings and over time, certain measures of inappropriate use, underuse, or overuse of care with safety implications can be studied.

## Medical Records and Electronic Health Records

### Data Sources

Medical records are as numerous as claims but much richer in information on patients and their healthcare experiences. Each healthcare encounter has a medical record associated with it to support diagnosis and justify services provided. Broadly speaking, a medical record may contain:

- Patient demographic information: name, address, date of birth, sex, race and ethnicity, legal status of any patient receiving behavioral healthcare services, and language and communication needs, including the preferred language for discussing healthcare issues
- Patient clinical information: reason(s) for admission; initial diagnosis; assessments; allergies to food or latex or medication; medical history; physical examination; diagnoses or conditions established during the patient's course of care, treatment, and services; consultation reports; observations relevant to treatment; patient's response to treatment; progress notes; medications ordered or prescribed; medications administered, including the strength, dose, frequency, and route; adverse drug reactions; treatment goals; plan of care and revisions to the plan of care; results of diagnostic and therapeutic tests and procedures; medications dispensed

or prescribed on discharge; discharge diagnosis; and discharge plan and discharge planning evaluation

- Other information: such as advance directives, informed consent, and records of communication with the patient, such as telephone calls or email.

Medical records can be handwritten, typed, or electronic and can be coded or written in open-text narratives. The rich clinical information makes medical records a good source for patient safety research, allowing identification of various medical injuries, adverse events, errors, and nearmisses and allowing analysis of circumstances and causes of various patient safety events. Earlier research on patient safety used medical records predominantly as the primary data source (Kohn et al. 1999). Those earlier studies mostly had to work with medical records in paper format or electronic format that was not readily usable for research and had to rely on medical experts to transform medical records into research data, a process that was resource intensive and required exceptional knowledge and skills in medical context and research. As a result, earlier patient safety research with medical records was usually limited in scope and statistical power.

The wide adoption of electronic medical records (EMRs) offers great promise for patient safety research. In the United States, a substantial percentage of hospitals and physicians have started to use EMR systems, with various levels of capacity and usability. In the United Kingdom, the National Health Service collects and stores data electronically on primary care encounters in the clinical information management system. Great efforts are being made in Canada and all over the world to move the healthcare industry into the Information Age.

### Potential Patient Safety Measures

In theory, EMRs hold much of what claims data can offer and much more. EMRs contain a great deal of information in structured, coded data similar to administrative data. The allure of EMR data in patient safety research lies with its rich clinical data, such as lab values, and narratives that record

**Table 2** Medical record-based screening for patient safety events: adverse drug events associated with warfarin

Description	Screening algorithm
Numerator	The subset of the denominator who during the hospital stay experienced:
	An INR $\geq 4.0$ with one or more of the following symptoms: cardiac arrest/emergency measures to sustain life, death, gastrointestinal bleeding, genitourinary bleeding, a hematocrit drop of three or more points more than 48 h after admission, intracranial bleeding (subdural hematoma), a new hematoma, other types of bleeding or pulmonary bleeding
	An INR $>1.5$ and an abrupt cessation/hold of warfarin with one or more of the above symptoms
	An INR $>1.5$ and administration of vitamin K or fresh frozen plasma (FFP) with one or more of the above symptoms
	An INR $>1.5$ and a blood transfusion absent a surgical procedure with one or more of the above symptoms
Denominator	All patients who received warfarin during hospitalization and had a documented INR result during the hospital stay

medical providers’ observations, judgments, treatment details, and outcomes. Screening algorithms can be designed to search for patient safety events in coded data as well as in text narratives. The search can look for falls, retrieve lab data on toxic serum levels of digoxin, or screen for international normalization ratios greater than 6 in patients on warfarin. It can entail a sophisticated, explicit, structured query of entire medical records. Table 2 shows an example that screens EMRs for possible adverse drug events for patients on warfarin.

Such algorithms can be used in manual review of medical records and can also be used to design automatic review of EMRs.

There are many challenges in implementing such explicit screening algorithms, and compromises are made. The Institute for Healthcare Improvement (Griffin and Resar 2009) has developed a set of global trigger tools that screen medical records for possible adverse events, including groups of triggers for medical, surgical, and medication-related patient harms. The tools screen coded data; look for the most significant, easy-to-detect signs; and can be applied by healthcare organizations to review paper-based and also electronic medical records. The trigger tools have been adopted by many countries and health systems. For example, Adventist Health System used the tools to gauge the number, types, and severity levels of adverse events in 25 hospitals that used a common EMR system and developed a centralized process to do so uniformly, including quarterly reports to participating

facilities to communicate findings and case studies illustrating the most egregious harms.

With regard to rich notes and other narratives in EMRs, there has been much hype but little real progress. The method to identify, extract, and encode relevant information from tremendous volumes of text narratives is called natural language processing (NLP). In general, EMR narratives are stored following internal structure; information extraction involves the selection of the relevant sections of EMR and then targeted text data processing. NLP systems, such as MEDSYNDIKATE, MetaMap, SemRep, MedLEE, and BioMedLEE, can extract data pertaining to patient safety events. In a recent study of adverse drug events attributable to six drugs, Wang et al. (2009) demonstrated the general process, which consists of five stages: (1) collecting the set of EMRs to be mined, (2) processing the summaries using NLP to encode clinical narrative data, (3) selecting data while co-occurrence match of a specific drug and its potential adverse drug events exist, (4) filtering data by excluding confounding information such as diseases or symptoms that occurred before the use of the drug, and (5) analyzing and determining the drug-adverse drug events association.

In theory, any type of errors and adverse events that can be recognized by a clinician going through a medical record can be captured electronically. However, this theory is far from being realized. There are many EMR systems that vary substantially in structure, format, and content, and there are legal and practical obstacles over data

sharing. However, some healthcare systems have started to pull together EMR data for research. It is expected that in the near future, research databases composed of large volume of medical records from many providers and cross care settings, databases resembling HCUP or CMS data navigator, will be created and made available to health services researchers.

## Reports and Surveillance of Patient Safety Events

### Data Sources

Alternative data sources for patient safety research include mandatory and voluntary reports of medical errors or adverse events, drug safety, or nosocomial infection surveillance systems and other data systems that government agencies and nongovernmental organizations use specifically to monitor patient safety. Spontaneous reporting systems have been created as the primary means for providing postmarket safety information on drugs since the 1960s, and some systems have also covered patient safety events due to inappropriate use of drugs. Such systems exist all over the world in various names and with various mandates.

This type of data sources records individual incidences of patient safety events and varies tremendously in formats and contents. One prominent example of such a reporting system is the US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS). FAERS contains information on adverse event and medication error reports submitted to the FDA by healthcare professionals and consumers voluntarily as well as by drug manufacturers who are required to send all adverse event reports they receive from healthcare providers and consumers. The database is designed to support the FDA's postmarketing safety surveillance program for drug and therapeutic biologic products, to help FDA look for new safety concerns that might be related to a marketed product, to evaluate a manufacturer's compliance with reporting regulations, and to respond to outside requests for information. Besides regulatory use, the FDA provides raw data consisting of individual case reports

extracted from the FAERS database to researchers inside and outside of the FDA. Similar to the FDA FAERS, the UK's Medicines and Healthcare Products Regulatory Agency institutes a Yellow Card Scheme that allows patients and health professionals to report suspected side effects. The reports are continually assessed by medicine safety experts, together with additional sources of information such as clinical trial data, the medical literature, and data from international medicines regulators, in order to identify previously unidentified safety issues or side effects.

MEDMARX is a similar system of voluntary reports but focuses on medication errors. Currently, MEDMARX contains over 1.3 million medication error records reported by over 400 healthcare facilities that voluntarily participate. The program collects information on medication errors, categorizing them into nine severity levels, ranging from errors that do not reach patients to errors that cause death. The reporting system contains up to 13 required data elements and 29 optional data elements to describe error types, causes, locations, staff involved, products involved, and patient characteristics. The system also asks about actions taken in response to the errors, including both individual procedural activities (i.e., actions to recover from the error) and practice-based changes (i.e., actions to prevent future errors). Most data elements are coded fields allowing single or multiple selection, and some data fields are for textual descriptions.

Some surveillance systems collect similar data but make reporting mandatory in order to accurately track incidences of patient safety events. The Centers for Disease Control and Prevention (CDC) National Nosocomial Infections Surveillance System is a prominent example of such a data source, which has continued gathering reports from a sample of hospitals in the United States on nosocomial infections since the 1970s. Another example is the National Electronic Injury Surveillance System (NEISS) at the CDC, composed of a national probability sample of hospitals in the United States that collect patient information for every emergency visit involving an injury associated with consumer products, including medical products. More recently, to address

heightened public concerns over drug safety, the system started a Cooperative Adverse Drug Event Surveillance Project (NEISS-CADES) to capture cases that are defined as those occurring in persons who sought emergency care for injuries linked by the treating physician to the outpatient use of a drug or drug-specific adverse effects. Using NEISS-CADES, Budnitz et al. (2011) were able to estimate that adverse drug events in older Americans accounted for about 100,000 emergency hospitalizations a year in the United States, and four medications (warfarin, insulins, oral antiplatelet agents, oral hypoglycemic agents) were implicated alone or in combination in two thirds of the cases.

### **Patient Safety Measures**

Because each record of this type is to provide details for one specific patient safety event, no effort is needed to identify or validate the reported event. The data allows various targeted research, such as the types of errors or adverse events most frequently occurring, the circumstances, the possible causes as reported, and the follow-up actions. But this type of data has some obvious limitations for patient safety research. First, the reported event (adverse event or medication error) may not be due to the product or a causal relationship with the product. Second, the reports do not always contain enough detail to properly evaluate an event. Third, because of the voluntary nature of data submission, the system does not receive reports for every adverse event or medication error that occurs; therefore, the data cannot be used to calculate the incidence of an adverse event or medication error in a population. Lastly, this type of data contains no controls (i.e., the patients without patient safety events), severely limiting its use in epidemiological research.

## **Surveys of Healthcare Encounters and Healthcare Experiences**

### **Data Sources**

Many government agencies conduct routine surveys to collect data in order to produce national statistics and track changes in the healthcare

sector. Some of the surveys collect data on personal encounters with healthcare systems and, therefore, are potential data sources for patient safety research.

In the United States, the National Center for Health Statistics, under the CDC, conducts a wide array of national surveys that contain healthcare encounter experiences. The National Ambulatory Medical Care Survey collects information about the provision and use of ambulatory medical care services, drawing a random sample of visits to nonfederal, office-based physicians who are primarily engaged in direct patient care. The National Hospital Ambulatory Medical Care Survey collects similar data, on the utilization and provision of ambulatory care services in hospital emergency and outpatient departments from a national sample of visits to the emergency departments and outpatient departments of noninstitutional, general, and short-stay hospitals. The National Hospital Discharge Survey collects data from a national sample of hospital discharges from nonfederal, short-stay hospitals. The National Hospital Care Survey, a relatively new database, integrates inpatient data formerly collected by the National Hospital Discharge Survey with the emergency department, outpatient department, and ambulatory surgery center data collected by the National Hospital Ambulatory Medical Care Survey, with personal identifiers linking care provided to the same patient in the emergency departments, outpatient departments, ambulatory surgical centers, and inpatient departments.

Beside surveys of healthcare encounters as listed above, some surveys ask patients and families directly for information on their healthcare experiences. CMS Medicare Current Beneficiary Survey is such a data source, containing survey responses from a random sample of Medicare beneficiaries and linking to their administrative data covering inpatient, outpatient, and other claims. AHRQ Medical Expenditure Panel Survey is a set of large-scale surveys of families and individuals, their medical providers, and employers on healthcare use and spending.

Similar surveys of healthcare encounters, residents, or families exist in various forms in

many other countries. For example, the Canadian Community Health Survey resembles the Medical Expenditure Panel Survey in general purposes and methods, collecting information annually on a large sample of the Canadian population on information related to health status, healthcare utilization, and health determinants.

### **Patient Safety Measures**

Surveys of healthcare encounters and healthcare use usually contain data on medical conditions, diagnoses, and procedures, coded by ICD-9-CM or other similar coding systems. As with claims data, some patient safety indicators can be derived from the coded data. Depending on the data collected, other screening algorithms can be designed. For example, many surveys collect data on medication prescriptions, and measures of inappropriate medication prescriptions can be derived by screening medications that generally should not be prescribed to patients with advanced age or with certain medical conditions. Once a patient safety event is identified with moderate specificity and sensitivity, survey data support a wide range of patient safety research, especially with national statistics, variation across regions and social strata, and changes over time.

### **Other Data Sources and Data Linkage**

Many other administrative data sources besides the four types discussed earlier contain information on individual events of patient safety concerns. Malpractice claims, for example, contain rich data for patient safety research. A malpractice claim is a written demand for compensation for a medical injury, alleging that an attending physician or a care provider is responsible for the injury due to missed or delayed or wrong diagnosis or treatment. A claims file captures information on an entire litigation, from statement of claim, depositions, interrogations, reports of internal investigations, root cause analyses, expert opinions from both sides, medical records and analysis, and final resolution and payments. Working with malpractice insurance companies, researchers can access closed malpractice claims to study the nature,

causes, and circumstances of the underlying errors and identify potential strategies to improve patient safety.

Combining multiple data sources for research has been a significant trend in recent years. The FDA's Mini-Sentinel Project is an example. Tasked with monitoring the safety of approved medical products, the postmarket surveillance system consists of claims data from 18 private health plans covering about 100 million people, supplemented by EMR data from 18 healthcare organizations, designed to answer the FDA's questions on postmarket safety. The claims data capture the complete records of individuals' exposure to a specific medical product in question and limited measures of patient outcomes such as death and major, codified complications. The linked EMR is then used to confirm a diagnosis and adverse events. The data are hosted locally with individual participants to protect privacy and confidentiality and are aggregated through common data formats and analytical modules. This complicates the data analysis somewhat, but with flexible design and proper stratification, such combined data can answer a great number of patient safety questions efficiently.

Some administrative data sources that are not concerned with patient safety events can be of great value to patient safety research. Data collected from providers for statistics, membership management, or licensing purposes can be merged with patient encounter data capable of identifying patient safety events. The American Hospital Association's Annual Survey, for example, contains hospital-specific data on approximately 6,500 hospitals and 400-plus systems, including as many as 1,000 data fields covering organizational structure, personnel, hospital facilities and services, and financial performance. By linking this data with data on personal healthcare encounters, researchers can study a variety of hospital-level factors in relationship to patient safety events. The American Medical Association maintains a suite of membership data, including the Physician Masterfile that contains extensive personal and practice-related data for more than 1.4 million physicians, residents, and medical students in the United States. By

linking this file with other data, researchers are able to examine physician-related factors in relation to patient safety events. Other types of organizations, such as nursing homes, home care agencies, hospice, and primary care practices, all maintain similar membership data, and, in theory, all can be linked to amplify patient safety research.

Population census data and geopolitical data can make similar contributions to patient safety research. Population surveys can provide denominator information such as total population and subpopulations by age, racial, economic, and other categories. The Area Resource File, compiled by the US government, contains information on health facilities, health professions, measures of resource, health status, economic activity, health training programs, and socioeconomic and environmental characteristics. By linking this file with other patient safety data through geographic codes, researchers can explore geographic variation in patient safety events and related econ-geo-political factors.

Data access to many of the above data sources can be challenging, but the challenges are fewer and less restricting compared with other data-gathering efforts. Government-owned data are usually available following straightforward processes. Data owned by private organizations can be obtained in many ways, including, through collaboration with the data owners or researchers intimate with the data owners.

---

### **Patient Safety Research Using Administrative Data: General Framework, Methods, and Tools**

Because administrative data are not collected or compiled following an a priori study design, efforts in choosing appropriate methods and in presenting the results in light of inherent limitations of various data sources are of great importance in generating valid information and knowledge on patient safety questions. This section offers a brief review of the general framework, methods, and tools for patient safety research using administrative data.

### **General Framework for Administrative Data-Based Patient Safety Research**

Generally speaking, there are two types of research: estimation and hypothesis testing. Since patient safety research is a relatively new field, most published studies since the landmark 1999 IOM report have been about estimating prevalence and incidence of patient safety events and distributions by categories, settings, causes, and circumstances. It is well recognized that each administrative data source has an inherent population, such as Medicare beneficiaries from Medicare claims, which is further refined by exclusion and inclusion criteria defined by the patient safety screening algorithms employed. The focus for a robust estimation study is to correctly identify the numerators (i.e., patient safety events) and the denominators (i.e., the underlying population at risk for the patient safety events), a seemingly straightforward but in reality rather tenuous process.

To test hypotheses, administrative data-based patient safety research usually follows the general framework of regression analysis in epidemiology.

To test hypotheses, administrative data-based patient safety research usually follows the general framework of regression analysis in epidemiology in which the occurrence of a patient safety event  $Y$  is related to possible causes being examined or interventions evaluated  $X$  and confounding factors  $Z$ . Within this framework two types of questions can be addressed. The first type of question is why a patient safety event occurs, and the second type of question is what are the consequences of such an event.

In answering both questions, the most critical task is to build an analytical dataset out of one or more administrative data sources for a specific patient safety research question. This step involves the correct identification and measurement of  $X$ ,  $Y$ , and  $Z$  in the context of study cohorts of selected study subjects and time-stamp data, matching the data sources (e.g., who is in the dataset and what  $X$ ,  $Y$ , and  $Z$  can be correctly measured and time-stamped) and the research questions to be answered. The second step is relatively easier,

using established statistical models or more advanced data-mining techniques to estimate the relational parameters in the equation. The third step, interpreting the results and making valid inferences in the full acknowledgment of data limitations, also demands great attention.

## Methodological Considerations

### Identification of Patients with Patient Safety Events

The previous section went through the list of potential administrative data sources and potential patient safety measures these data sources may offer. It is clear that the usefulness of an administrative data source in patient safety research depends, first of all, on the ability of the data source to correctly identify patient safety events. The validity of derived patient safety measures depends on carefully designed and validated indicators, screening algorithms, or triggers. Therefore, with the exception of medical error reports and malpractice claims where each record is, by definition, a patient safety event, a robust patient safety research project starts with the most critical task of screening, determining, and ascertaining patient safety events. This is a process of science, rooted in the researchers' understanding of the relevant medical knowledge, the data-generating process, the structure of the specific databases, and the specific purposes of the relevant research. It is also an art since there is usually no set formula for health services researchers to follow in completing this first step.

In general, the validity of an administrative data-based patient safety measure can be evaluated by specificity and sensitivity, with medical record review serving most often as the gold

standard. Specificity is defined by the positive predictive value (PPV), which is the proportion of patients flagged in the administrative data as having patient safety events who actually had such events, as confirmed by medical record review or other ascertaining methods. Sensitivity is the proportion of the patients with patient safety events that are actually flagged in the administrative data. Table 3 shows the calculation.

Zhan and his colleagues (2009) demonstrated the complexity of this issue in a study that attempted to determine the validity of identifying hospital-acquired catheter-associated urinary tract infections (CAUTIs) from Medicare claims, using medical record review as the gold standard. They found that ICD-9-CM procedure codes for urinary catheterization appeared in only 1.4 % of Medicare claims for patients who had urinary catheters. As a result, using Medicare claims to screen UTIs cannot be limited to claims that have a procedure code for urinary catheterization. Using major surgery as the denominator, Medicare claims had a PPV of 30 % and sensitivity of 65 % in identifying hospital-acquired CAUTIs. Because 80 % of the secondary diagnosis codes indicating UTIs were present on admission (POA), adding POA indicators in the screening algorithm would increase the PPV to 86 % and sensitivity to 79 % in identifying hospital-acquired CAUTIs. This study indicates that the screening algorithm based on the selected ICD-9-CM codes and POA code and confined to major surgery patients is a valid way to identify patients with hospital-acquired CAUTIs in Medicare claims data. Claims from private insurance do not currently contain POA codes and, therefore, are not suitable for research aimed at estimating CAUTI prevalence or hypothesis testing due to the 70 % false-positive rate.

**Table 3** Calculation of specificity and sensitivity of a patient safety measure based on administrative data, using medical record review as the gold standard

	Medical record review	
<b>Administrative data screening</b>	With patient safety event	Without patient safety event
With patient safety event	True positive (TP)	False positive (FP)
Without patient safety event	False negative (FN)	True negative (TN)
<b>Validity calculation</b>	<b>PPV = TP/(TP + FP); Sensitivity = TN/(TN + FN)</b>	

Because medical record review is labor intensive and expensive, researchers often cannot validate the screening algorithms they use and have to rely on what has been reported in the literature. In many cases, validity data are entirely unavailable. Nonetheless, researchers need to have a clear understanding of the specificity and sensitivity in the case identification algorithms they use based on relevant literature, context analysis, or experience and decide whether the patient safety measures are valid enough for their research purposes and discuss their results in light of these limitations.

### Construction of Analytical Dataset

Only with confidence that patient safety events can be identified with an acceptable level of specificity and sensitivity from an administrative data source should a researcher proceed to construct an analytical dataset. As discussed earlier, most administrative data contain measures of basic personal information, medical conditions, diagnosis, treatment, and disposition, and the administrative data can be expanded by linking to other data sources on patients (e.g., National Death Index), providers (e.g., AHA hospital surveys), local socioeconomic data (e.g., Area Resource Files), and so on (e.g., census population statistics), to form analytical files. From these extended datasets, arrays of variables of interest, such as dependent variables, explanatory variables, or confounding controls, can be constructed, including:

- Patient characteristics: age, sex, insurance coverage, etc.
- Medical conditions and diagnoses: primary diagnosis, secondary diagnoses, comorbidities, etc.
- Treatment or utilization: medical and surgical procedures, medications, outpatient visits, etc.
- Patient outcomes: disposition (including death), length of stay, charges or payments, complications, etc.
- Provider characteristics: ownership, practice size and composition, financial status, etc.
- Area characteristics: population statistics, market competitiveness, managed care market share, etc.

These variables support a wide range of cross-sectional analyses and longitudinal studies when the variables are time-stamped. Many claims databases, such as Medicare claims, allow researchers to build the complete profile of patient's healthcare experiences from multiple settings (e.g., inpatient, outpatient, pharmacy), over multiple years. Researchers can identify not only cases of patient safety events and controls but also cohorts to retrospectively follow over time, greatly expanding the capacity of any single administrative data source.

Besides identifying administrative databases with variables of interest, one crucial consideration in analytical data construction is the linkage of multiple data sources. The simplest kind of record linkage is through a unique identification number, such as social security number, or multiple variables that accurately identify a person, such as name, age, date of birth, gender, address, phone number, and so on. This method is called deterministic or rules-based record linkage. Sometimes, a personal identifier is combined with some personal demographic data in databases with missing data or errors in the identifier. Administrative data sources often do not contain or share common identifiers, and a new method called probabilistic record linkage can be used. Probabilistic record linkage takes into account a wider range of potential identifiers, computing weights for each identifier based on its estimated ability to correctly identify a match or a non-match, and uses these weights to calculate the probability that two given records refer to the same entity. Record pairs with probabilities above a certain threshold are considered to be matches, while pairs with probabilities below another threshold are considered to be non-matches; pairs that fall between these two thresholds are considered to be "possible matches" and can be dealt with accordingly (e.g., human reviewed, linked, or not linked, depending on the requirements).

### Data Analysis

For most patient safety studies using administrative data, the methods are simple and straightforward; the common statistical methods for observational studies, such as logistic regressions



with the dichotomous variable of having a patient safety event or not as the dependent variable and ordinary least-square regression with a continuous dependent outcome variable as dependent variable, apply. As with observational studies, administrative data-based patient safety research can fall into the following broad categories:

- Cross-sectional study, involving studying a population at one specific point in time
- Case-control study, in which two existing groups differing in outcome are identified and compared on the basis of some hypothesized causal attribute
- Longitudinal study, involving repeated observations of the same variables over long periods of time
- Cohort study, a particular form of longitudinal study where a group of patients is closely monitored over a span of time

However, administrative data-based patient safety research is unique in many ways. First, the number of observations is substantially larger than studies of experimental design or involving primary data collection. Second, because, by definition, patient safety events are unintended or unexpected; the cases of interest (i.e., patient safety events) are usually very small in numbers and rates. The standard approaches to causal inference or risk adjustment easily produce statistically significant findings that are small and clinically meaningless. Third, the cases of interest are identified with a certain level of uncertainty or misclassification errors, as discussed earlier. These particulars should be born in mind when devising analytical approaches.

The following general methods have been used in administrative data-based patient safety research:

- Matching: matching is a conceptually straightforward strategy, whereby confounders are identified and patients in the cases (e.g., those with patient safety events) are matched to the controls (e.g., those without safety events) on the basis of these factors so that, in the end, the

case group and control group are “the same” with regard to these factors. Matching can either be done on a one-to-one basis or one-to-many basis, and patients can be matched with respect to a single confounder or multiple confounders. This method is particularly applicable to administrative data-based patient safety research because patients with safety events are few and potential controls are many; therefore, it is relatively easy to find one or more matching controls for each case.

- Stratification: once a confounding variable is identified, the cohort is grouped by levels of this factor. The analysis is then performed on each subgroup within which the factor remains constant, thereby removing the confounding potential of that factor.
- Multivariable regression: regression analysis, the most commonly used analytical technique, is based on modeling the mathematical relationships between two or more variables in observed data. In the context of administrative data-based patient safety research, there are four types of outcome measures. The first type is a binary outcome, such as surgical site infections complicating total hip replacement, where multivariable logistic regression is the proper method to identify factors associated with the infections. The second type is a continuous outcome, such as functional status or costs, where multivariable linear regression is applicable to study the influence of various predictors of the outcomes. The third type is an incidence rate, such as nosocomial infection rates at individual hospitals, where Poisson regression may be the best method to identify hospital-level factors that predict higher or lower nosocomial infection rates. The fourth type is a time-to-event outcome, such as reoperation following initial operation, where Cox proportional hazards model may be most appropriate to study risk factors.
- Propensity score analysis: propensity score analysis entails two steps. In the first step, it summarizes multiple confounding variables into a probability or “propensity” of having a patient safety event or falling into an intervention group, usually generated by a logistic

regression model, with the propensity score ranging from 0 to 1. In the second step, the propensity score is used for matching or performing stratified analysis or to be inserted into multivariable regression to estimate the impact of a patient safety event or an intervention.

- Instrumental variable analysis: the instrumental variable approach is a method for confounding control that has been used by economists for decades but has only recently been implemented in health services research. The basic idea is that if a variable (the instrumental variable) can be identified, that has the ability to cause variation in the treatment of interest but that has no impact on outcome (other than through its direct influence on treatment). Then the variable can be used as an instrument in the regression analysis to control for unobserved or unobservable confounding variables on the outcome variable.
- Data-mining methodologies: data mining refers to an analytic process designed to explore data (usually large amounts of data, known as “big data”) in search of consistent patterns and systematic relationships between variables and then to validate the findings by applying the detected patterns to new subsets of data. One example of data-mining methods used in administrative data-based patient safety research is called disproportionality analysis, which creates algorithms that calculate observed-to-exposed ratios. For example, to find the link between a drug and a suspected adverse event, researchers can compare each potential drug-adverse event pair to background across all other drugs and events in the database and flag those pairs with disproportional ratios for further causal investigation. Unsupervised machine learning is another example, encompassing many data-mining methods purported to discover meaningful relationships between variables in large databases.
- Contextual analysis: some administrative data sources contain extensive narrative data. Screening text data for information on patient safety events is costly and, sometimes,

unproductive even with advanced NLP techniques. By cascading steps through coded data, researchers can narrow down the text data and read selected text narratives to gain valuable insights. For example, in their analysis of warfarin-related medication errors, Zhan et al. (2008) found that one hospital reported dispensing errors four times higher than average, two thirds of the errors occurred in the hospital’s pharmacy department, and 65 % of the errors were caused by inaccurate/omitted transcriptions. The textual descriptions in these reports clearly revealed the difficulties the pharmacists were having with the hospital’s new medication administration record system, therefore pinpointing the fix.

In summary, all methods for observational studies in epidemiology, sociology, and economics are applicable to administrative data-based patient safety research. Health services researchers should consult textbooks in these fields and also follow the advancement of methodologies in data mining, pattern recognition, and machine learning that are being developed and increasingly applied to extract information and knowledge from “big data” in the Information Age.

### Interpreting the Results

The results from administrative data-based patient safety research must be interpreted in light of the limitations implicit both in the data and in the methods. First of all, the specificity and sensitivity of the methods or algorithms used to screen or identify patient safety events must be adequately explained, and the potential bias due to misclassification of cases needs to be discussed. Similar measurement errors may also occur in other important variables derived from administrative data, and similar discussions need to be made.

Second, administrative data-based patient safety research shares the same flaws that all observational studies have. Regardless what methods are used, there is always the possibility that confounding remains in the results, due to a wide range of possible causes from unobserved or missed confounders, to measurement errors and mis-specifications of analytical models.

Furthermore, multiple other criteria are required to establish causation. For example, multivariable adjustment cannot give causation unless factors such as appropriate temporal ordering of predictors and outcome are ensured. Finally, health services researchers must completely report how the analyses were undertaken. From choice of confounders to the statistical procedure used, adequate information should be provided so that an independent analyst can reliably reproduce the reported results.

### **AHRQ Patient Safety Indicators: An Exemplary Tool for Administrative Data-Based Patient Safety Research**

The AHRQ patient safety indicators (AHRQ PSIs) are one of the most popular measurement tools for screening patient safety events in administrative data (AHRQ 2014). Developed in the United States in the context of claims data using ICD-9-CM coding system, this toolkit has been adopted worldwide. A case study of AHRQ PSIs serves to illustrate the general process, the potentials, the challenges, and the limitations of administrative data-based patient safety research.

AHRQ PSIs started with Iezzoni and colleagues' 1994 complication screening program (CSP) that relied on ICD-9-CM codes in claims data to identify 27 potentially preventable in-hospital complications, such as postoperative pneumonia, hemorrhage, medication incidents, and wound infection. In the mid-1990s, AHRQ broadened the CSP to include a set of administrative data-based quality indicators, including several measures of avoidable adverse events and complications. Realizing the potential value of administrative data-based measures in identifying patient safety events, AHRQ contracted with the Evidence-based Practice Center at the University of California, San Francisco, and Stanford University to further expand, test, and refine these measures as well as improve the evidence behind their use with extensive literature reviews and broad clinical consensus panels. The research team developed AHRQ PSIs through a five-step process (Romano et al. 2003). First, they reviewed

literature to develop a list of candidate indicators and collected information about their performance. Second, they formed several panels of clinician experts to solicit their judgment of clinical sensibility and their suggestions for revisions to the candidate indicators. Third, they consulted ICD-9-CM coding experts to ensure that the definition of each indicator reflects the intended clinical situation. Fourth, they conducted empirical analysis of the promising indicators using HCUP data. Last, they produced the software and documentation for public release by AHRQ.

Since its inception, AHRQ PSIs have been constantly validated and updated. The latest PSIs (AHRQ 2014) include 23 indicators and one composite indicator with reasonable face and construct validity, specificity, and potential for fostering quality improvement. Most indicators use per 1,000 discharges as the denominators, listed below. Some of the indicators are designed to capture event rates within a community:

- PSI 02 Death Rate in Low-Mortality Diagnosis Related Groups (DRGs)
- PSI 03 Pressure Ulcer Rate
- PSI 04 Death Rate among Surgical Inpatients with Serious Treatable Conditions
- PSI 05 Retained Surgical Item or Unretrieved Device Fragment Count
- PSI 06 Iatrogenic Pneumothorax Rate
- PSI 07 Central Venous Catheter-Related Blood Stream Infection Rate
- PSI 08 Postoperative Hip Fracture Rate
- PSI 09 Perioperative Hemorrhage or Hematoma Rate
- PSI 10 Postoperative Physiologic and Metabolic Derangement Rate
- PSI 11 Postoperative Respiratory Failure Rate
- PSI 12 Perioperative Pulmonary Embolism or Deep Vein Thrombosis Rate
- PSI 13 Postoperative Sepsis Rate
- PSI 14 Postoperative Wound Dehiscence Rate
- PSI 15 Accidental Puncture or Laceration Rate
- PSI 16 Transfusion Reaction Count
- PSI 19 Obstetric Trauma Rate-Vaginal Delivery Without Instrument
- PSI 21 Retained Surgical Item or Unretrieved Device Fragment Rate

Table 4 describes, as an example, the definition of the numerator, denominator, and key exclusions for PSI #13, postoperative sepsis.

AHRQ created software that implements evidence-based and consensus-approved algorithms; calculates raw rates, risk-adjusted rates

that reflect the US hospitalized population in age, sex, DRGs, and comorbidities; and estimates smoothed rates that dampen random fluctuations over time. Thirty comorbidity categories are automatically generated by the software and used as risk adjusters along with variables available in

**Table 4** Claims-based screening for patient safety events: AHRQ PSI #13, postoperative sepsis

Description	Screening algorithm
Numerator	Discharges, among cases meeting the inclusion and exclusion rules for the denominator, with any secondary ICD-9-CM diagnosis codes for sepsis. ICD-9-CM sepsis diagnosis code 1 0380 STREPTOCOCCAL SEPTICEMIA 0381 STAPHYLOCOCCAL SEPTICEMIA 03810 STAPHYLOCOCC SEPTICEM NOS 03811 METH SUSC STAPH AUR SEPT 03812 MRSA SEPTICEMIA 03819 STAPHYLOCC SEPTICEM NEC 0382 PNEUMOCOCCAL SEPTICEMIA 0383 ANAEROBIC SEPTICEMIA 78552 SEPTIC SHOCK 78559 SHOCK W/O TRAUMA NEC 9980 POSTOPERATIVE SHOCK 99800 POSTOPERATIVE SHOCK, NOS 99802 POSTOP SHOCK,SEPTIC 03840 GRAM-NEGATIVE SEPTICEMIA NOS 03841 H. INFLUENAE SEPTICEMIA 03842 E COLI SEPTICEMIA 03843 PSEUDOMONAS SEPTICEMIA 03844 SERRATIA SEPTICEMIA 03849 GRAM-NEG SEPTICEMIA NEC 0388 SEPTICEMIA NEC 0389 SEPTICEMIA NOS 99591 SEPSIS 99592 SEVERE SEPSIS
Denominator	Elective surgical discharges, for patients ages 18 years and older, with any-listed ICD-9-CM procedure codes for an operating room procedure. Elective surgical discharges are defined by specific DRG or MS-DRG codes with admission type recorded as elective (SID ATYPE=3) Exclude cases: With a principal ICD-9-CM diagnosis code (or secondary diagnosis present on admission) for sepsis (see above) With a principal ICD-9-CM diagnosis code (or secondary diagnosis present on admission) for infection With any-listed ICD-9-CM diagnosis codes or any-listed ICD-9-CM procedure codes for immunocompromised state With any-listed ICD-9-CM diagnosis codes for cancer With length of stay of less than 4 days MDC 14 (pregnancy, childbirth, and puerperium) With missing gender (SEX=missing), age (AGE=missing), quarter (DQTR=missing), year (YEAR=missing), or principal diagnosis (DX1=missing)

most administrative data systems. The PSI website also provides software (in Windows and SAS), benchmark tables, and risk-adjustment data for individual hospitals, hospital systems, health plans, state, and other interested parties to calculate their own risk-adjusted rates and make comparison to national benchmarks. Researchers can download the document and software for free (AHRQ 2014).

The specificity and sensitivity of these indicators have been evaluated, accounting for a substantial portion of published literature on AHRQ PSIs. It appears that the validity of AHRQ PSIs varies substantially from indicator to indicator, depending also on the data sources and gold standards used.

Broadly speaking, AHRQ PSIs have been used for:

- Internal hospital quality improvement: individual hospitals use them as a case finding trigger, to do root cause analyses, to identify clusters of potential safety lapses, to evaluate impact of local interventions, and to monitor performance over time.
- External hospital accountability to the community: local government, health systems, and insurance carriers such as Blue Cross/Blue Shield of Illinois produce hospital profiles to support consumers.
- National, state, and regional analyses: government and researchers used it to produce aggregate statistics, e.g., AHRQ's for National Healthcare Quality/Disparities Reports, for surveillance of trends over time, and for assessing disparities across areas, socioeconomic strata, ethnicities, and so on.
- Testing research hypotheses related to patient safety: researchers has used the PSIs to test various hypotheses on patient safety risk factors, such as those that support house staff work hours reform and nurse staffing regulation.
- Public reporting by hospital: several states (e.g., Texas, New York, Colorado, Oregon, Massachusetts, Wisconsin, Florida, and Utah) include AHRQ PSIs measures in their public reporting of hospital quality.

- Pay-for-performance by hospital: some reform initiatives, such as CMS/Premier Demonstration, include AHRQ PSIs measures in pay-for-performance determination.

AHRQ PSIs continue to evolve. Besides periodical refinements, one development hinges on the addition of time stamps on diagnosis codes (i.e., present-on-admission code) in claims or discharge abstracts. This code helps to separate hospital-acquired adverse events (i.e., events occurred after admission) from comorbidities (i.e., conditions present on admission). Another development is to include basic clinical data such as lab data, to improve risk adjustments, recognizing that such data exist alongside administrative data in many healthcare systems. The third direction is the conversion of ICD-9-CM based AHRQ PSIs to ICD-10, which most European countries use, with country-specific modifications (e.g., ICD-10-AM for Australian modification and ICD-10-GM for German modification).

These improvements, combined with advancements in administrative databases and computing technologies, will make AHRQ PSIs more useful in patient safety research in the future.

---

## Patient Safety Research Using Administrative Data: Potentials and Limitations

Administrative data-based patient safety research started with a very simple expectation: to flag the infrequent cases with potential patient safety concerns in the large volume of claims in order to guide further, in-depth investigation. As administrative data sources became more available and screening algorithms improved, researchers began to produce a variety of estimates and statistics and test various hypotheses related to patient safety. More recently, attempts are being made to create safety performance reports from administrative data for individual providers or healthcare systems, study variations across regions, and track progress over time. Previous sections have touched on many examples of such work. This section offers a more

detailed review of the types of patient safety studies, with examples, that administrative data can support and their limitations.

### **Screen Patient Safety Events for In-depth Examination**

First and foremost, AHRQ PSIs, the global trigger tools, and most screening algorithms, are considered *indicators*, not definitive measures, of patient safety concerns. These indicators are proposed to screen claims data for adverse events and to guide subsequent medical record reviews to determine whether safety concerns exist. AHRQ PSIs, for example, enable institutions to quickly and easily identify a manageable number of medical records for closer scrutiny. Ackroyd-Stolarz et al. (2014) developed an algorithm to screen the discharge abstract database of a Nova Scotia hospital for fall-related injuries. They compared cases identified in administrative data against cases identified in structured medical record review, finding that administrative data could identify fall-related injuries with sensitivity of 96 % and specificity of 91 %. Their work provided the hospital with a powerful tool to locate records for patients with fall-related injuries, explore causes, and search for solutions to the problem.

Screening cases of patient safety concerns is especially advantageous when the targeted events are rare. For example, it is not likely that one hospital provides enough data to study patterns, causes, or circumstances of foreign objects left in during surgery, because the events occur in less than 1 in 10,000 surgeries (Zhan and Miller 2003). Screening claims with AHRQ PSIs could quickly identify such rare events, and associated medical records could be obtained and abstracted for in-depth analysis. This two-step approach is particularly useful for individual providers or health systems in their search for localized safety lapses and improvement strategies.

### **Epidemiological Study**

A large proportion of administrative data-based patient safety research is aimed at discovering the

epidemiology of patient safety events, categorizing the events, assessing the prevalence, and understanding the causes and impacts, following the general framework and methodologies discussed earlier.

### **Prevalence of Patient Safety Events**

Because administrative data covers large populations, they are often the only available data sources to estimate national or state rates of patient safety events. The National Healthcare Quality Reports (AHRQ 2013), released annually, include, for example, the rate of postoperative sepsis based nationwide inpatient claims and the rates of ambulatory care visits due to adverse events based on the National Ambulatory Medical Care Survey and the National Hospital Ambulatory Medical Care Survey. The Medicare Current Beneficiary Survey, the Medical Expenditure Panel Survey, the National Ambulatory Medical Care Survey, and the National Hospital Ambulatory Medical Care Survey have been used to examine the prevalence of inappropriate use of medications in the United States (e.g., Zhan et al. 2001).

Similar studies on the prevalence of patient safety events are numerous in medical literature, covering all settings of care and types of problems. A more recent example is a study conducted by Owens et al. (2014). By examining claims of hospitalizations and ambulatory surgical visits for infections following ambulatory surgery, the authors were able to estimate the incidence of surgical site infections after ambulatory surgery procedures, highlighting safety concerns in the fast-growing outpatient surgery centers in the United States.

### **Causes of Patient Safety Events**

Many administrative data-based studies address the causes and circumstances of patient safety events. Gandhi et al. (2006) intended to find out how missed and delayed diagnoses in the outpatient setting led to patient injuries. For their purpose, the authors chose closed malpractice claims from four malpractice insurance companies. They selected 181 claims where patients sued doctors for injuries stemmed from

diagnosis errors and had a team of doctors review the closed documents, including statement of claims, depositions, interrogatories, reports of internal investigations, root cause analyses, expert opinions on both sides of the litigations, medical records, and other documents in the closed file to determine what kind of errors happened and what were the possible causes. They found that failure to order appropriate diagnostic tests, failure to create a proper follow-up plan, and failure to obtain adequate history or perform adequate physical examination (55 %, 45 %, and 42 %, respectively) were the leading types of diagnosis errors that resulted in the malpractice cases.

Zhan et al. (2008) examined warfarin-related medication errors voluntarily reported to the MEDMARX database. By tabulating and cross-tabulating coded variables in a cascading way and screening open-ended narratives in selected reports, the authors were able to construct a comprehensive understanding of errors in warfarin prescriptions and administration in hospitals and clinics. They found that, in outpatient settings, 50 % of errors in warfarin medication occurred in pharmacies and 50 % were intercepted by a pharmacist, indicating the critical role of pharmacists in helping patients with warfarin use.

### **Impact of Patient Safety Events**

Once patient safety events are identified with an acceptable level of validity in administrative data, it is relatively easy to examine the impacts of the events on various patient and social outcomes identifiable in the data. Using AHRQ PSIs, Zhan and Miller (2003) screened nationwide hospital claims and estimated the impacts of the selected patient safety events on length of stay, charges, and mortality. The authors found that postoperative sepsis, for example, extended hospital stay by about 11 days, added \$58,000 extra charges to the patients' hospital bills, and increased the in-hospital mortality rate by 22 %. In another study, Zhan et al. (2006) showed that when a case of postoperative sepsis occurred, Medicare actually paid \$9,000 extra. Taking the two studies together, it is easy to see that, once a postoperative

sepsis occurs, a hospital loses financially, establishing a case for collaboration among hospitals, payers, and patients or patient advocates to reduce postoperative sepsis. This type of study is common in health services research literature.

### **Interventions and Policies to Improve Safety**

Administrative data have been used to evaluate many system-wide interventions aimed at improving patient safety. Many studies have been conducted in the United States, Canada, and the United Kingdom, for example, to evaluate how various levels of nurse staffing, different staffing models, and nursing hours affect patient safety, by linking safety estimates from hospital claims or abstracts to nurse staffing data from hospital surveys. Rafferty et al. (2007) did such a study using data from 30 English hospital trusts. They used data from three sources: hospital structure (e.g., size and teaching status) from hospital administrative databases; patient outcomes, specifically, patient mortality and failure to rescue, from hospital discharge abstracts; and data on nursing staffing and nurse job satisfaction from surveys of the participating hospitals. Their finding that higher patient-to-nurse ratios were associated with worse patient outcomes could help hospitals plan their nurse staffing.

A study by Dimick et al. (2013) is an example of how administrative data can be useful to evaluate national health policies. Starting in 2006, CMS has restricted coverage of bariatric surgery to hospitals designated as centers of excellence by two major professional organizations. The authors wanted to explore if such coverage policy change improved patient safety as it intended. It would be difficult to design a study based on primary data collection or data sources other than nationwide administrative data to evaluate this policy. Using claims from 12 states covering 2004–2009, Dimick et al. (2013) were able to estimate risk-adjusted rates of complications and reoperations of bariatric surgery before versus after the implementation of the national policy restricting coverage, finding that the policy has had no impact with regard to patient safety.

## Public Reporting on Patient Safety Events

Using administrative data to measure patient safety of individual providers has been controversial. However, administrative data-based patient safety measures are increasingly used to profile hospitals and to support pay-for-performance programs. Ten of AHRQ PSIs are endorsed by the National Quality Forum as valid measures for public reporting. Many US states have used these measures as components of their hospital quality reports. CMS annually calculates seven AHRQ PSI rates as parts of public-reported outcome measures based on claims and administrative data, aimed at increasing the transparency of hospital care, providing information for consumers choosing care, and assisting hospitals in their quality improvement efforts.

There are many legitimate arguments against such use, such as coding differences across institutions, lack of specificity and sensitivity in the safety indicators, and lack of sufficient confounding adjustments, to list a few. These reasons raise some doubts whether differences between hospitals in administrative data-based patient safety event rates reflect true differences in patient safety. Because of these limitations, public reporting of such rates for institutions and

regions may lead to contentions over technicalities rather than facilitate quality improvement. Developers of AHRQ PSIs and similar administrative data-based indicators in general have expressed caution with regard to the use of the indicators for public reporting at an institutional level. Health services researchers must exert similar caution when using these measures as hospital performance measures in their research.

## Advantages and Challenges in Administrative Data-Based Patient Safety Research

Table 5 summarizes the major advantages and limitations of administrative data-based patient safety research.

The greatest advantage is that the data already exist and are mostly computerized, and the research effort requires properly acquiring the data, creating valid screening algorithms, and conducting robust analysis. Administrative data usually cover large populations, allowing estimations at county, state, or national levels and comparisons across different subpopulations. Many administrative databases allow linkage of patient records from multiple settings and over time, and researchers can construct large retrospective

**Table 5** Advantages and disadvantages of administrative data for patient safety research

Advantages of administrative data	Disadvantages of administrative data
Already collected for administrative purposes and therefore no additional costs of collection (besides data acquisition and cleaning costs)	Information collected is restricted to data required for administrative purposes
Large coverage of population of interest allowing estimation and comparison at regional and national levels	Collection process does not follow any research design, protocol, or procedure; lack of researcher control over content
Collection process not intrusive to target population	Algorithms, triggers, or indicators with variable validity, subject to coding errors and coding variation across institutions
Regularly, continuously updated	Claims, abstracts, and surveys lack contextual, clinical information, while malpractice claims and spontaneous reports lack data on denominator or population at risk
Mostly computerized	Results often statistically significant but clinically meaningless
Can be linked to form individual patient's complete healthcare experiences	
Malpractice claims and spontaneous reports contain rich contextual data not available elsewhere	



cohorts that mimic a prospective study design and test a wide range of hypotheses from risk factors to potential interventions.

The greatest limitation lies with the fact that the data were not collected with a research purpose, study protocol, or quality assurance procedure. Researchers have to creatively repurpose the data to meet their research needs and make great efforts in methodology design to minimize potential biases.

As discussed earlier, the most critical task of administrative data-based patient safety research is to design valid patient safety screening algorithms or indicators. Most of the indicators developed to date have relied on coded data in the administrative databases. Using ICD-9-CM codes as examples, many concerns exist. First, researchers can only find events for which there are corresponding ICD-9-CM codes. Second, there may be a substantial number of coding errors, due to misunderstanding of codes, or errors by physicians and coders, or miscommunications between them. Third, coding is very likely to be incomplete because of limited slots for coding secondary diagnoses and other reasons. Fourth, assignment of ICD-9-CM codes is variable because of the absence of precise clinical definitions and context. Last but not least, diagnoses are not dated in most administrative data systems, making it difficult to determine whether a secondary diagnosis occurs prior to admission (i.e., a comorbid disease) or during a hospitalization (i.e., a complication or medical error).

Administrative data have been repeatedly shown to have low sensitivity but fair specificity in identifying patient safety events. Focusing on specific adverse events for specific patient populations, as is built into the AHRQ PSIs, improves specificity appreciably. But, in most cases, researchers have to work with indicators that have modest validity in their research.

Lack of clinical details is another major limitation of most administrative data such as claims and discharge abstracts. Of special concern is the severity of illness that affects patient outcomes and conceivably affects the likelihood of patient safety events. Analyses of outcomes and risk factors associated with patient safety events are

limited to variables available from administrative data. On the other hand, malpractice claims and spontaneous medical error reports contain extensive details on specific events, but the denominator populations (i.e., patients at risk for those reported events) are unknown, severely limiting the data's ability to support estimation and hypothesis testing research.

There are also many analytical challenges. The sheer size of administrative data can give the illusion of great precision and power. Often times the differences found are statistically significant but of little clinical meaning. Coupled with missing important confounding variables and difficulty in choosing correct statistical models that fit the data, clinically insignificant but statistically significant results could lead to biased inferences and erroneous conclusions. Health services researchers must bear in mind these limitations when designing their administrative data-based patient safety studies and must interpret the results with full acknowledgment of these limitations.

---

## References

- Ackroyd-Stolarz S, Bowles SK, Giffin L. Validating administrative data for the detection of adverse events in older hospitalized patients. *Drug Healthc Patient Saf.* 2014;13(6):101–8.
- Agency for Healthcare Research and Quality (AHRQ). 2013 National healthcare quality report. <http://www.ahrq.gov/research/findings/nhqrdr/nhqr13/2013nhqr.pdf>. Accessed 1 Sept 2014.
- Agency for Healthcare Research and Quality (AHRQ). Patient safety indicators. [http://www.qualityindicators.ahrq.gov/Modules/psi\\_resources.aspx](http://www.qualityindicators.ahrq.gov/Modules/psi_resources.aspx). Accessed 1 Sept 2014.
- Budnitz DS, Lovegrove MC, Shehab N, et al. Emergency hospitalizations for adverse drug events in older Americans. *N Engl J Med.* 2011;365(21):2002–12.
- Centers for Medicare and Medicaid Services (CMS). CMS data navigator. [http://www.cms.gov/Research-Statistics-Data-and-Systems.html](http://www.cms.gov/Research-Statistics-Data-and-Systems/Research-Statistics-Data-and-Systems.html). Accessed 1 Sept 2014.
- Dimick JB, Nicholas LH, Ryan AM, et al. Bariatric surgery complications before vs after implementation of a national policy restricting coverage to centers of excellence. *JAMA.* 2013;309(8):792–9.
- Gandhi T, Kachalia A, Thomas E, et al. Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims. *Ann Intern Med.* 2006;145:488–96.

- Griffin FA, Resar RK. IHI global trigger tool for measuring adverse events (Second Edition). IHI innovation series white paper. Cambridge, MA: Institute for Healthcare Improvement; 2009.
- Healthcare Cost and Utilization Project (HCUP). <http://www.hcup-us.ahrq.gov/>. Accessed 1 Sept 2014.
- Iezzoni LI, Daley J, Heeren T, et al. Using administrative data to screen hospitals for high complication rates. *Inquiry*. 1994;31(1):40–55.
- Kohn LT, Corrigan JM, Donaldson M, et al. To err is human: building a safer health system. Washington, DC: Institute of Medicine; 1999.
- Owens PL, Barrett ML, Raetzman S, et al. Surgical site infections following ambulatory surgery procedures. *JAMA*. 2014;311(7):709–16.
- Quan H, Drösler S, Sundararajan V, et al. Adaptation of AHRQ patient safety indicators for use in ICD-10 administrative data by an international consortium. In: Henriksen K, Battles JB, Keyes MA, et al., editors. *Advances in patient safety: new directions and alternative approaches*. Rockville: Agency for Healthcare Research and Quality; 2008.
- Rafferty AM, Clarke SP, Coles J, et al. Outcomes of variation in hospital nurse staffing in English hospitals: cross-sectional analysis of survey data and discharge records. *Int J Nurs Stud*. 2007;44(2):175–82.
- Romano PS, Geppert J, Davies S, et al. A national profile of patient safety in US hospitals based on administrative data. *Health Aff*. 2003;22(2):154–66.
- Wang X, Hripcsak G, Markatou M, et al. Active computerized pharmacovigilance using natural language processing, statistics and electronic health records: a feasibility studies. *JAMIA*. 2009;16:328–37.
- Zhan C, Miller M. Excess length of stay, costs, and mortality attributable to medical injuries during hospitalization: an administrative data-based analysis. *JAMA*. 2003;190(4):1868–74.
- Zhan C, Sangl J, Bierman A, et al. Inappropriate medication use in the community-dwelling elderly: findings from 1996 Medical Expenditure Panel Survey. *JAMA*. 2001;286(22):2823–9.
- Zhan C, Friedman B, Mosso A, et al. Medicare payment for selected adverse events under the prospective payment system: building the business cases for investing in patient safety improvement. *Health Aff*. 2006;25(5):1386–93.
- Zhan C, Smith SR, Keyes MA, et al. How useful are voluntary medication error reports? The case of warfarin-related medication errors. *Joint Comm J Qual Patient Saf*. 2008;34(1):36–44.
- Zhan C, Elixhauser A, Richards C, et al. Identification of hospital-acquired catheter-associated urinary tract infections from Medicare claims: sensitivity and positive predictive value. *Med Care*. 2009;47(3):364–9.



# Health Services Information: Personal Health Records as a Tool for Engaging Patients and Families

# 12

John Halamka

## Contents

Introduction .....	266
A Short History of Personal Health Records .....	266
Policies .....	267
Products in the Marketplace .....	269
The Regulatory Environment: ARRA/HITECH, the HIPAA Omnibus Rule, and FDASIA .....	270
Myths .....	271
Digital Divide .....	272
Data Standards .....	272
The Role of Personal Medical Devices .....	273
Research: OpenNotes, ICU Harm Reduction, Care Plans, and Clinical Trials .....	273
Conclusion .....	276
References .....	276

## Abstract

Personal Health Records have evolved from stand alone websites requiring manual entry of data to automated mobile applications fully integrated into care management workflow. Technology issues such as interoperability, security, and patient identification have

matured. Policies such as who can see what for what purpose have been enumerated. Regulations now require a deeper level of interaction between care teams and patients. Many myths about the risks of engaging patients and families have been shattered. Research continues to expand the scope of information shared with families, enhance usability of patient facing applications, and improving the utility of solutions automating patient/provider workflow.

---

J. Halamka (✉)  
Department of Emergency Medicine, Harvard Medical School and Beth Israel Deaconess Medical Center, Boston, MA, USA  
e-mail: [jhalamka@bidmc.harvard.edu](mailto:jhalamka@bidmc.harvard.edu)

## Introduction

A key enabler to delivering safe, high-quality, efficient care is engaging patients and families by sharing healthcare records, codeveloping plans of treatment, and communicating preferences for care among the entire care team.

Over the past 20 years, patient portals, personal health records, electronic consumer education resources, wellness devices, and health-focused social networks have offered more transparency, shared decision-making, and communication than ever before.

This chapter examines the history of technologies that empower patients and families while also identifying important foundational policies and speculating how future innovations will provide even greater functionality. The chapter also reviews the evolving regulatory environment and discusses the impact of US national “Meaningful Use” requirements on the adoption of new tools.

---

## A Short History of Personal Health Records

Personal health records (PHRs) have the potential to make patients the stewards of their own medical data. PHRs may contain data from payer claims databases, clinician electronic health records, pharmacy-dispensing records, commercial laboratory results, and personal medical device data. They may include decision support features, convenience functions such as appointment making/referral requests/medication refill workflow, and bill paying.

Early personal health records were deployed at Beth Israel Deaconess Medical Center (PatientSite), Children’s Hospital (Indivo), and Palo Alto Medical Clinic (MyChart) in the late 1990s and early 2000s. In the mid-2000s, direct-to-consumer vendors such as Microsoft (HealthVault) and Google (Google Health) offered products. Since that time, most electronic health record (EHR) vendors (Epic, Meditech, Cerner, eClinicalWorks, Athena) have included patient portals in their products.

In 1999, a group of clinicians and patient advocates in New England suggested that Beth Israel Deaconess Medical Center (BIDMC) should share all of its electronic records with patients, since all healthcare data ultimately belongs to the patient. In 2000, BIDMC went live with a hospital-based personal health record, PatientSite (<http://www.patientsite.org>). PatientSite includes full access to problem lists, medications, allergies, visits, laboratory results, diagnostic test results, and microbiology results from three hospitals and numerous ambulatory care practices. In addition to these hospital and ambulatory clinic-provided data, patients can amend their own records online, adding home glucometer readings, over-the-counter medications, and notes. Secure patient-doctor messaging is integrated into the system. Convenience functions such as appointment making, medication renewal, and specialist referral are automated and easy to use. Clinical messaging is the most popular feature, followed by prescription renewals and followed by appointment making and referrals.

In 1998 researchers at the Children’s Hospital Informatics Program (CHIP) at Children’s Hospital Boston developed the concept of the Indivo Personally Controlled Health Record in a planning grant and began implementation in 1999. Critical to the success of the model, the code base of Indivo has always been open source, the application programming interface (API) is fully published, and all communication/messaging protocols adhere to freely implementable standards. Indivo enables patients to maintain electronically collated copies of their records in a centralized storage site. Access, authentication, and authorization all occur on one of several available Indivo servers, which are also responsible for encryption of the record. Individuals decide who can read, write, or modify components of their records.

In 1999, Epic Systems, an established vendor of EHR systems, decided to develop a patient portal, which they called MyChart. The Palo Alto Medical Foundation (PAMF) worked with Epic to develop the functionality requirements for a PHR that was integrated with their EHR. PAMF became the first customer of MyChart, which was implemented at the end of 2000.

MyChart enables the patient to review their diagnoses, active medications, allergies, health maintenance schedules, immunizations, test results, radiology results, appointments, and demographics. In many cases, relevant health educational resources are automatically linked to key terms or phrases in the patient's medical record, such as a diagnosis of diabetes. In addition, patients can communicate with the physician office to request an appointment, request a prescription renewal, update demographic information, update immunization status, or update a health maintenance procedure. The patient can also request advice from an advice nurse or from their own physicians.

Based on the success of these early adopters, many electronic health record companies began offering patient access to electronic records in the late 2000s. As is discussed below, the Federal HITECH Meaningful Use program now requires that patients be able to view, download, and transmit their medical records, accelerating market deployment of personal health record functionality.

---

## Policies

As personal health record technology was deployed, many novel policy questions arose. What information should be shared and when? Who should have access? Should parents have access to the records of their adolescent children? Over time, many best practices have evolved which have answered these questions.

Although the Health Insurance Portability and Accountability Act of 1996 (HIPAA) mandated that patients have access to their medical records, it did not require the release of data electronically. The HIPAA Omnibus Rule of 2013 does require electronic access, but it does not specify how quickly releases should occur. Should a cancer diagnosis be revealed to a patient in real time on a website or wait for a personal conversation with a physician?

At BIDMC, the majority of the record is shared with the patient immediately with minor exceptions, since it is the patient's data.

A small number of reports are delayed to enable a discussion between provider and patient to occur first. The Commonwealth of Massachusetts has specific regulatory restrictions on the delivery of HIV test results, so they are not shown on PatientSite. The tests and their delays are summarized below:

CT scans (used to stage cancer) 4 days  
 PET scans (used to stage cancer) 4 days  
 Cytology results (used to diagnose cancer)  
 2 weeks  
 Pathology reports (used to diagnose cancer)  
 2 weeks

HIV diagnostic tests: never shown

- Bone marrow transplant screen, including:  
 HIV-1 and HIV-2 antibody  
 HTLV-I and HTLV-II antibody  
 Nucleic acid amplification to HIV-I (NHIV)
- HIV-1 DNA PCR, qualitative
- HIV-2 and Western blot. Includes these results:  
 HIV-2 AB and EIA  
 HIV-2 and Western blot
- HIV-1 antibody confirmation. Includes these results:  
 Western blot  
 Anti-P24  
 Anti-GP41  
 Anti-GP120/160

We want the patient to own and be the steward of their own data, but we also want to support the patient/provider relationship and believe that bad news is best communicated in person. Over time, it is likely that even these delays and restrictions will be removed, making all data instantly available to the patient. When the wife of the author of this chapter was diagnosed with breast cancer in 2011, she wanted to see her pathology results immediately, even if they were bad news. In the future, the patient and provider may agree on data-sharing preferences as part of establishing a primary care relationship.

Other issues that arose during early experiences with personal health records included the access granted to adolescents and their parents.

As more and more practices and hospitals are making patient portals available to their patients, providers of adolescent patients are encountering a major hurdle: how to handle confidential adolescent information.

While adult patients generally maintain full personal control of their personal health record (PHR), adolescent PHRs are anything but personal. Adolescents rarely have full control of their record, but instead rely on parents and guardians to share control. The details around this shared access changes over time, depending on developmental and age-appropriate considerations, as well as guardianship arrangements.

The biggest challenge then becomes how to protect the adolescent's legal right to privacy and confidentiality within this hybrid/proxy-control model. Many medical encounters with adolescents come with the verbal assurance that what they tell us will (under most circumstances) remain entirely confidential, meaning we will not discuss personal health information pertaining to reproductive health, sexually transmitted diseases, substance abuse, and mental health with their parents or anyone else without their consent. As it turns out, this type of confidential information is pervasive through most EHRs.

Children's Hospital Boston spent a lot of time thinking about this issue and adolescent access to our patient portal and ultimately developed a custom-built solution to meet our and our patients' needs.

Their approach is built around differential access to the patient portal with the goal of mirroring current clinical practice and works as follows:

**Access to the patient portal:** Separate accounts are created for the patient and parent(s) that are linked. The parent has sole access to the patient's portal until the patient turns 13, at which point both the parent and the patient can have access. They chose 13 years as the cutoff based on a number of factors, including developmental maturity and other precedents at their institution based on their policies. At 18 years, the patient becomes the sole owner of the portal account, and Children's deactivates

the parent's link (unless they receive court documents stating that the parent remains the medical guardian).

**Health information contained in the patient portal:** Children's has identified and tagged certain information from their EHR that they consider sensitive, such as labs related to pregnancy, sexually transmitted illnesses, genetic results, select confidential appointments, and potentially sensitive problems and medications. This information is currently filtered from both parent and adolescent accounts, but in the near future, the sensitive information will flow to the adolescent account, but not to the parent account. So, even if a patient is less than 13 years, the parent would not have access to this information.

This solution does take a lot of time and effort, but best replicates the current clinical practice. Many current PHR applications in the marketplace do not allow for this type of differential access and only enable full proxy access.

Alternative solutions include the following:

1. Shared access for patient and parent, but filtering of sensitive information. One could then choose the age at which patients would gain access without worrying about the parent seeing sensitive information at any age. This makes the age at which the patient obtains access, whether it is 10 or 13 years, less important. Unfortunately, this option restricts adolescent access to confidential information and creates a fragmented and incomplete record.
2. Adolescent access only. This is trickier, because choosing the appropriate age when parental access is discontinued is difficult and may vary depending on patient characteristics. Many practices choose 12 or 13 years. However, if sensitive information is not being filtered, there may be an occasional 11-year-old with a sexually transmitted infection. Also, some parents object to being cut off from their child's medical information, and many play an important role in supporting their adolescent children and guiding them through healthcare decisions.

The issues and solutions involved with adolescent PHRs are certainly complex and will continue to evolve over time. However, I am hopeful that PHRs will start incorporating the unique needs of the adolescent population in the near future, allowing both parents and adolescents to share responsibility and engage in their healthcare.

---

## Products in the Marketplace

Over the nearly two decades that personal health records have been deployed, there have been four basic models.

**Provider-hosted patient portal to the electronic health record:** In this model, patients have access to provider record data from hospitals and clinics via a secure web portal connected to existing clinical information systems. Examples of this approach include the PatientSite and MyChart applications described above. The funding for provider-based PHRs is generally from the marketing department since PHRs are a powerful way to recruit and retain patients. Also, the Healthcare Quality Department may fund them to enhance patient safety since PHRs can support medication reconciliation workflows. Kaiser's implementation does not distinguish between the personal health record and electronic health record. Instead they call it a patient-/provider-shared electronic health record.

**Payer-hosted patient portal to the payer claims database:** In this model, patients have access to administrative claims data such as discharge diagnoses, reimbursed medications, and lab tests ordered. Few payer-hosted systems contain actual lab data, but many payers are now working with labs to obtain this data. Additionally, payers are working together to enable the transport of electronic claims data between payers when patients move between plans, enhancing continuity of care. The funding for payer-based PHRs is based on reducing total claims to the payer through enrollment of patients in disease management programs and

enhancing coordination of care. Many Blue Cross affiliates have made such sites available.

**Employer sponsored:** In this model, employees can access their claims data and benefit information via a portal hosted by an independent outsourcing partner. The funding for employer-based personal health records is based on reducing total healthcare costs to the employer through wellness and coordination of care. A healthy employee is a more productive employee. Keas is an example of an employer-sponsored employee engagement for health application.

**Vendor hosted:** Several vendors serve as a secure container for patients to retrieve, store, and manipulate their own health records. Microsoft's HealthVault includes uploading and storage of records as well as a health search engine. Google offered such services from 2007 to 2012, but discontinued the service because of lack of adoption. Humetrix is an example of a consumer-centered technology vendor, focused on mobile apps and healthcare information exchange. The business model for these PHRs is generally based on attracting more users to advertising-based websites, although the PHR itself may be advertising free. Vendor-hosted PHRs include HITECH-mandated privacy protections and must sign business associate agreements and agree to keep data private.

Here is the press release from Beth Israel Deaconess, describing the availability of HealthVault to its patients, which illustrates the value proposition communicated to the patients:

**BOSTON:** Beth Israel Deaconess Medical Center (BIDMC) is expanding options for users of its secure PatientSite portal by joining forces with Microsoft HealthVault to offer a new way to safely exchange medical records and other health data.

The affiliation follows an earlier commitment to offer a similar service through Google Health.

"We believe that patients should be the stewards of their own data," says John Halamka, MD,

BIDMC's chief information officer. BIDMC's PatientSite is wonderful if all care is delivered at BIDMC. However, many patients have primary care doctors, specialists, labs, pharmacies, and nontraditional providers at multiple institutions.

"Our vision is that BIDMC patients will be able to electronically upload their diagnosis lists, medication lists and allergy lists into a HealthVault account and share that information with health care providers who currently don't have access to PatientSite."

PatientSite, which currently has more than 40,000 patient users and 1,000 clinicians, enables patients to access their medical records online, securely email their doctors, make appointments, renew medications, and request referrals.

HealthVault is designed to put people in control of their health data. It helps them collect, store, and share health information with family members and participating health care providers, and it provides people with a choice of third-party applications and devices to help them manage things such as fitness, diet, and health.

HealthVault also provides a privacy- and security-enhanced foundation on which a broad ecosystem of providers – from medical providers and health and wellness device manufacturers to health associations – can build innovative new health and wellness solutions to help put people in increased control of their and their family's health.

"The end result will be when patients leave the BIDMC area or see a provider outside the area they can have all their medical data located in one safe place," adds Halamka.

---

## **The Regulatory Environment: ARRA/ HITECH, the HIPAA Omnibus Rule, and FDSIA**

The American Recovery and Reinvestment Act (ARRA) of 2009 included the HITECH provisions which launched the national Meaningful Use program. Meaningful Use includes certification for products, ensuring they are good enough, and attestation for clinicians that they are using the technology wisely.

In stage 1 of Meaningful Use, vendor software was certified to provide basic health information access to patients. Providers were optionally able to attest to use of personal health records as part of meeting criteria for stimulus payment. In stage 2 of Meaningful Use, use of personal health record technology became a mandatory part of attestation. The three provider requirements related to PHRs include:

- Providers must offer online access to health information to more than 50 % of their patients with more than 5 % of patients actually accessing their information.
- More than 5 % of patients must send secure messages to their provider.
- Providers must use the EHR to identify and provide educational resources to more than 10 % of patients.

Although some institutions have offered personal health records for many years, others have not yet established the workflow, created the policies, or experienced the cultural changes that are foundational to provider/patient electronic interaction. Many organizations have suggested that requiring actual use of the personal health record by the patient is beyond provider control and thus is unfair.

Beth Israel Deaconess has already achieved patient participation rates of 25 % for record viewing and 15 % for secure messaging without significant advertising or educational effort. Patients find value in the timeliness and convenience of these transactions, so participate enthusiastically. Admittedly, BIDMC had 15 years to refine the application, modify medical staff bylaws to require PHR use, and overcome some of the doubts and myths described below.

In addition to the Meaningful Use requirements, the HIPAA Omnibus Rule expands an individual's rights to receive electronic copies of his or her health information and to restrict disclosures to a health plan concerning treatment for which the individual has paid out of pocket in full. Many healthcare organizations are struggling with the self-pay disclosures workflow, since modifying data flows based on how the patient pays is not



currently supported by commercial EHR products. There are also ongoing national efforts to refine the Omnibus Rule language for “accounting of disclosures,” when a patient requests a list of all who have accessed or received copies of their record. Implementing such accounting for all disclosures including treatment, payment, and operations requires capabilities not present in most commercial EHR products.

The Food and Drug Administration issued a report in April 2014 outlining the Food and Drug Administration Safety and Innovation Act (FDASIA) regulatory framework that is relevant to personal health records because of the increasing popularity of using mobile devices to access health-related resources. Mobile devices will be discussed in detail later in this chapter.

The FDA stratified mobile devices/apps into three categories:

Administrative apps – an application that reminds you about an appointment, describes costs/benefits such as co-pays, or helps you find a doctor.

Wellness apps – an application that measures your daily exercise, suggests weight loss strategies, or offers healthcare coaching via a social network.

Medical devices – an application that measures a body parameter such as pulse, blood pressure, or EKG and may offer therapeutic suggestions based on directly gathered diagnostic data.

The FDA reaffirmed its intent to regulate Medical devices and not administrative apps/wellness apps.

It is unlikely that the FDA will regulate personal health records in the near future, but it will likely regulate the apps and devices which collect patient telemetry and transmit it to personal health records.

---

## Myths

Many providers and patients have concerns about the impact of increased electronic data sharing and automated workflows. After nearly 20 years of

experience with personal health records, it is clear that most of those concerns have not appeared in practice.

Providers were concerned that sharing electronic health records would result in more assertions of malpractice as patients found errors in their records. At BIDMC and other Harvard-associated hospitals, the opposite has been true. Informed and engaged patients do find errors and work with their providers to correct inaccuracies before harms occur. Malpractice assertions decrease when personal health records are deployed.

Providers were concerned that they would be overwhelmed with secure email or other electronic requests from patients. Electronic requests have replaced phone calls and have reduced time spent on “phone tag” and accelerated the resolution of simple administrative matters than can be delegated to others.

Patients were concerned that increased electronic access would create new security risks. While it is true that the Internet is increasingly a mire of viruses and malware, keeping electronic data centrally managed on secure servers is less risky than exchanging paper copies, storing PDFs on laptops, or exchanging electronic copies on USB flash drives because centrally stored information can be better audited and controlled.

Patients and providers were concerned that more transparency could jeopardize the clinician/patient relationship because of misunderstandings in the interpretation of electronic health records. Instead, providers have been careful to write comprehensible summaries with fewer abbreviations because they know a patient is likely to read their work.

There have been lessons learned along the way. Sharing inaccurate or confusing data with patients does not add value. For example, administrative billing data is a coded summary of the clinical care that lacks perfect specificity and time references, i.e., just because you had a diagnosis of low potassium 5 years ago does not imply it is a problem today.

Thus, we must be thoughtful about what data is sent to PHRs and how that data is presented to patients. The problem list is useful clinical

information as long as clinicians keep it current. BIDMC removes ICD-9 administrative data feed so that the clinician's problem list is the only data which populates the patient view. Also, BIDMC improved its problem list functionality so that it maps to a standardized terminology, SNOMED CT, enabling BIDMC to provide medical information and decision support based on a controlled vocabulary instead of just free text.

As long as the PHR software is usable and the data presented is relevant, supplemented by educational materials, the experiencing of provider/patient data sharing will be positive.

---

## Digital Divide

As we offer more electronic resources to patients and encourage the use of mobile technology and home medical devices, we must be careful not to create a digital divide – the technology haves and have nots. In the Boston area, there are many academic and technology professionals with fast Internet connections and the latest mobile devices. There are also Medicaid patients without the funding to purchase personal devices and those who feel technology requires expertise beyond their comfort zone. Research done in the Boston area discovered that the large majority of Medicaid patients have phones capable of receiving text messages and most patients have access to the Internet at work, at a local library, or a community center. We must engineer our personal health records so they run anywhere on anything, but also protect privacy by not leaving behind cached data that could be viewed inappropriately. PatientSite and most vendor applications are web based so they can be accessed regardless of location or platform, with specific protections to ensure data is encrypted and not stored in web browsers. Engineering for those with disabilities, failing eyesight, or limited computer skills is also essential.

---

## Data Standards

The HITECH Meaningful Use program requires the use of specific standards for transition of care summary transmission, public health reporting,

and e-prescribing. Although the standards for personal health records are not explicitly stated, it is logical that personal health records should mirror the standards used in electronic health records themselves. Standards can generally be lumped into three different categories.

**Vocabulary** – the terminology used in each part of the record to communicate meaning between sender and receiver. The Meaningful Use Common Data Set requires LOINC codes for labs, RxNorm codes for medications, SNOMED CT for problem lists, CVX for immunization names, and ISO 639–2 for primary language. The same standards should be used in personal health records and medical devices connecting to personal health records. Mappings to patient friendly terminology, available for the National Library of Medicine's Value Set Authority Center, are likely to be helpful to patients.

**Content** – the container used to package a collection of data to be transported between a sender and receiver. The Consolidated Clinical Document Architecture (CCDA) is used for all EHR transition of care summaries and is appropriate to use for sending data to PHRs and collecting data from patients. Medical devices may additionally use the IEEE 11073 standard to transfer data to and from PHRs.

**Transmission** – the secure protocol to transport content from one place to another without modification or interception. Meaningful Use stage 2 requires the Direct Protocol (SMTP/SMIME or SOAP/HTTPS) to be used for transport. These standards are also appropriate for personal health records and medical devices.

As standards become increasingly constrained, ease of interfacing improves and the value of interoperable products increases. Ideally, Meaningful Use certification should create an ecosystem of personal health record products, leveraging the liquidity of data to foster innovation. Later stages of Meaningful Use likely encourage “modular” EHR and PHR products that plug into large commercial systems through the use of simple application

programming interfaces (APIs). The April 2014 JASON report, requested by AHRQ and facilitated by MITRE corporation, provides a roadmap for evolution of healthcare apps that expand the use of today's EHRs and PHRs.

---

## The Role of Personal Medical Devices

As Accountable Care Organizations move from fee for service to risk contracts, providers will be reimbursed for keeping patients healthy and not for delivering more care. Personal medical devices that report on patient activities, functional status, and body parameters between clinician visits will be increasingly important.

Such devices include electronic scales for measuring fluid retention in CHF patients, blood pressure measurement for refractory hypertension, glucometers for diabetics, and home spirometry for patients with COPD or asthma.

The current challenge is that home medical devices communicate using proprietary protocols that make interfacing to personal health records and electronic health records very challenging.

The Continua Alliance is a group of 60 companies that collaboratively develops standards for incorporation into products with the goal that devices available at the local drugstore will "plug and play" with the diversity of current EHRs and PHRs without complex engineering or custom software development.

Future stages of Meaningful Use will likely include a requirement for patient-generated data. Payers, providers, and patients will all have incentives to include device from home telemetry in electronic medical records that provide coordinated, optimized care further personalized via access to personal medical devices.

Here's an example. The father of the author of this chapter had multiple sclerosis for 23 years. His mobility declined but there was no easy way to measure that decline. To complicate the situation, he self-medicated with over-the-counter and prescription medications to episodically reduce his symptoms. During personal visits his level of function seemed very high. Imagine that a Fitbit or other home device provided data about

his mobility to an EHR or PHR. It would be clear that on some days he walked 50 ft and other days he walked 5,000 ft. The trend would be clear – fewer good mobility days and more limited function. Care plans, medications, and supportive therapies would be informed by this objective data.

Just as personal computing has evolved from terminals to PCs to mobile smartphones/tablets, it is likely that personal health records will increasingly run on mobile technology with interfaces to home care devices.

---

## Research: OpenNotes, ICU Harm Reduction, Care Plans, and Clinical Trials

When BIDMC's PatientSite was originally released, it included patient access to the entire health record except for the clinic notes a physician wrote about a patient. That changed in 2011 when notes were added via the OpenNotes project. Here's the press release about it.

BOSTON – A Beth Israel Deaconess Medical Center-led study has found that patients with access to notes written by their doctors feel more in control of their care and report a better understanding of their medical issues, improved recall of their care plan, and being more likely to take their medications as prescribed.

Doctors participating in the OpenNotes trial at BIDMC, Geisinger Health System in Danville, PA, and Harborview Medical Center in Seattle reported that most of their fears about an additional time burden and offending or worrying patients did not materialize, and many reported enhanced trust, transparency, and communication with their patients.

"Patients are enthusiastic about open access to their primary care doctors' notes. More than 85 % read them, and 99 % of those completing surveys recommended that this transparency continue," says Tom Delbanco, MD, co-first author, a primary care doctor at BIDMC and the Koplrow-Tullis Professor of General Medicine and Primary Care at Harvard Medical School. "Open notes may both engage patients far more

actively in their care and enhance safety when the patient reviews their records with a second set of eyes.”

“Perhaps most important clinically, a remarkable number of patients reported becoming more likely to take medications as prescribed,” adds Jan Walker, RN, MBA, co-first author and a Principal Associate in Medicine in the Division of General Medicine and Primary Care at BIDMC and Harvard Medical School. “And in contrast to the fears of many doctors, few patients reported being confused, worried or offended by what they read.”

The findings reflect the views of 105 primary care physicians and 13,564 of their patients who had at least one note available during a year-long voluntary program that provided patients at an urban academic medical center, a predominantly rural network of physicians, and an urban safety net hospital with electronic links to their doctors’ notes.

Of 5,391 patients who opened at least one note and returned surveys, between 77 % and 87 % reported OpenNotes made them feel more in control of their care, with 60–78 % reporting increased adherence to medications. Only 1–8 % of patients reported worry, confusion, or offense, three out of five felt they should be able to add comments to their doctors’ notes, and 86 % agreed that availability of notes would influence their choice of providers in the future.

Among doctors, a maximum of 5 % reported longer visits, and no more than 8 % said they spent extra time addressing patients’ questions outside of visits. A maximum of 21 % reported taking more time to write notes, while between 3 % and 36 % reported changing documentation content.

No doctor elected to stop providing access to notes after the experimental period ended.

“The benefits were achieved with far less impact on the work life of doctors and their staffs than anticipated,” says Delbanco. “While a sizeable minority reported changing the way their notes addressed substance abuse, mental health issues, malignancies and obesity, a smaller minority spent more time preparing their notes, and some commented that they were improved.”

“As one doctor noted: ‘My fears? Longer notes, more questions and messages from patients . . . In reality, it was not a big deal.’”

Walker suggests that so few patients were worried, confused, or offended by the note because “fear or uncertainty of what’s in a doctor’s ‘black box’ may engender far more anxiety than what is actually written, and patients who are especially likely to react negatively to notes may self-select to not read them.”

“We anticipate that some patients may be disturbed in the short term by reading their notes and doctors will need to work with patients to prevent such harms, ideally by talking frankly with them or agreeing proactively that some things are at times best left unread.”

“When this study began, it was a fascinating idea in theory,” says Risa Lavizzo-Mourey, MD, president and CEO of the Robert Wood Johnson Foundation, the primary funder of the study. “Now it’s tested and proven. The evidence is in: Patients support, use, and benefit from open medical notes. These results are exciting – and hold tremendous promise for transforming patient care.”

Although PatientSite provides great transparency into ambulatory and inpatient records, the ICU is still an area with limited patient and family engagement. Patient-connected devices in the ICU provide a dizzying array of data but rarely provide an interpretation of that data that is useful to families, especially while making end-of-life decisions. The Moore Foundation recently funded a grant for several hospitals, including BIDMC, to create unique patient dashboards that make the process of care in ICUs more transparent and reduce harms. Here’s an example.

As discussed previously, the father of the author of this chapter had multiple sclerosis for 23 years. He also had myelodysplastic syndrome for 2 years, had 3 myocardial infarctions since 2009, and died in mid-March of 2013.

When the family arrived at his ICU bedside in early March, they spoke with all his clinicians to create a mental dashboard of his progress. It looked something like this

Cardiac – history of 2 previous myocardial infarctions treated with 5 stents. New myocardial infarction resulting in apical hypokinesis and an ejection fraction of 25 %. No further stent placement possible, maximal medical therapy already given

Pulmonary – new congestive heart failure post recent myocardial infarction treated with diuretics, nitroglycerine drip, afterload reduction, upright position, and maximal oxygenation via bilevel positive airway pressure. O<sub>2</sub> saturation in the 90s and falling despite maximal therapy (other than intubation)

Hematologic – failing bone marrow resulting in a white count of 1, a platelet count of 30, and a hematocrit of 20

Neurologic – significant increase in muscle spasticity, resulting in constant agitation. Pain medication requirements escalating. Consciousness fading.

Renal – creatinine rising

Although the family did not have real-time access to his records, they gathered enough data to turn this mental dashboard into a scorecard green, yellow, and red indicators.

Cardiac – red due to irreversible low ejection fraction

Pulmonary – red due to the combination of falling O<sub>2</sub> saturation despite aggressive therapy

Hematologic – red due to lack of treatment options available for myelodysplastic syndrome and an inability to transfuse given the low ejection fraction and congestive heart failure

Neurologic – yellow due to the potential for successful symptom control with pain medications

Renal – yellow due to treatment options available for renal failure

The patient had expressed his wishes in a durable power of attorney for healthcare – do not intubate, do not resuscitate, no pressors, no feeding tubes, and no heroic measures.

From the combination of the dashboard, scorecard, and his end-of-life wishes, it was clear that hospice was the best course of action.

Ideally, all patients and families should have the tools needed to make such decisions regardless of their medical sophistication.

The Moore Foundation project includes an automated ICU dashboard/scorecard for patients and families updated in real time based on data aggregated from the medical record and patient-connected telemetry. The architecture includes a cloud-hosted decision support web service. Hospitals send data in and the web service returns the wisdom of a graphical display.

Although OpenNotes and the Moore Foundation ICU project implement new ways to share data and its interpretation, we still need additional ways to involve patients and families in shared decision-making through the creation of shared care plans. BIDMC created the Passport to Trust initiative, in collaboration with a commercial PHR software vendor. Patients and doctors use a secure PHR website to develop a shared care plan, and then that plan is sent to the EHR using Meaningful Use standards and it is made part of the permanent medical record and integrated into care delivery. This kind of third-party PHR to EHR integration is likely to increase now that Meaningful Use requires EHRs to receive externally generated data. Also, care plan exchange is likely to be part of future stages of Meaningful Use.

An area in which more patient and family engagement could be beneficial is in the area of clinical trial enrollment. Today, most patients are unaware of the new treatments that could provide a cure or breakthrough. Many are willing to enroll in clinical trials but do not know how. Clinicians may be unaware of matching criteria or a patient's suitability for a given trial. BIDMC has worked with a company called TrialX that enables patients and providers to use PHRs and EHRs with innovative electronic connections to clinical trial databases to facilitate the process. Not only can direct patient involvement in clinical trial enrollment accelerate research, it is likely that patient sharing their experiences with other patients will enable new discoveries to be rapidly disseminated for the benefit of all.

## Conclusion

From 1999 to the present, personal health records have transitioned from a research project to the mainstream and are now required by several federal programs. Patients and families increasingly expect access to their records, a role in decision-making, and the convenience of using electronic workflows to manage their care. Consumer platforms continue to rapidly evolve, accelerated by market demand and new interoperability standards incorporated into electronic health records.

As important as the technology has been, the breakthroughs of the past 5 years have been in culture and policy. Clinicians no longer fear sharing the record or participating in secure messaging. There are available policy solutions to tricky problems like sharing adolescent records with their parents.

The next few years will be an important turning point for the medical industry as care becomes increasingly focused on continuous wellness rather than episodic sickness. Patient-generated healthcare data and patient involvement in the entire process is essential to achieving our national and international policy goals for quality, safety, and efficiency. Patients, acting as stewards of their own data, will facilitate data sharing, discovery of new therapies, and innovation as part of a connected learning healthcare system.

## References

- AHIMA e-HIM Personal Health Record Work Group. The role of the personal health record in the EHR. *J AHIMA*. 2005;76(7):64A–D.
- Archer N, Fevrier-Thomas U, Lokker C, McKibbin KA, Straus SE. Personal health records: a scoping review. *J Am Med Inform Assoc*. 2011;18(4):515–22. <https://doi.org/10.1136/amiajnl-2011-000105>. Review.
- Beard L, Schein R, Morra D, Wilson K, Keelan J. The challenges in making electronic health records accessible to patients. *J Am Med Inform Assoc*. 2012;19(1):116–20.
- Bourgeois FC, Taylor PL, Emans SJ, Nigrin DJ, Mandl KD. Whose personal control? Creating private, personally controlled health records for pediatric and adolescent patients. *J Am Med Inform Assoc*. 2008a;15(6):737–43. <https://doi.org/10.1197/jamia.M2865>. Epub 2008 Aug 28.
- Brennan PF, Downs S, Casper G. Project HealthDesign: rethinking the power and potential of personal health records. *J Biomed Inform*. 2010;43 Suppl 5:S3–5. <https://doi.org/10.1016/j.jbi.2010.09.001>.
- Britto MT, Wimberg J. Pediatric personal health records: current trends and key challenges. *Pediatrics*. 2009;123 Suppl 2:S97–9. <https://doi.org/10.1542/peds.2008-17551>.
- Collins SA, Vawdrey DK, Kukafka R, Kuperman GJ. Policies for patient access to clinical data via PHRs: current state and recommendations. *J Am Med Inform Assoc*. 2011;18 Suppl 1:i2–7. <https://doi.org/10.1136/amiajnl-2011-000400>. Epub 2011 Sep 7.
- Council on Clinical Information Technology. Policy Statement—Using personal health records to improve the quality of health care for children. *Pediatrics*. 2009;124(1):403–9. <https://doi.org/10.1542/peds.2009-1005>.
- Forsyth R, Maddock CA, Iedema RA, Lassere M. Patient perceptions of carrying their own health information: approaches towards responsibility and playing an active role in their own health – implications for a patient-held health file. *Health Expect*. 2010;13(4):416–26. <https://doi.org/10.1111/j.1369-7625.2010.00593.x>.
- Goel MS, Brown TL, Williams A, Cooper AJ, Hasnain-Wynia R, Baker DW. Patient reported barriers to enrolling in a patient portal. *J Am Med Inform Assoc*. 2011;18 Suppl 1:i8–12. <https://doi.org/10.1136/amiajnl-2011-000473>. Epub 2011 Nov 9.
- Haggstrom DA, Saleem JJ, Russ AL, Jones J, Russell SA, Chumbler NR. Lessons learned from usability testing of the VA's personal health record. *J Am Med Inform Assoc*. 2011;18 Suppl 1:i13–7. <https://doi.org/10.1136/amiajnl-2010-000082>. Epub 2011 Oct 8.
- Kaelber J. A research agenda for personal health records. *Am Med Inform Assoc*. 2008;15:729–36.
- Kim EH, Stolyar A, Lober WB, Herbaugh AL, Shinstrom SE, Zierler BK, Soh CB, Kim Y. Challenges to using an electronic personal health record by a low-income elderly population. *J Med Internet Res*. 2009;11(4), e44. <https://doi.org/10.2196/jmir.1256>.
- Poulton M. Patient confidentiality in sexual health services and electronic patient records. *Sex Transm Infect*. 2013;89(2):90. <https://doi.org/10.1136/ssextrans-2013-051014>.
- Rudd P, Frei T. How personal is the personal health record?: comment on “the digital divide in adoption and use of a personal health record”. *Arch Intern Med*. 2011;171(6):575–6. <https://doi.org/10.1001/archinternmed.2011.35>. No abstract available.
- Saparova D. Motivating, influencing, and persuading patients through personal health records: a scoping

- review. *Perspect Health Inf Manag*. 2012;9:1f. Epub 2012 Apr 1.
- Sittig DF, Singh H. Rights and responsibilities of users of electronic health records. *CMAJ*. 2012;184(13):1479–83.
- Sittig DF, Singh H, Longhurst CA. Rights and responsibilities of electronic health records (EHR) users caring for children. *Arch Argent Pediatr*. 2013;111(6):468–71.
- Wynia M, Dunn K. Dreams and nightmares: practical and ethical issues for patients and physicians using personal health records. *J Law Med Ethics*. 2010;38(1):64–73. <https://doi.org/10.1111/j.1748-720X.2010.00467.x>.
- Yamin CK, Emani S, Williams DH, Lipsitz SR, Karson AS, Wald JS, Bates DW. The digital divide in adoption and use of a personal health record. *Arch Intern Med*. 2011;171(6):568–74. <https://doi.org/10.1001/archinternmed.2011.34>.



# A Framework for Health System Comparisons: The Health Systems in Transition (HiT) Series of the European Observatory on Health Systems and Policies

# 13

Bernd Rechel, Suszy Lessof, Reinhard Busse, Martin McKee, Josep Figueras, Elias Mossialos, and Ewout van Ginneken

## Contents

<b>Introduction</b> .....	280
The Ljubljana Charter: HiTs and Health Systems in Transition .....	281
The Observatory Partnership: HiTs and Policy Relevance .....	281
The Observatory Functions: HiTs in a Wider Work Plan .....	282

---

B. Rechel (✉)  
European Observatory on Health Systems and Policies,  
London School of Hygiene and Tropical Medicine,  
London, UK  
e-mail: [Bernd.Rechel@lshtm.ac.uk](mailto:Bernd.Rechel@lshtm.ac.uk)

S. Lessof · J. Figueras  
European Observatory on Health Systems and Policies,  
Brussels, Belgium  
e-mail: [szy@obs.euro.who.int](mailto:szy@obs.euro.who.int); [jfi@obs.euro.who.int](mailto:jfi@obs.euro.who.int)

R. Busse  
Technische Universität Berlin, Berlin, Germany  
Department Health Care Management, Faculty of  
Economics and Management, Technische Universität,  
Berlin, Germany  
e-mail: [rbusse@tu-berlin.de](mailto:rbusse@tu-berlin.de)

M. McKee  
London School of Hygiene and Tropical Medicine,  
London, UK  
e-mail: [Martin.McKee@lshtm.ac.uk](mailto:Martin.McKee@lshtm.ac.uk)

E. Mossialos  
London School of Economics and Political Science,  
London, UK  
e-mail: [e.a.mossialos@lse.ac.uk](mailto:e.a.mossialos@lse.ac.uk)

E. van Ginneken  
Berlin University of Technology, Berlin, Germany  
European Observatory on Health Systems and Policies,  
Department of Health Care Management, Berlin  
University of Technology, Berlin, Germany  
e-mail: [ewout.vanginneken@tu-berlin.de](mailto:ewout.vanginneken@tu-berlin.de)



<b>The HiT Template: Structuring, Populating, and Signposting a Comparative Framework</b> .....	283
Structure .....	283
Scope and Content .....	284
Signposting .....	285
<b>HiT Processes: Making Sure Frameworks Are Used Consistently and Comparably</b> .....	285
Data Sources .....	286
Authors, Author Teams, and the Role of (Contributing) Editors .....	286
Long-Term Relationships .....	286
Flexibility, Consistency, and Signaling Gaps .....	286
Review .....	287
<b>Dissemination and Policy Relevance: Helping Frameworks Achieve Their Objectives</b> .....	287
Timeliness .....	287
Visibility .....	287
Signaling Credibility .....	289
<b>Lessons Learned</b> .....	289
The Value of a Template .....	289
The Importance of Author and Editor Roles .....	289
The Need to Build In “Accessibility” and Relevance .....	289
The Need to Signal Credibility .....	294
The Need to Build in a Review Process .....	294
<b>Conclusions</b> .....	294
<b>References</b> .....	296

### Abstract

Comparing health systems across countries allows policy-makers to make informed decisions on how to strengthen their systems. The European Observatory on Health Systems and Policies produces a series of profiles that systematically describe health systems – the HiTs. These capture how a health system is organized, how funds flow through the system, and what care it provides. They follow a common template and are updated periodically. They allow policy-makers and academics to understand each system individually in light of its previous development and in the context of other European health systems. In effect, the HiTs provide a framework for comparison across countries. This chapter describes the Observatory’s experience in developing the framework. It explores the role of the HiT template, the processes put in place to support consistency and comparability, and the efforts to build in policy relevance. It highlights the

lessons learned so far and what they might contribute to the development of other comparative frameworks.

### Introduction

The European Observatory on Health Systems and Policies (Observatory) is a partnership of countries, international organizations, and academic institutions that was set up to provide evidence for the policy-makers shaping Europe’s health systems. A central pillar of its work is the Health Systems in Transition (HiT) series – a set of highly structured and analytic descriptions of country health systems that are updated periodically. This experience of monitoring and comparing country health systems and policies, which stretches back over 20 years, provides insights into the challenges researchers face in developing and applying any framework for health system comparisons. Understanding the background to the HiT series and the Observatory

helps explain the specific approach taken to HiTs, but also speaks of the significance of context in developing comparative frameworks.

### **The Ljubljana Charter: HiTs and Health Systems in Transition**

The Observatory can trace its origins to the early 1990s and the challenges Europe faced as western European expectations (and health-care costs) rose and as the countries emerging in the wake of the Soviet Union looked to overhaul their own health systems. The World Health Organization (WHO) Regional Office for Europe facilitated a process that culminated in the 1996 Ljubljana conference on European Health Care Reforms and the Ljubljana Charter, in which health ministers from across the European region committed themselves to a set of principles for health system reform. These reflected a growing understanding of health's part in the wider society and economy, the importance of people and patients, the need for policy to be "based on evidence where available," and the role of monitoring and learning from experience (Richards 2009).

The original HiTs were developed as part of the preparations for the Ministerial Conference. They were addressing a postcommunist Europe in which more than 15 new countries had emerged and many more were making a transition from state-managed to market economies with all the accompanying economic upheaval. There were also growing challenges to the sustainability of established and wealthy health systems and to notions of solidarity. The HiTs had therefore to establish a common vocabulary for describing health systems and to make sure that the terms used could be explained and understood in countries with very different traditions. They had also to provide for the fact that the systems to be compared were contending with significant discontinuities and ongoing change. This prompted the development of a template to describe health systems that would set down the bases on which to make comparisons across countries. It was comprehensive, allowed for very different path developments, and offered detailed explanations to guide authors.

### **The Observatory Partnership: HiTs and Policy Relevance**

Many of (what came to be) the Observatory team were involved in developing evidence for Ljubljana. The Observatory, which took formal shape in May 1998, was designed to take forward the approach to evidence for policy, after the Charter was agreed (Box 1). The original Partners were WHO Europe, the government of Norway, the European Investment Bank, the World Bank, the London School of Economics and Political Science (LSE), and the London School of Hygiene & Tropical Medicine (LSHTM). The exact composition of the partnership has changed over the years, so that the Observatory today also includes the European Commission, more national governments (Austria, Belgium, Finland, Ireland, Slovenia, Sweden, and the United Kingdom), a regional government (Veneto), and the French National Union of Health Insurance Funds (UNCAM); but the concept of a partnership that brings different stakeholders together remains the same. The idea is that the Observatory, like a good health system, is informed by the people who use its services as well as those providing them. The Partners have genuine experience of shaping health systems, and this has prompted a focus on policy relevance and how decision-makers can access and use the evidence generated. They have insisted that the HiT series should be "accessible" to a nonspecialist, non-academic audience and, more specifically, be readable, clearly structured, consistent (so that readers can move from one HiT to another and find comparable information), and timely, that is, available while the data and analysis are still current.

#### **Box 1: The European Observatory on Health Systems and Policies**

The core mission of the Observatory is to support and promote evidence-based health policy-making through the comprehensive and rigorous analysis of the dynamics of

*(continued)*

**Box 1:** (continued)

health systems in Europe and through brokering the knowledge generated.

The Observatory is overseen by a steering committee, made up of representatives of all its Partners, which sets priorities and emphasizes policy relevance. Work is shared across four teams with a “secretariat” in Brussels that coordinates and champions knowledge brokering and analytic teams in Brussels, London (at LSE and LSHTM), and in Berlin (at the University of Technology).

The core staff team carries out research and analysis but depends (often) on secondary research and (almost always) on the Observatory’s extensive academic and policy networks. Over 600 researchers and practitioners provide country- and topic-specific knowledge, insights, and understanding. Collectively the Observatory and those who contribute to it equip Europe’s policy-makers and their advisors with evaluative and comparative information that can help them make informed policy choices.

### The Observatory Functions: HiTs in a Wider Work Plan

The Observatory has four core functions: country monitoring, analysis, comparative health system performance assessment, and knowledge brokering (Box 2). The HiTs are a fundamental part of country monitoring, supported by a (relatively) new initiative to provide online updates – the Health Systems and Policy Monitor (HSPM). They are, to some extent, a stand-alone exercise. However, the fact that the Observatory’s portfolio of work is broader than country monitoring has done much to strengthen the comparative framework. The analysis program runs in-depth studies of issues like governance, insurance mechanisms, staff mobility, hospitals, primary care, care for chronic conditions, and the economics of prevention, using HiTs, but also reviews of the academic literature

and secondary data collection. These have provided insights into important health system dimensions and how they impact on each other. At the same time, they create (a positive) pressure on the HiT series to deliver consistent and comparable information that can feed into more in-depth analysis. The performance assessment work has given the Observatory the tools to understand the use (and misuse) of performance measures and address how far systems achieve their goals. The contribution of these “other” functions to the HiT makes clear the value of wide-ranging inputs from different specialist and thematic perspectives in developing a comparative framework.

#### Box 2: The Observatory’s Core Functions

- Country monitoring generates systematic overviews of health systems in Europe (and in key OECD countries beyond) in the form of Health Systems in Transition (HiT) reviews. All HiTs are available on the web, listed in PubMed, and disseminated at launches and conferences.

The Health Systems and Policy Monitor (HSPM) is a new initiative to update HiTs online. It is a web platform that hosts 27 “living HiTs.” These are regularly updated by the expert members of the HSPM network with short “reform logs” and longer “health policy updates.” These give users news and insights into policy processes and developments <http://www.hspm.org/mainpage.aspx>.

The HSPM also allows users to extract and merge specific sections from the HiTs for several countries at the same time as a single file, facilitating comparisons <http://www.hspm.org/searchandcompare.aspx>.

- Analysis provides for in-depth work on core health system and policy issues. The Observatory brings together teams of academics, policy analysts, and practitioners from different institutions, countries, and disciplines to ensure rigorous meta-analysis and secondary research on

(continued)

**Box 2:** (continued)

the issues that matter most to decision-makers. All evidence is available “open access” to facilitate its use in practice.

- Performance assessment includes a package of methodological and empirical work designed to respond to country needs. There have been two key studies looking at the policy agenda for performance comparison to improve health services and separate work on the domains that comprise performance (efficiency, population health, responsiveness).
- Knowledge brokering involves engaging with policy-makers to understand what evidence they need and then assembling and communicating the relevant information at the right time. The Observatory combines an extensive publication program with face-to-face and electronic dissemination to convey evidence on what might work better or worse in different country and policy contexts.

### **The HiT Template: Structuring, Populating, and Signposting a Comparative Framework**

HiTs use a standard questionnaire and format to guide authors – referred to as the HiT template. It guides the production of detailed descriptions of health system and policy initiatives so that every HiT examines the organization, financing, and delivery of health services, the role of key actors, and the challenges faced in the same way, establishes a comparable baseline for reviewing the impact of reforms, and takes a standardized approach to health system assessment. This structure is central to the ability of HiTs to inform comparative analysis and facilitates the exchange of reform experiences across countries. Arriving at a robust template is not straightforward, but the Observatory’s experience suggests some elements that can help.

### **Structure**

The HiT template benefits from a clear structure, based on a functional perspective of health systems. It works from the premise that all health systems perform a number of nonnormative core functions (Duran et al. 2012), including the organization, the governance, the financing, the generation of physical and human resources, and the provision of health services. The first HiT template was developed in 1996. It was revised in 2007 and again in 2010, but all iterations have used the notion of core functions and have drawn on the literature and prevailing debate to interpret what those functions are.

All revisions have involved input from staff (editors) and national authors, based on their work on the country profiles, but they have also included consultation with a wider group of users and stakeholders (Observatory Partners, various units of WHO and of the European Commission’s health directorate, and, more recently, members of the HSPM network). These review stages have helped strengthen the template and build some consensus around its structure and approach.

Table 1 shows the changes over time and the very marked structural consistency between versions. This is in part because of a conscious decision to adapt rather than rethink the structure completely so that HiT users can read backwards in time as well as across countries. It is also a testament to the robustness of the first iteration. The adjustments reflect on a wider rethinking on how different elements fit into the whole and on what seemed more or less important at particular times.

The initial template placed more emphasis on the political, economic, and sociodemographic context and on a country’s historical background, because of the proximity to transition for so many eastern European countries. The 2004–2007 revision consolidated financing in one chapter, bringing together the collection and allocation of funds, and split the chapter on organization and management to address planning and regulation separately, reflecting shifts in emphasis at the time in wider academic and policy thinking. In addition, a

**Table 1** The evolution of the HiT template structure

Version 1: developed 1995–1996 <sup>a</sup>	Version 2: developed 2004–2007 <sup>b</sup>	Version 3: developed 2009–2010 <sup>c</sup>
Introduction and historical background	Introduction	Introduction
Organizational structure and management	Organizational structure	Organization and governance
Health-care finance and expenditure	Financing	Financing
	Planning and regulation	
	Physical and human resources	Physical and human resources
Health-care delivery system	Provision of services	Provision of services
Financial resource allocation		
Health-care reforms	Principal health-care reforms	Principal health-care reforms
	Assessment of the health system	Assessment of the health system
Conclusions	Conclusions	Conclusions
References	Appendices	Appendices

<sup>a</sup>Figueras and Tragakes (1996)

<sup>b</sup>Mossialos et al. (2007)

<sup>c</sup>Rechel et al. (2010)

new chapter was added, on the assessment of the health system, again a response to the more explicit way this issue was being addressed at the time. The 2010 template condensed organization, governance, planning, and regulation into a single chapter again and revised and extended the section on performance assessment as policy-makers became increasingly interested in understanding and contextualizing the evaluations of their health systems that they were being confronted with.

## Scope and Content

There were of course other changes to the template between iterations in terms of the detail addressed within the relatively stable overall structure. New questions and issues were added because areas like mental health, child health services, and palliative care (2007) or public health and intersectorality (2010) came to the policy fore and as a wide group of experts and users were consulted. The 2007 template was particularly heavily laden with new additions and contributed to longer and more time-consuming HiTs. Certainly there was a marked growth in the length of HiTs in successive iterations with Estonia, for example, growing from 67 pages in 2000, to 137 pages in 2004, and 227 pages in 2008. This was addressed to some extent in 2010 with a

tightening of the template (see Box 3) after which the 2013 Estonia HiT dropped to 195 pages, and it is being revisited again in the 2015–2016 update.

### Box 3: The 2010 Template, Structure and Contents

1. **Introduction:** the broader context of the health system, including economic and political context, and population health
2. **Organization and governance:** an overview of how the health system in the country is organized, the main actors and their decision-making powers, the historical background, regulation, and levels of patient empowerment
3. **Financing:** information on the level of expenditure, who is covered, what benefits are covered, the sources of health-care finance, how resources are pooled and allocated, the main areas of expenditure, and how providers are paid
4. **Physical and human resources:** the planning and distribution of infrastructure and capital stock, IT systems, and human resources, including registration, training, trends, and career paths
5. **Provision of services:** concentrates on patient flows, organization and delivery

(continued)

**Box 3:** (continued)

of services, addressing public health, primary and secondary health care, emergency and day care, rehabilitation, pharmaceutical care, long-term care, services for informal carers, palliative care, mental health care, dental care, complementary and alternative medicine, and health care for specific populations

6. **Principal health reforms:** reviews reforms, policies, and organizational changes that have had a substantial impact on health care, as well as future developments
7. **Assessment of the health system:** provides an assessment based on the stated objectives of the health system, financial protection, and equity in financing; user experience and equity of access to health care; health outcomes, health service outcomes, and quality of care; health system efficiency; and transparency and accountability
8. **Conclusions:** highlights the lessons learned from health system changes and summarizes remaining challenges and future prospects
9. **Appendices:** includes references, further reading, and useful web sites

and easier to read and update. The editorial team also drew up word limits for chapters, although these have not been included in the published template yet; they are used with authors to agree the length of HiTs. The changes in the way terms are explained reflect the fact that they are now familiar to authors and readers alike.

Key changes that have been aimed at readers include the reorganization of several subsections to increase accessibility and clarity and the introduction of summary paragraphs with key messages at the start of chapters, an abstract (of less than one page), and an executive summary (of three to five pages). These pull out (or signpost) findings in a way that allows policy-makers and their advisers quick access and is in line with the Observatory's growing understanding of knowledge brokering (Catallo et al. 2014) and the testing of "HiT Summaries" between 2002 and 2008.

There is a further round of revision which started in 2015 and is now being piloted, which will fine-tune the HiT template. It will signpost still more explicitly how health systems are doing by integrating more evaluative elements in the broadly "descriptive" sections rather than keeping them all for a single, policy-focused, assessment section.

---

### HiT Processes: Making Sure Frameworks Are Used Consistently and Comparably

#### Signposting

The HiT template has also seen a number of significant changes to layout and design. These have aimed firstly to make the template itself more user-friendly for authors and editors and secondly to create easier to read HiTs.

Key changes from the perspective of authors have been clear signposting of sections or sets of questions that are "essential" and of those which are only "discretionary" and some reworking of the glossary elements and examples that characterized the 1996 template. The intention in flagging what is and what is not essential is to help authors and editors to focus and keep HiTs short

The HiT template in its various iterations has guided the writing of country profiles, providing a clear overall structure, as well as detailed notes on what belongs in the various subsections. However, despite its definitions and advice on how to produce a HiT, it is not a tool that can ensure consistency and comparability on its own. This is because health systems are so complex and there are so many layers of information that could be deemed relevant. The Observatory has therefore developed a range of practice over the last 20 years that helps make the template into a framework that supports health system comparisons.

## Data Sources

Data is of course a constant issue in seeking to make comparisons, particularly across countries. The Observatory has chosen to supply quantitative data in the form of a set of standard comparative tables and figures for each country, drawing on the European Health for All Database (HFA-DB) of the WHO Regional Office for Europe, as well as the databases from the Organization for Economic Co-operation and Development (OECD), the Eurostat, and the World Bank. All of these international databases rely largely on official figures provided and approved by national governments. These are not unproblematic. The WHO Europe HFA database covers the 53 countries of its European region and Eurostat the 28 EU member states and the 4 members of the European Free Trade Association, while OECD Health Statistics covers the 34 OECD countries (of which only 26 are in WHO's European region and 22 in the EU). There are also differences in definitions and data collection methods. However, they have the merit of being consistently compiled and rigorously checked. National statistics are also used in the HiTs, although they may raise methodological issues, as are national and regional policy documents, and academic and gray literature, although these do not of course have comparability built in. Data in HiTs is discussed and assessed, and there is explicit attention given to discrepancies between national and international sources.

## Authors, Author Teams, and the Role of (Contributing) Editors

HiTs are produced by country experts in close collaboration with Observatory (analytic) staff. Having a national author is important because the framework covers so much ground it is extremely difficult to marshal the range of information needed to complete it from "outside." It also creates ownership within the country and the national academic community which encourages subsequent use of the profile. The choice of national experts is important and needs to reflect

research expertise and signal credibility. Appointing small teams of national authors can also be a helpful way of bringing different skills and knowledge into the process. However experienced the author team, writing a HiT is a complex process. The role of the editor is extremely important and a crucial factor in applying the HiT framework so that it can support comparisons. Observatory editors play a proactive role and are expected to address not just the quality of the individual profile they are working on but its fit with the rest of the series. They are often credited as authors because of the contribution they make.

## Long-Term Relationships

The HiTs are updated periodically, and the Observatory has found that building long-term relationships with its author teams is efficient in terms of minimizing the learning curve (and costs) of new iterations and, as importantly, is effective in promoting focus and consistency. The template is a complicated instrument and familiarity with it (and a role in shaping it) makes a difference in authors' ability to use it. It also fosters a sense of co-ownership of and commitment to the outputs.

The HSPM initiative (Box 3) has strengthened these links, engaging authors and contributors by sharing ownership, creating publishing opportunities (with its dedicated series in Health Policy <http://www.hspm.org/hpj.aspx>), and holding annual meetings which let authors and editors meet and exchange ideas. Efforts to properly integrate national experts into thinking on a comparative framework and to acknowledge their contribution are demonstrably worthwhile.

## Flexibility, Consistency, and Signaling Gaps

The experience of writing HiTs makes clear that no two health systems are identical. There needs to be an ability, therefore, to apply the template thoughtfully. Each profile should bring out what is important in a country without slavishly rehearsing details that are not pertinent while

simultaneously maintaining comparability with other countries. It has proved to be helpful to flag up where data is missing or an element of a system is not yet developed rather than simply avoiding mention of it, as it helps readers understand gaps. Editors have an important role in steering HiTs between flexibility and consistency and deciding what should be included or omitted. They meet regularly to exchange experience and discuss practice.

## Review

Review is an essential element of the HiT process. Each HiT editor works with their supervisor and the Brussels secretariat as needed to resolve issues. When the draft HiT is complete to their satisfaction, the Observatory combines external review by academics (typically one national and one international) with that of policy-makers. This means quality is addressed not only through academic criteria but also in terms of readability, credibility, and policy relevance. The draft is also sent to the Ministry of Health and the Observatory Partners for comment. Ministries of Health are allowed 2 weeks to flag any factual concerns, but they do not approve HiTs. In the same way, Partners can comment but do not have a clearance function. Any feedback provided is handled by the editor and introduced only where it is consistent with the evidence. Completed HiTs are given a final check by one of the Observatory's codirectors or hub coordinators to ensure that they achieve expectations on quality and objectivity and fulfill the aims of the series.

---

## Dissemination and Policy Relevance: Helping Frameworks Achieve Their Objectives

The HiTs are designed to allow comparisons across countries, but they are not intended purely to feed into (academic) research and analysis. HiT audiences are often national policy-makers who use the HiT to take stock of their own health system and to reach a shared understanding of

what is happening which different sectors, ministries, and levels of the health service (primary, secondary, regional, local) can all subscribe to. They use HiTs in considering reforms, as the basis for policy dialogue and to explore policy options, and to set their own health system's performance in a European context. Other users are foreign analysts or consultants trying to get a comprehensive understanding of a health system, and researchers and students. HiTs are a single source of information and pull together different strands of analysis which otherwise can be surprisingly hard to find in "one place."

## Timeliness

Any comparative evidence will have more impact if it is delivered when it is still "current" and if it can coincide with a window of opportunity for reform. The Observatory tries to turn HiTs around in the shortest possible time, although this is not always easy. The Health Systems and Policy Monitor is, in part, a response to this and provides a log of policy developments and reforms online in between formal HiT updates. Other steps to ensure that material is not superseded by developments before it is published include agreeing a schedule with authors in advance, efforts to keep HiTs short and focused, and quick turnaround on review stages, all of which must be underpinned by strong project management on the part of the HiT editor. Linking HiTs to an entry point where they are likely to be considered by policy-makers is both a way of motivating authors to deliver on time and a way of securing impact when they do. The Observatory has successfully tied HiTs and HiT launches to EU Presidencies (Denmark 2012, Lithuania 2013, Italy 2014, Luxembourg 2015), to moments of political change (Ukraine 2015), and to major reform programs in countries (Slovenia 2016).

## Visibility

HiTs can only be used when potential users are aware of their existence. The Observatory has developed a mix of dissemination approaches to encourage uptake. There are launch events,



typically in the country and in collaboration with national authors, partner institutions, and Ministries of Health. These work particularly well if linked to a policy dialogue (a facilitated debate about policy options for decision-makers) or a major national or international conference (like the Polish annual National Health Fund meeting or the Czech Presidency of the Visegrad Group) or a workshop or meeting held by other agencies (European Commission meeting on health reform in Ukraine).

All HiTs are available as open access online on the Observatory's web site and there are e-bulletins and tweets to draw attention to new publications <http://www.euro.who.int/en/about-us/partners/observatory>. A list of the latest available HiTs for the various countries is shown in Box 4.

**Box 4: Latest Available HiTs, September 2016**

Albania HiT (2002)  
 Andorra HiT (2004)  
 Armenia HiT (2013)  
 Australia HiT (2006)  
 Austria HiT (2013)  
 Azerbaijan HiT (2010)  
 Belarus HiT (2013)  
 Belgium HiT (2010)  
 Bosnia and Herzegovina HiT (2002)  
 Bulgaria HiT (2012)  
 Canada HiT (2013)  
 Croatia HiT (2014)  
 Cyprus HiT (2012)  
 Czech Republic HiT (2015)  
 Denmark HiT (2012)  
 Estonia HiT (2013)  
 Finland HiT (2008)  
 France HiT (2015)  
 Georgia HiT (2009)  
 Germany HiT (2014)  
 Greece HiT (2010)  
 Hungary HiT (2011)  
 Iceland HiT (2014)  
 Ireland HiT (2009)  
 Israel HiT (2015)  
 Italy HiT (2014)  
 Italy, Veneto Region HiT (2012)  
 Japan HiT (2009)

**Box 4: (continued)**

Kazakhstan HiT (2012)  
 Kyrgyzstan HiT (2011)  
 Latvia HiT (2012)  
 Lithuania HiT (2013)  
 Luxembourg HiT (2015)  
 Malta HiT (2014)  
 Mongolia HiT (2007)  
 Netherlands HiT (2016)  
 New Zealand HiT (2001)  
 Norway HiT (2013)  
 Poland HiT (2011)  
 Portugal HiT (2011)  
 Republic of Korea HiT (2009)  
 Republic of Moldova HiT (2012)  
 Romania HiT (2016)  
 Russian Federation HiT (2011)  
 Slovakia HiT (2011)  
 Slovenia HiT (2016)  
 Spain HiT (2010)  
 Sweden HiT (2012)  
 Switzerland HiT (2015)  
 Tajikistan HiT (2016)  
 The former Yugoslav Republic of Macedonia HiT (2006)  
 Turkey HiT (2011)  
 Turkmenistan HiT (2000)  
 Ukraine HiT (2015)  
 United Kingdom HiT (2015)  
 United Kingdom, England HiT (2011)  
 United Kingdom, Northern Ireland HiT (2012)  
 United Kingdom, Scotland HiT (2012)  
 United Kingdom, Wales HiT (2012)  
 United States of America HiT (2013)  
 Uzbekistan HiT (2014)

Translations can also be extremely helpful in facilitating national access, and HiTs have been translated from English into 11 other languages, including Albanian, Bulgarian, Estonian, French, Georgian, Polish, Romanian, Russian, Spanish, and Turkish. However, translation is expensive and requires careful review by national authors as concepts and policy terms often pose problems.

## Signaling Credibility

Securing visibility alone cannot ensure uptake. It is helpful also to demonstrate credibility. The Observatory has gone about this in a number of ways. It invests considerable resources in “presentation,” i.e., copy-editing and typesetting, so that the HiTs signal professionalism. It also endorses all aspects of the International Committee of Medical Journal Editors’ Uniform Requirements for Manuscripts Submitted to Biomedical Journals ([www.ICMJE.org](http://www.ICMJE.org)) that are relevant to HiTs. It has also taken time to make the HiTs compatible with PubMed/Medline requirements, and the Health Systems in Transition series has been recognized as an international peer-reviewed journal and indexed on PubMed/Medline since 2010.

---

## Lessons Learned

The experience of the HiT series suggests a number of lessons for frameworks for health system comparisons. These include:

### The Value of a Template

A template that follows a rational and defensible structure, establishes a common vocabulary with clearly defined terms (supported by examples when appropriate), and is mindful of the way researchers from different disciplines and national traditions may understand it is an invaluable tool. It needs to include clear and sensible explanations on how to use it, be sufficiently robust to accommodate change over time, and allow a certain degree of flexibility. It should also reflect on what the final output is expected to be and who will use it.

### The Importance of Author and Editor Roles

Comparative work demands data collection and analysis in different settings and national expertise is key to this. Selecting authors with appropriate skills and credibility is therefore essential and is boosted by clear criteria, by using teams

rather than single authors, and by building long-term relationships, which is possible through a network like the HSPM. Good authors must be complemented by equally skilful editors who can support the authors and ensure consistency. Bringing editors and authors together to agree expectations around timing and quality can be extremely effective, as is keeping editors in touch with each other.

The experience of the Observatory suggests that it is useful to provide for two roles analogous to national author and HiT editor, to have clear (academic) criteria for guiding the choice of author, to schedule an initial meeting between the editor and author(s) to go through the template and clarify expectations, and to agree a clear timetable. In the case of the HiT template, there is often discussion of how to tailor the HiT to national circumstances (and specifically of which areas will be addressed in more detail and which in less), but this may not apply to other comparative frameworks. The experience with HiTs also suggests there needs to be allowance for numerous drafts and iterations before the overall manuscript is ready for review. While this may be less of an issue in frameworks with a narrower coverage, plans should include sufficient opportunities for authors and editors to exchange views.

### The Need to Build In “Accessibility” and Relevance

Users need to be considered in designing the template, the processes to deliver the comparisons, and the way findings are disseminated. Readable, well-structured, well-presented reports that allow users to move from one report to another and find comparable information easily will increase uptake and impact. Abstracts, summaries, and key messages will all help different users access the things they need. An example of a cover and an executive summary of a HiT are shown in Fig. 1 and Box 5. Delivering timely (current) data and analysis is also important if the evidence generated is to have an impact. Reports that are overly long and

**Fig. 1** Cover of the 2014 German HiT (Source: Busse and Blümel 2014)



detailed can still be useful, but they may tend to be used by academics rather than policy-makers. Furthermore, those developing comparative frameworks need to have an explicit debate as to how best to balance the comprehensive against the manageable and the timely. A mix of approaches to dissemination should be considered, paying attention to ease of access, free download from the Internet, and translation into other languages.

**Box 5: Executive Summary from Germany, Health System Review, 2014**

The Federal Republic of Germany is in central Europe, with 81.8 million inhabitants (December 2011), making it by some distance the most populated country in the

**Box 5:** (continued)

European Union (EU). Berlin is the country's capital and, with 3.5 million residents, Germany's largest city.

In 2012 Germany's gross domestic product (GDP) amounted to approximately €32 554 per capita (one of the highest in Europe). Germany is a federal parliamentary republic consisting of 16 states (Länder), each of which has a constitution reflecting the federal, democratic, and social principles embodied in the national constitution known as the Basic Law (Grundgesetz).

By 2010, life expectancy at birth in Germany had reached 78.1 years for men and

(continued)

**Box 5:** (continued)

83.1 years for women (slightly below the Eurozone average of 78.3 years for men and 84.0 years for women, although the gap with other similar European countries has been narrowing). Within Germany, the gap in life expectancy at birth between East and West Germany peaked in 1990 at 3.5 years for men and 2.8 years for women, but narrowed following reunification to 1.3 years for men and 0.3 years for women. Moreover, differences in life expectancy in Germany no longer follow a strict east–west divide. The lowest life expectancy for women in 2004, for example, was observed in Saarland, a land in the western part of the country.

A fundamental facet of the German political system – and the health-care system in particular – is the sharing of decision-making powers between the Länder, the federal government, and civil society organizations. In health care, the federal and Länder governments traditionally delegate powers to membership-based (with mandatory participation), self-regulated organizations of payers and providers, known as “corporatist bodies.” In the statutory health insurance (Gesetzliche Krankenversicherung (SHI)) system, these are, in particular, sickness funds and their associations together with associations of physicians accredited to treat patients covered by SHI. These corporatist bodies constitute the self-regulated structures that operate the financing and delivery of benefits covered by SHI, with the Federal Joint Committee (Gemeinsamer Bundesausschuss) being the most important decision-making body. The Social Code Book (Sozialgesetzbuch (SGB)) provides regulatory frameworks; SGB V has details decided for SHI.

Since 2009, health insurance has been mandatory for all citizens and permanent residents, either through SHI or private health insurance (PHI). SHI covers 85% of

**Box 5:** (continued)

the population – either mandatorily or voluntarily. Cover through PHI is mandatory for certain professional groups (e.g., civil servants), while for others it can be an alternative to SHI under certain conditions (e.g., the self-employed and employees above a certain income threshold). In 2012, the percentage of the population having cover through such PHI was 11%. PHI can also provide complementary cover for people with SHI, such as for dental care. Additionally, 4% of the population is covered by sector-specific governmental schemes (e.g., for the military). People covered by SHI have free choice of sickness funds and are all entitled to a comprehensive range of benefits.

Germany invests a substantial amount of its resources in health care. According to the Federal Statistical Office (Statistisches Bundesamt), which provides the latest available data on health expenditure, total health expenditure was €300.437 billion in 2012, or 11.4% of GDP (one of the highest in the EU). This reflects a sustained increase in health-care expenditure even following the economic crisis in 2009 (with total health expenditure rising from 10.5% of GDP in 2008).

Although SHI dominates the German discussion on health-care expenditure and reform(s), its actual contribution to overall health expenditure was only 57.4% in 2012. Altogether, public sources accounted for 72.9% of total expenditure on health, with the rest of public funding coming principally from statutory long-term care insurance (Soziale Pflegeversicherung). Private sources accounted for 27.1% of total expenditure. The proportion of health care financed from taxes has decreased throughout the last decades, falling from 10.8% in 1996 to 4.8% in 2012. The most significant decrease of public expenditure was recorded

*(continued)*

**Box 5:** (continued)

for long-term care (over 50%) with the introduction of mandatory long-term care insurance in 1993 shifting financing away from means-tested social assistance.

The 132 sickness funds collect contributions and transfer these to the Central Reallocation Pool (Gesundheitsfonds; literally, “Health Fund”). Contributions increase proportionally with income to an upper threshold (a monthly income of €4050 in 2014). Since 2009 there has been a uniform contribution rate (15.5% of income). Resources are then redistributed to the sickness funds according to a morbidity-based risk-adjustment scheme (morbidity-oriented *Risikostrukturausgleich*; often abbreviated to *Morbi-RSA*), and funds have to make up any shortfall by charging a supplementary premium.

Sickness funds pay for health-care providers, with hospitals and physicians in ambulatory care (just ahead of pharmaceuticals) being the main expenditure blocks. Hospitals are financed through “dual financing,” with financing of capital investments through the *Länder* and running costs through the sickness funds, private health insurers, and self-pay patients – although the sickness funds finance the majority of operating costs (including all costs for medical goods and personnel). Financing of running costs is negotiated between individual hospitals and *Länder* associations of sickness funds and primarily takes place through diagnosis-related groups (*Diagnose-bezogene Fallpauschale*; DRGs). Public investment in hospital infrastructure has declined by 22% over the last decade and is not evenly distributed; in 2012, hospitals in the western part of Germany received 83% of such public investment.

Payment for ambulatory care is subject to predetermined price schemes for each profession (one for SHI services and one

**Box 5:** (continued)

for private services). Payment of physicians by the SHI is made from an overall morbidity-adjusted capitation budget paid by the sickness funds to the regional associations of SHI physicians (*Kassenärztliche Vereinigungen*), which they then distribute to their members according to the volume of services provided (with various adjustments). Payment for private services is on a fee-for-service basis using the private fee scale, although individual practitioners typically charge multiples of the fees indicated.

In 2012, there were 2017 hospitals with a total of 501 475 beds (6.2 beds per 1000; higher than any other EU country). Of these, 48% of beds were in publicly owned hospitals, 34% in private non-profit, and 18% in private for-profit hospitals. Both SHI and PHI (as well as the two long-term care insurance schemes) use the same providers. Although acute hospital beds have been reduced substantially since 1991, the number of acute hospital beds is still almost 60% higher than the EU15 (15 EU Member States before May 2004) average. The average length of stay decreased steadily between 1991 and 2011, falling from 12.8 to 7.7 days.

Health care is an important employment sector in Germany, with 4.9 million people working in the health sector, accounting for 11.2% of total employment at the end of 2011. According to the WHO Regional Office for Europe’s Health for All Database, 382 physicians per 100 000 were practicing in primary and secondary care. Thus, the density of physicians in Germany was slightly above the EU15 average and substantially higher than the EU28 (Member States at 1 July 2013) average; the relative numbers of nurses and dentists are also higher than the EU average. With the EU enlargements of 2004 and 2007, a growing migration of health professionals to

(continued)

**Box 5:** (continued)

Germany had been expected. In fact, the number of foreign health workers grew from 2000 and reached its peak in 2003, thus before the enlargements. The extent of migration to Germany is relatively small compared with that to other destination countries in the EU.

Ambulatory health care is mainly provided by private for-profit providers. Patients have free choice of physicians, psychotherapists (including psychologists providing psychotherapy, since 1999), dentists, pharmacists, and emergency room services. Although patients covered by SHI may also go to other health professionals, access to reimbursed care is available only upon referral by a physician. In 2012, of the 121 198 practicing SHI-accredited physicians in Germany (psychotherapists not included), 46% were practicing as family physicians and 54% as specialists. German hospitals have traditionally concentrated on inpatient care, with strict separation from ambulatory care. This rigid separation has been made more permeable in recent years and now hospitals are partially authorized to provide outpatient services and to participate in integrated care models and disease management programs (DMPs).

For pharmaceuticals, while hospitals may negotiate prices with wholesalers or manufacturers, the distribution chain and prices are much more regulated in the pharmacy market. In both sectors, manufacturers are free in theory to set prices without direct price controls or profit controls. However, there is a reference pricing system for SHI reimbursement, which has been steadily strengthened over recent years, whereby “reference” prices are defined nationally for groups of similar pharmaceuticals with reimbursement capped at that level. Although prices can be set higher (with the patient paying the difference), in practice very few

**Box 5:** (continued)

drugs exceed the reference price. For pharmaceuticals with an additional benefit beyond existing reference price groups, reimbursement amounts are negotiated between the manufacturer and the Federal Association of Sickness Funds (GKV-Spitzenverband). Patients generally pay co-payments for pharmaceuticals of €5–10; there are also other cost-saving measures, such as provisions for generic substitution. Of the pharmaceutical industry’s total turnover in 2011 of €38.1 billion, €14.3 billion was gained in the domestic market and €23.8 billion from exports (62.5%); Germany is the third largest producer of pharmaceuticals in the world after the United States and Japan.

Public health is principally the responsibility of the Länder, covering issues such as surveillance of communicable disease and health promotion and education. Historically, the Länder have resisted the influence of the federal government on public health, and although some elements of public health have been included in SHI in recent decades (such as cancer screening), and other interventions have separate agreements (e.g., immunizations), a “prevention act” at federal level intended to consolidate and clarify responsibilities in this area in 2005 was ultimately rejected by the Federal Assembly (Bundesrat).

Governmental policy since the early 2000s has principally focused on cost containment and the concept of a sustainable financing system. The government in office at the time of writing, again a grand coalition of Christian Democrats and Social Democrats, has agreed a focus on quality, especially in hospitals.

In international terms, the German health-care system has a generous benefit basket, one of the highest levels of capacity as well as relatively low levels of cost

*(continued)*

**Box 5:** (continued)

sharing. Expenditure per capita is relatively high, but expenditure growth since the early 2000s has been modest in spite of a growing number of services provided both in hospital and ambulatory care, an indication of technical efficiency. In addition, access is good – evidenced by low waiting times and relatively high satisfaction with out-of-hours care.

However, the German health-care system also shows areas in need of improvement if compared with other countries. This is demonstrated by the low satisfaction figures with the health system in general; respondents see a need for major reform more often than in many other countries. Another area is quality of care, in spite of all reforms having taken place. Germany is rarely placed among the top OECD or EU15 countries, but usually around average, and sometimes even lower.

In addition, the division into SHI and PHI remains one of the largest challenges for the German health-care system – as risk pools differ and different financing, access, and provision lead to inequalities.

Source: Busse and Blümel (2014)

## The Need to Signal Credibility

If evidence is to be used, the reader needs to have confidence in it. Using expert inputs and consultation in developing the template can support this. External review stages of the HiT are of course also important and ideally will include academic and practitioner perspectives. It is also crucial that any review by governments or other authorities with a potential conflict of interests is handled in such a way that it is not seen to compromise the integrity of the work. Professional presentation, launches and links to major events, as well as other efforts to “publicize” the materials may also enhance the reputation of the work. Those developing comparative frameworks will also

have to be clear about the sources of data they use, their quality, and the extent to which they are compatible with each other.

## The Need to Build in a Review Process

The experience of the Observatory has shown the value of a comprehensive review process for developing templates for health system comparisons. While it is clear that consulting widely brings new perspectives and creates acceptance for a model, it does run the risk of diluting the framework’s focus. The Observatory has found that making it clear in advance that there are space constraints and giving those consulted some explanation of how or why their suggestions have been acted on (or not) lessens the pressure to expand the framework indefinitely and helps those consulted see that their inputs are valued even if they are not always used.

---

## Conclusions

The HiT series is, at least in Europe, in many respects a “gold standard” for comparing health systems. It has a long and positive track record with HiTs for 56 European and OECD countries, often in several editions, and a total of some 130 HiTs overall. It has made information on health systems and policies publicly available in a format that cannot be found elsewhere and supported comparative analysis across countries, including analytic studies, more detailed country case studies, and explicitly comparative works, for example on countries emerging from the Soviet Union (Rechel et al. 2014), the Baltic states (van Ginneken et al. 2012), the central Asian countries (Rechel et al. 2012), or the Nordic countries (Magnussen et al. 2009). HiTs are some of the most downloaded documents held on the WHO web site and are used not just in Europe but beyond. They have served as a guide for the Asia Pacific Observatory on Health Systems and Policies (which was mentored by the European Observatory and launched in 2011) which uses an adapted version of the template to produce

country reviews for its region. The average impact factor of (European Observatory) HiTs, calculated internally using Thomson Reuters methodology, was 3.6 between 2012 and 2014, with a high of 4.26 in 2013 although this only captures citations in journals listed on PubMed/Medline. Google Scholar, which also recognizes the gray literature, shows that some HiTs achieve several hundred citations per edition.

The Observatory's experience with HiTs has generated insights that others developing frameworks for health system comparison might usefully draw on. It demonstrates the importance of a user-friendly template that helps authors and editors produce accessible, relevant, and credible outputs with a focus on what is expected from the comparisons and on who is going to use them. However, it also suggests that no template is perfect. There are different ways of categorizing and grouping key functions (of a health or any other system) or of conceptualizing systems and different levels of tackling and reporting evaluation. To some extent these are a matter of preference. There are also and always tradeoffs between comprehensiveness and accessibility, completeness and timeliness, and inclusiveness and readability. The current HiT template can be seen as a pragmatic trade-off based on almost 20 years of experience. How other teams chose to balance these will depend on the focus of their comparisons and the people who are to use their work.

The Observatory has also found ways of combining (excellent) national authors with its own technical editors. This is not always straightforward as not all European countries have the same capacity in health system research and national experts with strong analytical and English writing skills can be hard to find (Santoro et al. 2016) and may move on rapidly. Moreover, HiT and HSPM authors are not normally remunerated but, at "best," receive only small honoraria. The HiT series has addressed these challenges by identifying and linking formally with leading institutions, cultivating long-term relationships with HiT author teams, and, most recently, through its HSPM network. This mix of approaches may have helped build capacity in countries. It has certainly developed the understanding and

research (and people) management skills of the editorial team. Other comparative initiatives with limited resources might also want to consider what they can do in terms of sharing ownership and recognition to create non-monetary incentives for national counterparts and to develop their own team.

Comparability is and will remain a challenge, despite the standard template, tables, and figures, and is likely to be an issue for all other comparative projects. This is somewhat obvious when it comes to quantitative data given the divergent geographic coverage of international databases and the differences in definitions and data collection methods, not to mention the challenges at the individual country level. While it is clear that caution must be exercised when comparing quantitative data from different sources, it is also true, if less obvious, that qualitative data and the descriptive elements of the HiTs raise issues of comparability. In some areas there are broadly accepted tools (OECD et al. 2011) that help, but in many there are no agreed standard definitions (with health professionals being a case in point). Other comparative projects will need both to draw on the latest available knowledge and frameworks and to invest in methodological work as the Observatory team has done, for example, with the conceptual model (the three-dimensional cube) to explain coverage (Busse et al. 2007; Busse and Schlette 2007). They will also need to tailor responses to data and evidence availability in parts of Europe (particularly but by no means exclusively in central, eastern, and southeastern Europe) and to hope that EC/OECD/WHO initiatives on data will ultimately fill the gaps. There will still and inevitably be differences in the information available in countries, in the issues which are important to them, and in the interests and strengths of authors. Those developing frameworks for comparison will have to address these tensions in light of their overarching objectives and in the knowledge that health systems are constantly evolving. They may also find, as the Observatory has, that a comparative framework simply cannot capture everything and that analysis for more specialized issues may require separate study.



Despite the challenges, the Observatory would hold that there is real value in a framework for health system comparison, particularly one that relates to a defined “user” need and which can be sustained over time. Much follows from knowing who will use a set of comparisons and why. Longevity allows a framework to evolve – to improve, strengthen comparability, and build up successive levels of knowledge. Combining the two means a framework can move beyond the descriptive to the truly evaluative so that it captures and assesses aspects of health system performance in ways that speak to policy-makers or the research community or, ideally, both.

## References

- Busse R, Blümel M. Germany: health system review. *Health Syst Transit*. 2014;16(2):1–296.
- Busse R, Schlette S, editors. Health policy developments issue 7/8: focus on prevention, health and aging, and human resources. Gütersloh: Verlag Bertelsmann Stiftung; 2007.
- Busse R, Schreyögg J, Gericke CA. Analyzing changes in health financing arrangements in high-income countries: a comprehensive framework approach, Health, Nutrition and Population (HNP) discussion paper. Washington, DC: World Bank; 2007.
- Catalo C, Lavis J, The BRIDGE study team. Knowledge brokering in public health. In: Rechel B, McKee M, editors. *Facets of public health in Europe*. Maidenhead: Open University Press; 2014. p. 301–16.
- Duran A, et al. Understanding health systems: scope, functions and objectives. In: Figueras J, McKee M, editors. *Health systems, health, wealth and societal well-being: assessing the case for investing in health systems*. Maidenhead: Open University Press; 2012. p. 19–36.
- Figueras J, Tragakes E. Health care systems in transition: production template and questionnaire. Copenhagen: World Health Organization Regional Office for Europe; 1996.
- Magnussen J, Vrangbak K, Saltman RB, editors. *Nordic health care systems. Recent reforms and current policy challenges*. Maidenhead: Open University Press; 2009.
- Mossialos E, Allin S, Figueras J. Health systems in transition: template for analysis. Copenhagen: WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies; 2007.
- OECD, Eurostat, WHO. A system of health accounts. Paris: OECD Publishing; 2011. <https://doi.org/10.1787/9789264116016-en>.
- Rechel B, Thomson S, van Ginneken E. Health systems in transition: template for authors. Copenhagen: WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies; 2010.
- Rechel B, et al. Lessons from two decades of health reform in Central Asia. *Health Policy Plan*. 2012;27(4):281–7.
- Rechel B, Richardson E, McKee M, editors. *Trends in health systems in the former Soviet countries*. Copenhagen: World Health Organization; 2014 (acting as the host organization for, and secretariat of, the European Observatory on Health Systems and Policies).
- Richards T. Europe’s knowledge broker. *BMJ*. 2009;339: b3871.
- Santoro A, Glonti K, Bertollini R, Ricciardi W, McKee M. Mapping health research capacity in 17 countries of the former Soviet Union and South Eastern Europe: an exploratory study. *Eur J Pub Health*. 2016;26:349–54.
- van Ginneken E, et al. The Baltic States: building on 20 years of health reforms. *BMJ*. 2012;345:e7348.



# Health Services Knowledge: Use of Datasets Compiled Retrospectively to Correctly Represent Changes in Size of Wait List

# 14

Paul W. Armstrong

## Contents

<b>Introduction</b> .....	298
<b>Why Does the Waiting List Shrink (or Swell)?</b>	
<b>The Primary Hypothesis</b> .....	302
<b>What Happens to Enrolment and Admission in a Waiting List Initiative?</b> .....	302
<b>Does Size Shrink if Admission Exceeds Enrolment (and Does Size Swell if Enrolment Exceeds Admission)?</b> .....	304
In South Glamorgan, Wales .....	304
In INSALUD, Spain .....	305
In England .....	307
In Victoria, Australia .....	308
In Winnipeg, Canada .....	309
In Sweden .....	311
In England .....	312
<b>The Balance of Enrolments and Admissions (Plus Other Removals) Equals the Change in Size. Why?</b> .....	319
If the Model Is Not Complicated, the Data Must Be Simple! .....	319
The Number of ‘Starts’ and ‘Stops’ Must Be the Same .....	324
<b>Secondary Hypotheses</b> .....	325
Inexplicably Complicated .....	325
Supplier-Induced Demand .....	325
<b>Why has the Effect of Enrolment Confounded Analyses to Date?</b> .....	326
Some Assumed Enrolment Was Fixed and Unvarying .....	326
Some Only Registered Discharge (and Death) .....	328
Some Compiled Returns .....	333
Some Made Hay .....	338

---

P. W. Armstrong (✉)  
London, UK  
e-mail: [P.W.Armstrong@outlook.com](mailto:P.W.Armstrong@outlook.com)

<b>The Primary Hypothesis Has Not Been Falsified</b> .....	339
<b>References</b> .....	340

### Abstract

This chapter introduces two items which are mandatory for any dataset which hopes to support the analysis of waiting lists. It was once understood that the direction and extent of any change in size was determined by the balance of ‘enrolments’ and ‘admissions.’ We can assess the effect on the size of the list of any increase (or decrease) in the number of admissions, if we know the numbers on the list at two points in time (at the beginning, and at the end, of the period) and if we know the number enrolled on the list and the number admitted from it during the interval. In the chapter, we show that one cannot determine how changes in the number of admissions affect the size of the list, if the number of enrolments is not known or we do not know how it has changed.

### Introduction

It is not always possible to start treatment the moment that a clinician decides it is desirable. Delay is sometimes unacceptable, and the work of the clinician is arranged to expedite the assessment, investigation, or treatment of such cases wherever possible. But delay is sometimes acceptable. Patients experiencing such delay are said to require assessment, investigation, and treatment on an elective basis and to belong to the waiting list. The waiting list identifies all of those in this state of limbo at any particular moment in time. It functions as an order book, allowing clinicians to keep track of their outstanding obligations.

The limits of each delay are defined by a ‘start date’ and an ‘end date.’ There are a variety of these to choose from. For example, the start date might be the date of receipt of a referral, and the end date might be the date of the relevant consultation, if we are interested in the wait for an expert opinion. Or the start date might be the date on which the

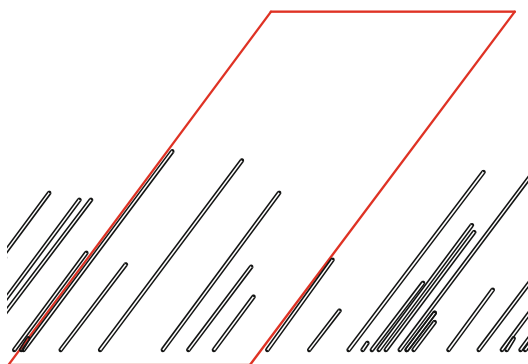
clinician recommended – and the patient agreed to – hospitalization, i.e., the date of the clinician’s ‘decision to admit’ to the list, and the end date might be the date of the relevant admission, if we are interested in the wait for investigation or treatment on an inpatient or day-case basis.

The delay is the interval between the start and end dates. This interval tends to lengthen whenever it involves the coordination of multiple players or the scheduling of a scarce resource. It is helpful to visualize the delay as a line connecting the start and end dates. Demographers refer to this as a ‘lifeline’ (Hinde 1998). It is particularly helpful if lifelines are orientated to display the passage of time (on the horizontal axis) and the acquisition of experience (on the vertical axis) in what is known as a Lexis diagram (Hinde 1998).

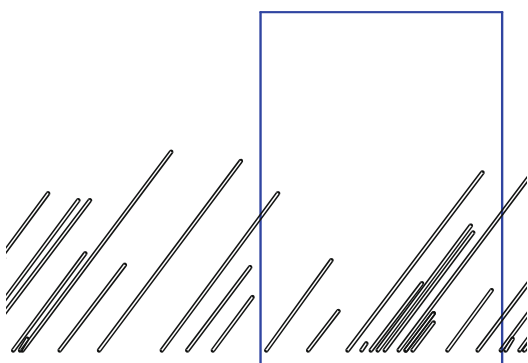
The number of start dates and end dates in any cohort must be the same because they are two different ways of counting the same set of completed lifelines. There are seven ‘starts’ and seven ‘stops’ in Fig. 1. But it is more difficult to enumerate the relevant start and end dates when we restrict attention to the lifelines eligible over any period.

Some lifelines had a start date during the period of interest (Fig. 2). We will refer to these as the newly ‘enrolled’ ( $E = 13$ ) on the waiting list. Others had a start date which preceded the earlier census: the end date of some of those also preceded the census, and they are of no further interest, but the end date of others succeeded the census. We will refer to these as the lifelines counted at the time of the earlier census ( $C^{\text{then}} = 1$ ). In this example,  $C^{\text{then}} + E = 14$  lifelines became *eligible*, e.g., for elective admission, at some point during the Period.

Some lifelines had an end date during the period (Fig. 2). We will refer to these as the newly ‘admitted’ ( $A = 11$ ) from the list. Others had an end date which succeeded the later census: the start date of some of those also succeeded the census, and they are of no further interest, but the start date of others preceded the census. We will



**Fig. 1** Counting the start and end dates in a cohort



**Fig. 2** Counting the start and end dates in a period

refer to these as the lifelines counted at the time of the later census ( $C^{now} = 3$ ). In this example,  $A + C^{now} = 14$  lifelines became *ineligible* at some point during the period.

It is evident that

$$C^{then} + E = A + C^{now}. \tag{1}$$

There are several conditions which have to be satisfied if we want our set of records to provide a valid representation of even the simplest waiting list.

- The start date must be **complete**. If it is the decision to admit which determines whether or not a patient has been added to the waiting list, the date of the decision to admit must be complete; otherwise, the dataset might exclude someone who required elective admission.

- The start and end dates must be **accurate**. The date recorded as the date of the decision to admit to the list must be the same as that on which the patient and clinician agreed to elective investigation or treatment on an inpatient or day-case basis. The date recorded as the date of admission from the list must be the same as that on which the patient was admitted to hospital for elective investigation or treatment. Otherwise, those considered eligible on a specified date might exclude some who were in fact available and might include others who were not.
- No record must be **duplicated**.

Additional items of data may be needed if we want to assess whether delay is acceptable (or not), to determine who should be invited ‘to come in’ next or to allow the comparison of like lifelines with like. Our records must show the same entry as the waiting list in each of these fields.

But there may be difficulty demonstrating that the number of starts equals the number of stops (formula (1)), even if the records are entirely accurate (Goldacre et al. 1987). We cannot expect to obtain counts which are consistent unless they enumerate the same lifelines, e.g., those which represent individuals eligible for the same service and in the same period. Moreover, the counts may appear inconsistent if the records are more complicated than the model. We have assumed (a) that there is never more than one record per patient, (b) that no one joins the list without eventually receiving the service desired, and (c) that the interval between the start and end dates is unbroken. So if any lifelines are removed ( $R$ ) from the list without experiencing the outcome desired ( $A$ ), counting the stops will not give the same result as counting the starts.

But if we are not able to demonstrate empirically that the number becoming eligible exactly equals the number becoming ineligible (formula (1)), we must doubt our processing of the records. This is not trivial. The implications for the subsequent analysis are important.

The number of starts must equal the number of stops, that is,

$$C^{\text{then}} + E = A + C^{\text{now}}. \quad (1)$$

If we subtract  $A$  from both sides of formula (1), we obtain

$$C^{\text{then}} + E - A = C^{\text{now}}. \quad (2)$$

The size of the list when the period closes ( $C^{\text{now}}$ ) is determined by the size of the list when the period opened ( $C^{\text{then}}$ ) and by the balance of enrolments ( $E$ ) and admissions ( $A$ ) during the interval. Formula (2) (Moral and de Pancorbo 2001) is a simplified version of what is known as the basic demographic equation (Newell 1988) or the balancing equation (Pressat 1985).

If we subtract  $C^{\text{then}}$  from both sides of formula (2), we obtain

$$E - A = C^{\text{now}} - C^{\text{then}}. \quad (3)$$

Formula (3) indicates that it is the net *in*-flow ( $E - A$ ) which determines the direction and extent of any change in stock ( $C^{\text{now}} - C^{\text{then}}$ ).

If we add  $A$  to both sides of formula (3), we obtain

$$E = A + C^{\text{now}} - C^{\text{then}}. \quad (4)$$

Formula (4) has been used to estimate the number of enrolments when the relevant count is not available (Naylor et al. 1997).

If we subtract  $E$  from both sides of formula (4), we obtain

$$0 = A + C^{\text{now}} - C^{\text{then}} - E. \quad (5)$$

If the four counts enumerate the same set of lifelines and are accurate, formula (5) gives an ‘error of closure’ of zero (Newell 1988).

If we subtract  $C^{\text{now}}$  from both sides of formula (1), we obtain

$$C^{\text{then}} + E - C^{\text{now}} = A. \quad (6)$$

It is often asserted that the size of the list ought to shrink if there is an increase in the number of

admissions (Naylor 1991). This ‘primary hypothesis’ is often used to justify requests for additional resources when the size of the list is thought to be a problem. But this relationship has proven so difficult to substantiate that health economists (and others) have cast around for ‘secondary hypotheses’ which better fit the available data.

Formula (3),  $E - A = C^{\text{now}} - C^{\text{then}}$ , indicates that the effect of admissions may be confounded by the effect of enrolments. The size of the list may swell despite an increase in the number of admissions, if enrolments exceed admissions; and the size of the list may shrink despite a reduction in the number of admissions, if admissions exceed enrolments.

The literature includes a number of studies in which investigators found little evidence of an inverse relationship between the number of admissions and changes in the size of the list (Fowkes et al. 1983; Goldacre et al. 1987; Niinimäki 1991; Harvey et al. 1993; Nordberg et al. 1994; Newton et al. 1995). Some investigators were unwilling to surrender the hypothesis. They preferred to attribute the results to the effect of enrolments and were prepared to infer a pattern of variation in the number of enrolments consistent with their data. But few have assembled the information needed to test this surmise. Few supplied their readers with a qualitative judgment as to whether the number of enrolments was fixed or varying, and few supplied quantitative data such as the numbers enrolled from one period to the next (White 1980; Newton et al. 1995; Street and Duckett 1996; Armstrong 2000; NAO 2001b; Moral and de Pancorbo 2001; House of Commons Health Committee 2010; Armstrong 2010; Kreindler and Bapuji 2010).

We cannot use the relationship implied by formula (4) to obtain estimates of the numbers enrolled and then use these estimates to test the relationship implied by any of the other formulae. The estimate is necessarily consistent with the count of admissions and with changes in size under formula (4) and therefore cannot provide independent verification of another version of the formula, e.g., formula (3). So the authors of these studies assert that it is not possible to evaluate the effect of admissions without making some allowance for the

effect of enrolment and find that they are unable to test the primary hypothesis having failed to collect the necessary data.

Other investigators chose to infer that the relationship did not have the form hypothesized when they found no evidence of an inverse relationship between the number of admissions and changes in the size of the list (Feldstein 1967; Culyer and Cullis 1976; Snaith 1979; Buttery and Snaith 1980; Kenis 2006).

Culyer and Cullis (1976, 244) invoked “Say’s Law of Hospitals . . . that additions to the supply of inpatient capacity create equal additions to the demand for that capacity.” “[A]s the price of a good or service is lowered . . . the quantity demanded in any . . . period . . . will rise” (p. 244), so the authors hypothesized that an increase in “throughput capacity” would be accompanied by a fall in “mean waiting time” and “time price” which “would lead one to expect demand to increase” (p. 247). As a result, the authors expected “a positive relation between throughput capacity and waiting lists” (p. 247), and they reported significant direct correlations between these two variables for seven (out of 15) “hospital regions” over time (p. 249). Under this hypothesis, an increase in the number of admissions is about the worst line a policy-maker can take if his/her goal is to reduce the size of the list.

Buttery and Snaith (1980, 58) hypothesized “a self-regulating system . . . in which waiting times for patients and waiting lists per surgeon are relatively constant” (Feldstein 1967; Smethurst and Williams 2002). They thought that “[w]aiting times must provide a constraint on unmet need, preventing patients from coming forward or surgeons from putting them on their waiting lists if they do.” In other words, the number of enrolments shrank to counter any increase in wait as a result of any decrease in the number of admissions; and the number of enrolments swelled in response to any decrease in wait as a result of any increase in the number of admissions. They understood that if such a system exists, a “further increase in the number of surgeons will further increase the national waiting list” and “a diminution [will] reduce it.”

Culyer and Cullis (1976) imagine enrolments driven by admissions, i.e., ‘demand’ for a service induced by its ‘supply,’ and Buttery and Snaith (1980) imagine enrolments constrained by the length of wait. The former list is driven by the public’s appetite for consumption and the latter by the clinicians’ desire to limit their commitments. But the mechanisms and the outcomes envisaged are the same. An increase in the number of admissions is thought to reduce the length of wait, a reduction in the length of wait is thought to increase the number of enrolments, and an increase in the number of enrolments is thought to increase the size of the list. It is change in the number of admissions which is thought to evoke change in the number of enrolments in both instances, and both hypotheses predict a direct correlation between admissions and enrolments.

Enrolments are thought to contribute to self-regulation and to supplier-induced demand, but neither pair of investigators assembled data which allowed them to confirm whether the number of enrolments was fixed or to establish whether any variation had the pattern hypothesized. Instead, they assume relationships which are consistent with the primary hypothesis. They expect the size of the list to swell if there has been an increase in the number of enrolments, and they expect the size of the list to shrink if there has been a decrease in the number of enrolments.

The authors of most of these investigations dismissed the primary hypothesis without fair trial. It would have been reasonable to restrict attention to the effect of admissions on size – and to draw conclusions accordingly – when the effect of enrolment had been given no thought. It would also have been reasonable to restrict attention to the effect of admissions on size when counts of enrolments were thought to be unvarying. But it was not reasonable to dismiss the primary hypothesis without attempting to adjust for enrolments (formula 3) once the effect had been surmised and the variation acknowledged.

The relationship between the change in the size of the list and the balance of enrolments and admissions over a period is simple, not complex; exact, not approximate; and mathematical, not behavioral. The relationship is not affected by

the location of the delay (outpatient or inpatient/day case) or its cause (assessment, investigation, or treatment); by diagnosis or procedure; by clinician, specialty, or provider; or by any other classification of the lifelines. So the seeming variety of our case studies contributes a little color to the account but adds nothing to the veracity of the argument presented in formulae (1) through (5).

An increase in the number of admissions may either make matters better, under the primary hypothesis, or else may make matters worse, under the secondary hypothesis. The two views have influenced the thinking of contemporary commentators (Carvel 2004) and policy-makers but are contradictory. We wish to establish which studies provide relevant data, whether the results are trustworthy or suspect and whether they confirm or refute the primary hypothesis.

The literature is clogged with citations. So, in this chapter, we have given the floor to studies which use empirical data to explore the effect of admission on the size of the list while allowing for the effect of enrolment. We have not knowingly omitted any study which offers relevant data. We are widely read, and we have used reviews (Faulkner and Frankel 1993; Sanmartin et al. 1998; Hurst and Siciliani 2003; Finn 2004; Kreindler 2010) and reference lists (Harvey et al. 1993) to identify potentially relevant material. We are looking for one, well-substantiated, exception to the rule, a set of data which invalidates formula (3). But the literature is extensive, and we have not had time to run a systematic search of our own. Nevertheless, it should be easy to find examples given the eagerness with which alternative hypotheses have been adopted and the primary hypothesis dismissed.

---

### **Why Does the Waiting List Shrink (or Swell)? The Primary Hypothesis**

A number of authorities claim that it is the balance of enrolments and admissions which determines whether there is an increase (or a decrease) in the size of a waiting list (DHSS 1981a; Sykes 1986; Naylor 1991; Worthington 1991; Street and Duckett 1996; Hanning and Lundström 1998;

Torkki et al. 2002; NWTU 2003; National Audit Office Wales 2005; Kenis 2006; Kreindler 2010). They represent a variety of stakeholders, e.g., clinicians (DHSS 1981a; Sykes 1986; Naylor 1991; Hanning and Lundström 1998; Torkki et al. 2002), managers (Worthington 1991; Street and Duckett 1996; Hanning and Lundström 1998; Kenis 2006), and policy-makers (DHSS 1981a; NWTU 2003); and they represent a variety of paradigms, e.g., health economics (Street and Duckett 1996; Hanning and Lundström 1998), organizational science (Kenis 2006), and system dynamics (Worthington 1991). We thought it remarkable that they should all agree: when a consensus is the result of independent (and rigorous) evaluation by various stakeholders and different disciplines, their agreement adds weight to the evidence. But independent (and rigorous) evaluation is not the only way in which we reach a consensus. Some authors also claim that the relationship (formula 3) has never been observed in practice (Culyer and Cullis 1976). Given that the proposition of a second hypothesis implies the failure of the first, it is perhaps not surprising that this view is one echoed by many of those who have contributed to this literature. So the variety of the stakeholders provides no assurance of the independence of their judgment if it is the failure of the first hypothesis on which they are all agreed; and the variety of the paradigms provides us with no assurance of the rigor of their evaluation if the failure of the first hypothesis is assumed by each approach. We begin this chapter by conducting a fresh assessment of claims that the primary hypothesis has failed.

---

### **What Happens to Enrolment and Admission in a Waiting List Initiative?**

Very few researchers have reported the number of enrolments. We know of only seven instances. Four appeared in print having been subject to peer review (White 1980; Street and Duckett 1996; Armstrong 2000, 2010), and three are contributions to the grey literature, two of which are in the public domain (Hamblin et al. 1998; Moral and de Pancorbo 2001) and the other of which is not (Kreindler and Bapuji

2010). This observation indicates that most researchers have not examined the effect of the balance of enrolments and admissions on the size of a waiting list and implies that the validity of the first hypothesis has not been widely evaluated.

Many researchers believe that an increase in the numbers admitted from a waiting list ought to be accompanied by a decrease in its size (Culyer and Cullis 1976; Goldacre et al. 1987; Naylor 1991). They rarely declare that enrolment may be a factor (Newton et al. 1995) or state that the number of enrolments is assumed to be fixed and unvarying, i.e., stationary (MoH 1963a). Instead, the effect of enrolment is conceded only when researchers are obliged to explain how an increase in the numbers admitted from a waiting list appears to have been accompanied by an increase in its size (MoH 1964; Goldacre et al. 1987; Hamblin et al. 1998; Sanmartin et al. 1998). (The size of the list and the numbers admitted may both increase yet still be consistent with the primary hypothesis if enrolment exceeded admission.) But at the close of their investigations, these researchers are unable to substantiate the claims they wish to make because they neither eliminated the variation in enrolments nor did they collect the counts which would have allowed them to adjust for it. The limitations of these studies have not been sufficiently appreciated. Their results have nothing to contribute to our understanding of the relationship between enrolments, admissions, and the size of the list. Their discussions have nothing to contribute to our methodology because they fail to acknowledge that variation in enrolment confounds the apparent relationship between admissions and the size of the list (Newton et al. 1995).

Some analysts anticipate that an increase in expenditure, intended to increase the numbers admitted from the waiting list, ought to reduce its size. So the Irish Minister for Health and Children authorized the expenditure of an additional €246 million between 1993 and 2002 (Purcell 2003), on the understanding that this would buy substantial numbers of additional elective procedures and, as a consequence, would reduce the numbers who had been on the list for a long time. The Comptroller and Auditor General found invoices for work in the private sector

commissioned by the hospitals he visited but, apart from these (Purcell 2003), was obliged to confess that “it is difficult to verify the reported number of procedures carried out under the [waiting list initiative] and to ascertain the extent to which they are over and above core-funded activity” (Purcell 2003, 26). (© Government of Ireland 2003.) He also reported that “the overall level of elective inpatient treatment ... fell between 1998 and 2001,” which “suggests that the Initiative did not result in an increase in elective inpatient activity over and above existing levels” (Purcell 2003, 28). (© Government of Ireland 2003.) It is possible that the fault was due to a failure in the system of bookkeeping, that financial control mechanisms were blameless (Purcell 2003), and that additional activity failed to have the effect desired. But before accepting that this is the case, we would like to know whether the funds awarded to each hospital were apportioned in line with the intended contribution of each department and whether these additional resources appeared under the appropriate budget headings in time to pay for the activity planned. If a study does not quantify the effect of additional expenditure on elective admission, we doubt its ability to provide empirical evidence about the effect of elective admission (or additional expenditure) on the size of the list. Newton et al. (1995, 784) report that “[i]n only six [out of 44 waiting list initiatives] was additional funding followed by a rise in admissions and a fall in list size.”

Some commentators believe that an increase in the amount of a resource, particularly one thought to be in critically short supply, ought to reduce the size of the waiting list (DHSS 1975, 1981a). Our evaluation of the effect of an increase in such a resource proceeds along the same lines as our evaluation of the effect of an increase in expenditure. We expect a reduction in the size of the list only when the number of admissions exceeds the number of enrolments. We therefore want to know what effect the increase in resources had on the number of admissions (and on the number of enrolments). This data is the minimum required for any evaluation, and the effect of a waiting list initiative on the size of the list



cannot be established without it. Regrettably, this information has rarely been assembled for the benefit of readers (Hamblin et al. 1998).

### Does Size Shrink if Admission Exceeds Enrolment (and Does Size Swell if Enrolment Exceeds Admission)?

The veracity of the primary hypothesis can only be tested by studies which report the number of enrolments alongside the number of admissions and changes in the size of the list. These studies are therefore rather more important than has hitherto been recognized.

#### In South Glamorgan, Wales

White (1980) reports a study of the combined elective activity of three or four consultants in one surgical specialty (unspecified) at one public hospital in South Glamorgan, Wales. He thinks that the size of the outpatient waiting list ought to have something to do with the number of new outpatients seen and the number of GP referrals (White 1980), and he uses column charts to explore the relationship. He seems also to have thought that the size of the inpatient waiting list might have something to do with the number of discharges (or deaths) and the number of new outpatients seen.

White (1980) does not make full use of his data. So he looks for a relationship between the size of the outpatient waiting list and new outpatients seen and between the size of the outpatient waiting list and GP referrals, but he does not consider the combined effect on the size of the outpatient waiting list of new outpatients seen and GP referrals. Worse, he looks for a relationship between the size of the inpatient waiting list and deaths and discharges, but he does not consider even the univariate effect of new outpatients seen.

White (1980) does not reason correctly from the data he has assembled. He is not satisfied with the relationship he observes between the size of the outpatient waiting list and the number of new

outpatients seen, but he does not express the same dissatisfaction with the relationship he observes between the size of the inpatient waiting list and discharges (or deaths). In the former instance, he attributes inconsistency to those who walk-in without having been entered on the list: in other words, he understands that not every new outpatient seen represented a unit reduction in the numbers waiting (White 1980). In the latter, he does not acknowledge the difference between the number of discharges (or deaths) and the number of elective admissions or between the number of decisions to admit to the list and the number of new outpatients seen. Instead, he expresses himself satisfied that “[f]ewer deaths and discharges in the specialty coincide with a lower in-patient waiting list” which “indicates that long in-patient waiting lists combine with greater in-patient activity” (White 1980, 274).

White (1980) uses surrogate measures to describe activity over 15 periods each of 3 months duration. He counts GP referrals rather than all referrals and referrals received rather than referrals accepted. He counts new outpatients booked rather than decisions to admit to the inpatient waiting list, and he counts discharges (or deaths) rather than elective admissions. Now, if it is the balance of enrolments and admissions which determines whether there is an increase (or a decrease) in the size of a waiting list, we would expect  $E - A = C^{\text{now}} - C^{\text{then}}$  (3). But 2.39 new outpatients were booked per discharge (or death). Therefore, where  $E$  represents new outpatients booked,  $A$  represents discharges (or deaths), and  $C$  represents the size of the inpatient waiting list, we do not expect  $E - A$  to exactly equal  $C^{\text{now}} - C^{\text{then}}$  (Table 1, right-hand side). However, if the surrogate measures have the effect anticipated on the size of the waiting list, we would expect a direct correlation between the two sides of formula (3), i.e.,  $E - A \propto C^{\text{now}} - C^{\text{then}}$ .

We obtained the quantities  $E - A$  and  $C^{\text{now}} - C^{\text{then}}$  by calculation from counts charted by White (1980), and we used Spearman's *rho* to assess the direction and strength of association between them. (We used the number of “deaths and

**Table 1** Was the change in size directly correlated with the balance of enrolments and admissions in South Glamorgan, Wales?

Year	Qtr	Waiting for out-patient assessment					Waiting for in-patient admission				
		No. of 'GP referrals'	No. of 'new out-patients booked'	Size of 'out-patient waiting list'	Net in-Flow	Change in Stock	No. of 'new out-patients booked'	No. of 'discharges and deaths'	Size of 'in-patient waiting list'	Net in-Flow	Change in Stock
		$E$	$A$	$C^{now}$	$E-A$	$C^{now}-C^{then}$	$E$	$A$	$C^{now}$	$E-A$	$C^{now}-C^{then}$
		[3]	[4]	[6]	[9]	[10]	[3]	[4]	[6]	[9]	[10]
1976	1	1,076	1,109	712	-33						
	2	1,112	923	813	189	1,109	363	290	746		
	3	1,197	1,296	495	-99	923	455	271	468	-19	
	4	1,028	1,117	105	-89	1,296	412	237	884	-34	
1977	1	1,080	826	346	254	1,117	451	233	666	-4	
	2	1,108	583	857	525	826	423	264	403	31	
	3	1,044	615	1,126	429	583	354	195	229	-69	
	4	1,068	575	1,360	493	615	377	231	238	36	
1978	1	1,020	672	1,649	348	575	385	222	190	-9	
	2	1,229	555	2,076	674	672	346	207	326	-15	
	3	1,092	704	2,084	388	555	344	114	211	-93	
	4	1,205	1,385	1,489	-180	704	321	152	383	38	
1979	1	1,237	1,381	1,247	-144	1,385	300	222	1,085	70	
	2	876	583	1,505	293	1,381	247	237	1,134	15	
	3	1,036	704	1,505	332	583	358	203	225	-34	
						704	323	217	381	14	

Source: White (1980)

discharges” (White 1980, Fig. 14) as our surrogate for the number of admissions from the inpatient waiting list, and we used the number of “new out-patients booked” (White 1980, Fig. 8) as our surrogate for the number of enrolments on it. We used the number of “new out-patients booked” (White 1980, Fig. 8) as our surrogate for the number of admissions from the outpatient waiting list, and we used the number of “GP referrals” (White 1980, Fig. 2) as our surrogate for the number of enrolments on it.)

We found that the correlation for the inpatient waiting list – while not statistically significant – had the direction anticipated (Spearman’s  $\rho = +0.46$ ,  $n = 14$ ,  $p = 0.10$ ) and that the correlation for the outpatient waiting list was strong and statistically significant as well as having the direction anticipated (Spearman’s  $\rho = +0.82$ ,  $n = 14$ ,  $p < 0.01$ ). White’s data is compatible with formula (3) and the primary hypothesis – an increase in ‘admission’ (net of ‘enrolment’) may accompany a reduction in the size of the list. It is noteworthy that the author overlooked the effect of enrolment despite assembling relevant data when trying to identify “[w]hat [f]actors [i]nfluence [the size of w]aiting [l]ists” (White 1980, 270).

## In INSALUD, Spain

Moral and de Pancorbo (2001) report a study of the combined elective activity in six surgical specialties funded by INSALUD, Spain. When the initiative began, the waiting list comprised orthopedics (27%), general surgery (21%), ophthalmology (17%), ENT surgery (10%), urology (7%), gynecology (6%), and other specialties (12%).

Unlike White (1980), the authors do not report the inputs and outputs at some distance from the waiting list down the referral pathway. Instead, they count “entries” to the list, i.e., that number by which the count of patients on the list ought to swell; and they count “exits” from the list, i.e., that number by which the count of patients on the list ought to shrink. Moral and de Pancorbo (2001, 48) use counts of “entries” and “exits” to describe activity over four periods each of 12 months duration.

Moral and de Pancorbo (2001) emulate the approach modeled by White (1980). They look for a relationship between the size of the list and the number of “exits”, but they do not consider the effect of the number of “entries”. We obtained the quantities  $E - A$  and  $C^{now} - C^{then}$  from counts charted by Moral and de Pancorbo (2001), and we

used Spearman’s *rho* to assess the direction and strength of association between them (Spearman’s  $\rho = +1.00, n = 4, p < 0.01$ ).

The error of closure reports the difference between the number of dates of entry and the number of dates of exit (1) often as a percentage of all of those eligible for admission during the period of interest, i.e.,

$$\text{error of closure (\%)} = 100 \times \frac{(C^{\text{then}} + E) - (A + C^{\text{now}})}{(C^{\text{then}} + E + A + C^{\text{now}})/2} \tag{5.1}$$

The initiative was associated with a reduction in the size of the list in its early years (in 1997 and 1998, according to  $E - A$  and to  $C^{\text{now}} - C^{\text{then}}$ ). But if we are prepared to credit the initiative with success in its early years – claiming that the initiative reduced the size of the list (Hanning and Lundström 1998) – we should also be prepared to credit it with failure in its later years – acknowledging that the initiative increased the size of the list (in 1999 and 2000, according to net in-flow and to change in stock).

The number of dates of entry ( $C^{\text{then}} + E$ ) did not equal the number of dates of exit ( $A + C^{\text{now}}$ ): the difference ranges from a shortfall of  $-15,148$  to a surplus of  $+2009$ . Although these differences are small, a little less than 2.5% when compared with the number of lifelines enumerated over the period, they should not occur and require some attempt at explanation. If there were a systematic error in one of the three counts, we would expect the direction and the extent of the error to be consistent.

- If the number awaiting surgery is always over-reported, e.g., by a factor of 1.05, the apparent change from one census to the next will correctly indicate whether the size of the list decreased or increased, but the size of the apparent change will be exaggerated by a factor of 1.05. If this were the only source of error,  $C^{\text{now}} - C^{\text{then}}$  would always be greater than  $E - A$  but would have the same direction.
- If the number of entries is always overreported, e.g., by a factor of 1.05, then  $E - A$  will always be too positive. When the size of the list is increasing,  $E - A$  will maximize the amount, and when the size of the list is decreasing,  $E - A$  will minimize the amount sometimes to the extent of reporting an increase in size where there has been a decrease.
- If the number of exits is always overreported, e.g., by a factor of 1.05, then  $E - A$  will always be too negative. When the size of the list is increasing,  $E - A$  will minimize the amount sometimes to the extent of reporting a decrease in size where there has been an increase, and when the size of the list is decreasing,  $E - A$  will maximize the amount.

Unfortunately, none of these scenarios fit Table 2 in which  $E - A$  is more negative than  $C^{\text{now}} - C^{\text{then}}$  in the first and second periods, less positive in the third, and more positive in the fourth. This implies either that there is systematic error in more than one count or that the error is not systematic.

There are several problems with the counts available. We have not been able to reconcile the

**Table 2** Did the balance of “entries” and “exits” adequately account for the change in size in INSALUD, Spain?

Year	Waiting for Admission					error of closure			
	No. of “entries”	No. of “exits”	Size of list	Net in-Flow	Change in Stock	Counting dates of entry	Counting dates of exit	difference	(%)
	<i>E</i>	<i>A</i>	<i>C<sup>now</sup></i>	<i>E</i> − <i>A</i>	<i>C<sup>now</sup></i> − <i>C<sup>then</sup></i>	<i>C<sup>then</sup></i> + <i>E</i>	<i>A</i> + <i>C<sup>now</sup></i>		
	[3]	[4]	[6]	[9]	[10]	[11]	[12]	[13]	[14]
30-Jun-96			190,000						
31-Dec-96			165,735		−24,265				
31-Dec-97	445,816	478,452	148,247	−32,636	−17,488	611,551	626,699	−15,148	−2.45
31-Dec-98	489,331	509,414	132,221	−20,083	−16,026	637,578	641,635	−4,057	−0.63
31-Dec-99	557,950	552,929	141,827	5,021	9,606	690,171	694,756	−4,585	−0.66
31-Dec-00	616,527	598,117	158,228	18,410	16,401	758,354	756,345	2,009	0.27

Source: Moral and de Pancorbo 2001

heights of the columns representing exits from the target population with the numbers reported in the text. (This undermines our confidence in the authors' presentation of their data.) The chart records the suspiciously tidy 190,000 as the size of the list in June 1996, whereas the text reports a count of 168,265. (We have chosen to tabulate the numbers obtained from the chart which provides information on entries as well as exits.) As a consequence, we report a change in stock of -24,265 rather than of -2530, but this affects neither the correlation nor the error of closure. More importantly, we have had to read the counts of entries and exits off the printed version of the column chart. We enlarged this so that 1 mm represented 1674 patients on the vertical axis instead of 9412. The errors of closure are therefore equivalent to heights of 9.0, 2.4, 2.7, and 1.2 mm for the periods 1997, 1998, 1999, and 2000, respectively. If a measurement may be out by as much as ±1.0 mm, then  $E - A$  and  $C^{now} - C^{then}$  may be out by as much as ±2.0 mm, and the error of closure by as much as ±4.0 mm. So one of these differences is not trivial. The waiting list initiative claims to have funded an additional 35,883 surgical procedures in 1997, when it recorded 15,148 too many exits (or too few entries) for the change in size observed.

The data provided by Moral and de Pancorbo (2001) is compatible with formula (3) and the primary hypothesis – an increase in “exits” (net of

“entries”) may accompany a reduction in the size of the list.

### In England

Hamblin et al. (1998) tabulate counts which describe activity over six periods each of 12 months duration and invite their readers to examine “[t]he effects of the Waiting Time Initiative” (1998, 13). They supply three different counts of ‘enrolments,’ two different counts of ‘admissions,’ and a count of the numbers awaiting elective admission on a day-case, or an inpatient, basis. When we used their counts of “[s]pecialist referring ... with no date” as a measure of enrolment, and “[w]aiting list episodes” as a measure of admission, we obtained a perfect correlation between the change in stock and the net in-flow (Table 3a: Spearman’s  $\rho = +1.00$ ,  $n = 5$ ,  $p < 0.01$ ).

Similarly, when we used “[t]otal elective episodes” as a measure of admission, and combined “[s]pecialist referring ... with no date” and “[s]pecialist referring ... with date” as a measure of enrolment, we obtained a perfect correlation between the change in stock and the net in-flow (Table 3b: Spearman’s  $\rho = +1.00$ ,  $n = 5$ ,  $p < 0.01$ ).

We consider this result suspicious although it is everything we are looking for. If ‘enrolments,’

**Table 3a** Did the balance of “[s]pecialist referring” and “episodes” adequately account for the change in size in England?

Year-end waiting list	Waiting for Admission to hospital					error of closure				$\hat{E}$
	'Specialist referring to waiting list with no date'	'waiting list episodes'	'waiting list size'	Net in-Flow	Change in Stock	Counting dates of entry	Counting dates of exit	difference	(%)	
	$E$	$A$	$C^{now}$	$E - A$	$C^{now} - C^{then}$	$C^{then} + E$	$A + C^{now}$			
	[3]	[4]	[6]	[9]	[10]	[11]	[12]	[13]	[14]	
1989/90	2,189,437	2,163,709	912,800	25,728			3,076,509			*
1990/91	2,094,683	2,101,089	906,394	-6,406	-6,406	3,007,483	3,007,483	0	0.00	2,094,683
1991/92	2,261,086	2,251,873	915,607	9,213	9,213	3,167,480	3,167,480	0	0.00	2,261,086
1992/93	2,362,393	2,283,026	994,974	79,367	79,367	3,278,000	3,278,000	0	0.00	2,362,393
1993/94	2,455,038	2,384,643	1,065,369	70,395	70,395	3,450,012	3,450,012	0	0.00	2,455,038
1994/95	2,493,649	2,514,977	1,044,041	-21,328	-21,328	3,559,018	3,559,018	0	0.00	2,493,649

Source: Hamblin et al. 1998

\*The authors were able to enter a value for 1989/90 in column 3, but we were unable to provide an estimate of the value for 1989/90 in the last column on the right. This suggests that the authors knew the ‘waiting list size’ for 1988/89 but opted not to report it.

**Table 3b** Did the balance of “[s]pecialist referring” and “episodes” adequately account for the change in size in England?

Year-end	Waiting for Admission to hospital					error of closure				$\hat{E}$
	'Specialist referring to ... with no date' or 'with date'	Total elective episodes'	'waiting list size'	Net in-Flow	Change in Stock	Counting dates of entry	Counting dates of exit	difference	(%)	
	$E$	$A$	$C^{now}$	$E-A$	$C^{now}-C^{then}$	$C^{then}+E$	$A+C^{now}$	[13]	[14]	
	[3]	[4]	[6]	[9]	[10]	[11]	[12]	[13]	[14]	
1989/90	3,361,737	3,336,009	912,800	25,728			4,248,809			*
1990/91	3,288,594	3,295,000	906,394	-6,406	-6,406	4,201,394	4,201,394	0	0.00	3,288,594
1991/92	3,684,057	3,674,844	915,607	9,213	9,213	4,590,451	4,590,451	0	0.00	3,684,057
1992/93	3,914,759	3,835,392	994,974	79,367	79,367	4,830,366	4,830,366	0	0.00	3,914,759
1993/94	4,065,606	3,995,211	1,065,369	70,395	70,395	5,060,580	5,060,580	0	0.00	4,065,606
1994/95	4,139,168	4,160,496	1,044,041	-21,328	-21,328	5,204,537	5,204,537	0	0.00	4,139,168

Source: Hamblin et al. 1998

\*The authors were able to enter a value for 1989/90 in column 3, but we were unable to provide an estimate of the value for 1989/90 in the last column on the right. This suggests that the authors knew the ‘waiting list size’ for 1988/89 but opted not to report it.

‘admissions,’ and ‘size’ had enumerated the same lifelines and if ‘admission’ was the inevitable and, therefore, the only outcome of ‘enrolment,’ we might hope for a perfect correlation and for errors of closure of zero. But Hamblin et al. (1998) present counts obtained from *Hospital Episode Statistics* alongside counts from the KH07 return, i.e., counts of the number of episodes of investigation or treatment alongside counts of people awaiting admission, and they omit to report counts of “removals other than admissions” (CRIR 1998, 3 of KH06). We are told that “specialists . . . may either refer with a date for admission (these patients are known as ‘booked admissions’) or . . . without a date – the true ‘waiting list’ admissions” (Hamblin et al. 1998, 13). But the distinction between “booked admissions” and “waiting list admissions” was made by *Hospital Episode Statistics* among finished consultant episodes, and the distinction between those “with a date” and those “with no date” was made by the KH07 return in its count of the number of patients awaiting admission. The KH06 return, which counted the number of “decisions to admit” to the list (and the number “admitted” and the number of “removals other than admissions” from it), made no such distinction (CRIR 1998, 3 of KH06).

We know the authors were prepared to fill the gaps in their table by calculation because they indicate that they have done so for two of the eight items. The numbers in the column headed  $\hat{E}$  (on the right-hand side of Tables 3a and 3b)

are estimates obtained using formula (4):  $\hat{E} = A + C^{now} - C^{then}$ . (The reader can check these by adding the content of columns 4 and 10 in each row. We cannot estimate the number enrolled during 1989/90 without the size of the list at the start of that financial year.) We think that the numbers tabulated as “[s]pecialist referrals . . . with no date” and “[s]pecialist referrals . . . with a date” are estimates rather than counts.

If this is correct, then the number of ‘enrolments’ presented in Tables 3a and 3b were obtained by assuming that the counts of ‘enrolments,’ ‘admissions,’ and ‘size,’ are perfectly consistent. The results therefore cannot be used to test whether this is true. At best, the table presented by Hamblin et al. (1998) provides an example which shows how the three counts ought to be related were the primary hypothesis true (Mason 1976; Fordham 1987). At worst, the table presented by Hamblin et al. (1998) invites readers to imagine that this is what actually happened to ‘enrolments’ when counts of finished consultant episodes and of patients awaiting admission varied in the manner indicated.

### In Victoria, Australia

Street and Duckett (1996) report a study of the combined elective activity of surgeons at public hospitals dealing with patients in categories 1–3 in Victoria, Australia. The authors feared that an

increase in elective procedures would increase the size of the list (Street and Duckett 1996, 4). They use counts of “additions” and “deletions” to describe activity over a single period of 12 months duration (31 July 1993 to 31 July 1994).

Street and Duckett (1996, 12) claim that “hospitals have achieved waiting list reduction in the face of increases in the number of elective surgery patients: the number of additions to the list is . . . offset by increases in the number of patients . . . deleted from the list. . .” They report that the number of category 1 patients waiting shrank from 1298 on 31 July 1993 to 195 on 31 July 1994 and that the number of category 2 patients waiting shrank from 12,115 on 31 July 1993 to 8506 on 31 July 1994 (Street and Duckett 1996), and they present an intuitively helpful plot of the number of “additions” to, and the number of “deletions” from, the surgical waiting list each month (31 December 1991 to 31 July 1994) (Street and Duckett 1996). This appears to describe the movement of people on and off the combined waiting list, although this is not clearly stated in the text.

It is true that the size of the list has diminished, despite more additions to the list (85,259, 1 Aug 1993–31 Jul 1994 incl.) than in the previous year (77,820, 1 Aug 1992–31 Jul 1993). But the published data permit only a single comparison, i.e., of the change in size between 31 Jul 1993 and 31 Jul 1994, with the difference in additions and deletions over the intervening period. It is therefore not possible to assess the strength of association between change in stock and net in-flow. The error of closure is small (335, or 0.29%, of those on the list at any point during the year).

The authors were unable to verify the number of additions and deletions we obtained from their plot (Street and Duckett 1996) 20 years after its publication but kindly volunteered the additional census counts reported in column 6 of Table 4. This allows us to describe elective activity over 32 periods each of one calendar month duration. The correlation between the change in size and the balance of enrolments and admissions was positive, strong, and statistically significant (Spearman’s  $\rho = +0.99$ ,  $n = 32$ ,  $p < 0.01$ ). But the count of dates of entry ( $C^{\text{then}} + E$ ) did not equal the count

of dates of exit ( $A + C^{\text{now}}$ ). If we ignore the grossest error, a shortfall of  $-2002$  cases ( $-6.70\%$ ) occurring in December 1991, the difference ranged from a shortfall of  $-105$  ( $-0.29\%$ ) to a surplus of  $+109$  ( $+0.32\%$ ) cases and was less than  $\pm 0.20\%$  in 28 (out of 32) instances.

If a measurement may be out by as much as  $\pm 0.5$  mm, then  $E - A$  and  $C^{\text{now}} - C^{\text{then}}$  may be out by as much as  $\pm 1.0$  mm and the error of closure by as much as  $\pm 2.0$  mm. 28 out of 32 errors cannot be attributed to this level of inaccuracy in reading the number of “additions” and “deletions” off a scale of 1 mm per 37 cases. While Street & Duckett’s data may not be entirely consistent with formula (3) and the primary hypothesis, the difference between “additions” and “deletions” accounts very well for the change in size.

## In Winnipeg, Canada

Kreindler and Bapuji (2010) report a study of the elective replacement of hips and knees in Winnipeg, Canada. Winnipeg Regional Health Authority thought that an increase in elective procedures ought to reduce the size of the list (Kreindler and Bapuji 2010). Kreindler and Bapuji (2010) use counts of “arrivals” and “departures” to describe activity over 11 periods each of 3 months duration. They emulate Street and Duckett (1996) in presenting a similarly helpful plot of the number of “arrivals” and the number of “departures” during each quarter (31 Mar 2005–31 Mar 2008) (Kreindler and Bapuji 2010) alongside a plot of the number of joints still awaiting surgery at the close of each month (31 Jan 2005–31 Jan 2008) (Kreindler and Bapuji 2010). They appreciate that they ought to count the arrival and the departure of joints if they are interested in the number of joints requiring surgery (Table 5) or count the arrival and the departure of people if they are interested in the number of people awaiting surgery.

The correlation between  $E - A$  and  $C^{\text{now}} - C^{\text{then}}$  was positive, strong, and statistically significant (Spearman’s  $\rho = +0.90$ ,  $n = 11$ ,  $p < 0.01$ ). But the number of dates of entry ( $C^{\text{then}} + E$ ) did not equal the number of dates of exit ( $A + C^{\text{now}}$ ):

**Table 4** Did the balance of “additions” and “deletions” adequately account for the change in size in Victoria, Australia?

Year	Month-end	Waiting in Victoria, Australia					Change in Stock	Counting dates of entry	Counting dates of exit	error of closure			
		No. of 'additions'	No. of 'deletions'	Size of list	Net in-Flow					difference	(%)		
		$E$	$A$	$C^{now}$	$E-A$	$C^{now}-C^{then}$				$C^{then}+E$	$A+C^{now}$	[13]	[14]
		[3]	[4]	[6]	[9]	[10]				[11]	[12]	[13]	[14]
1992	31-Dec	5,988	4,574	26,323	1,414	3,416	28,895	30,897	-2,002	-6.70			
	31-Jan	4,946	4,686	26,563	260	240	31,269	31,249	20	0.06			
	29-Feb	6,397	6,248	26,757	149	194	32,960	33,005	-45	-0.14			
	31-Mar	6,490	6,527	26,689	-37	-68	33,247	33,216	31	0.09			
	30-Apr	5,671	6,322	26,025	-651	-664	32,360	32,347	13	0.04			
	31-May	6,118	6,545	25,539	-427	-486	32,143	32,084	59	0.18			
	30-Jun	6,136	6,136	25,532	0	-7	31,675	31,668	7	0.02			
31-Jul	6,545	6,025	26,098	520	566	32,077	32,123	-46	-0.14				
1993	31-Aug	6,192	5,969	26,299	223	201	32,290	32,268	22	0.07			
	30-Sep	6,322	6,360	26,206	-38	-93	32,621	32,566	55	0.17			
	31-Oct	6,564	6,322	26,463	242	257	32,770	32,785	-15	-0.05			
	30-Nov	6,564	5,541	27,436	1,023	973	33,027	32,977	50	0.15			
	31-Dec	6,601	4,426	29,634	2,175	2,198	34,037	34,060	-23	-0.07			
	31-Jan	5,002	5,002	29,671	0	37	34,636	34,673	-37	-0.11			
	28-Feb	6,471	6,471	29,776	0	105	36,142	36,247	-105	-0.29			
	31-Mar	7,271	6,955	30,121	316	345	37,047	37,076	-29	-0.08			
	30-Apr	6,341	6,694	29,827	-353	-294	36,462	36,521	-59	-0.16			
	31-May	6,192	6,899	29,088	-707	-739	36,019	35,987	32	0.09			
	30-Jun	7,085	7,550	28,618	-465	-470	36,173	36,168	5	0.01			
31-Jul	7,215	7,122	<b>28,745</b>	93	127	35,833	35,867	-34	-0.09				
1994	31-Aug	6,917	7,847	27,740	-930	-1,005	35,662	35,587	75	0.21			
	30-Sep	7,494	7,810	27,391	-316	-349	35,234	35,201	33	0.09			
	31-Oct	6,843	7,140	27,113	-297	-278	34,234	34,253	-19	-0.06			
	30-Nov	7,178	7,736	26,549	-558	-564	34,291	34,285	6	0.02			
	31-Dec	7,029	6,360	27,164	669	615	33,578	33,524	54	0.16			
	31-Jan	5,839	6,285	26,678	-446	-486	33,003	32,963	40	0.12			
	28-Feb	7,252	7,940	25,881	-688	-797	33,930	33,821	109	0.32			
	31-Mar	7,903	7,959	25,850	-56	-31	33,784	33,809	-25	-0.07			
	30-Apr	6,583	7,308	25,093	-725	-757	32,433	32,401	32	0.10			
	31-May	7,624	7,921	24,776	-297	-317	32,717	32,697	20	0.06			
	30-Jun	7,512	8,014	24,271	-502	-505	32,288	32,285	3	0.01			
31-Jul	7,085	7,308	<b>24,041</b>	-223	-230	31,356	31,349	7	0.02				

Source: Street and Duckett 1996

the error of closure ranged from a shortfall of -62 (-1.70%) to a surplus of +62 (+1.58%) cases and was less than ±1.00% in 7 (out of 11) instances.

Kreindler and Bapuji’s data is compatible with formula (3) and the primary hypothesis – an increase in “departures” (net of “arrivals”) may accompany a reduction in the size of the list. But we used a scale of 1 mm per 9.5 cases to estimate the size of the list and a scale of 1 mm per 6.5 cases to estimate the number of “arrivals” and “departures,” so nine out of 11 errors cannot be attributed to inaccuracy in reading the relevant plot.

When entry ( $C^{then} + E$ ) and exit ( $A + C^{now}$ ) dates are used to enumerate the same lifelines

(Fig. 2), it is inconceivable that they give different counts. It is therefore reasonable to suspect the data when the counts appear inconsistent. Kreindler and Bapuji (2010, 76) recognized that their count of new “arrivals” might be considered inflated if admission was the only outcome of interest, so they calculated net “arrivals” (2005–2007) by deducting those “removed from the wait list without surgery” (2005–2007).

Kreindler and Bapuji (2010) may have deducted the number “removed” from the list during a 3 months period from the number known to have enrolled on the list in the same quarter. It is likely that some of those deducted in this fashion had enrolled earlier. If so, the net

**Table 5** Was the change in size directly correlated with the balance of “arrivals” and “departures” in Winnipeg, Canada?

Year	Month-end	Waiting in Winnipeg, Canada					Counting dates of entry $C^{then} + E$	Counting dates of exit $A + C^{now}$	error of closure	
		No. of 'arrivals'	No. of 'departures'	Size of list	Net in- Flow	Change in Stock			difference	(%)
		$E$	$A$	$C^{now}$	$E - A$	$C^{now} - C^{then}$				
		[3]	[4]	[6]	[9]	[10]			[11]	[12]
2005	31-Jan			3,076						
	28-Feb			3,171						
	31-Mar	800	600	3,200	200		3,800			
	30-Apr			3,276						
	31-May			3,271						
	30-Jun	797	710	3,338	87	138	3,997	4,048	-51	-1.27
	31-Jul			3,352						
	31-Aug			3,390						
	30-Sep	745	681	3,400	64	62	4,083	4,081	2	0.05
	31-Oct			3,424						
	30-Nov			3,414						
	31-Dec	674	739	3,371	-65	-29	4,074	4,110	-36	-0.88
2006	31-Jan			3,352						
	28-Feb			3,271						
	31-Mar	679	892	3,190	-213	-181	4,050	4,082	-32	-0.79
	30-Apr			3,133						
	31-May			3,062						
	30-Jun	769	868	3,029	-99	-161	3,959	3,897	62	1.58
	31-Jul			3,043						
	31-Aug			2,957						
	30-Sep	769	816	3,024	-47	-5	3,798	3,840	-42	-1.10
	31-Oct			2,995						
	30-Nov			2,957						
	31-Dec	677	790	2,881	-113	-143	3,701	3,671	30	0.81
2007	31-Jan			2,881						
	28-Feb			2,867						
	31-Mar	732	842	2,833	-110	-48	3,613	3,675	-62	-1.70
	30-Apr			2,771						
	31-May			2,681						
	30-Jun	716	865	2,662	-149	-171	3,549	3,527	22	0.62
	31-Jul			2,629						
	31-Aug			2,581						
	30-Sep	616	677	2,614	-61	-48	3,278	3,291	-13	-0.40
	31-Oct			2,562						
	30-Nov			2,562						
	31-Dec	685	748	2,519	-63	-95	3,299	3,267	32	0.97
2008	31-Jan			2,500						

Source: Kreindler and Bapuji 2010

“arrivals” will sometimes underestimate (and sometimes overestimate) the number which actually enrolled ( $E$ ) and proceeded to receive surgery. As a result,  $E - A$  would sometimes yield too positive, and sometimes too negative, a value. Moreover, Kreindler and Bapuji (2010) do not report deducting those “removed” from each census which followed their enrolment so the balance of net “arrivals” and “departures” could not entirely account for any change in the size of the list even if it were correct.

## In Sweden

Armstrong (2010) reports a study of cataract extraction across Sweden. He claims that “[t]he stock-flow model . . . predicts that the size of the list will increase when there is a decrease in admissions (and removals) net of enrolment, and vice versa” (Armstrong 2010, 113). Armstrong (2010) uses counts of enrolments and admissions to describe activity over 64 periods each of 3 months duration.



The change in stock correlated perfectly with net *in*-flow (Spearman's  $\rho = +1.00$ ,  $n = 64$ ,  $p < 0.01$ ) (Armstrong 2010). The number of dates of entry ( $C^{\text{then}} + E$ ) equals the number of dates of exit ( $A + C^{\text{now}}$ ), and there was no error of closure in any of the quarters studied.

It seems that the *National Cataract Register* for Sweden is entirely consistent with formula (3) and the primary hypothesis – the relationship between enrolments, admissions, and the size of the list was found to be mathematically exact.

None of the numbers presented in columns 3, 4, and 6 of Table 6 were obtained by calculation. The count of enrolments was obtained by enumerating records with a start date in the period of interest, and the count of admissions was obtained by enumerating records with an end date in the relevant period. The count of those awaiting admission was obtained by enumerating records where the start date preceded, and where the end date succeeded, the date and time of the relevant census.

It is helpful, on this occasion, that the dataset registers extractions and is compiled retrospectively. It does not contain any record where a patient was removed from the list without having received treatment, and it does not contain any record where the outcome is not yet known. So if we want to know how many cataracts were enrolled during a particular quarter, or how many – at a specified date – were still awaiting extraction, we have to allow sufficiently lengthy follow-up to ensure that each of them received treatment. (Armstrong (2010) restricted his analysis to the set of cataracts extracted less than 2 years after enrolment.) But no count has to be adjusted in the manner described by Kreindler and Bapuji (2010) to exclude those removed from the list. As a result, the records are consistent with the model.

## In England

The four studies cited here provide different compilations from the same series of counts. These counts were obtained from the *Patient Administration System* for each provider and used to complete a set of standard forms, which described

the size of the list at the close of each quarter (the KH07) and the amount of activity over its course (the KH06 and KH07A). These central returns were collated by the Department of Health and used to produce aggregate counts for England.

## Twelve Periods Each of 3 Months Duration

Newton et al. (1995) reports a study of elective inpatient activity combined across NHS hospitals in England. The authors acknowledge that “studies . . . have so far failed to show a strong inverse correlation between admission rates and list size” (Newton et al. 1995, 784). Newton et al. (1995) describe activity over 12 periods each of 3 months duration using counts of additions and admissions from the KH06 return and counts of the number still waiting from the KH07 return. They report that “changes in the number of admissions correlated inversely with changes in list size ( $r = -0.62$ ;  $P < 0.001$ ) . . . [a]fter adjusting for changes in the number of additions to lists” (Newton et al. 1995, 783). They obtain an inverse relationship because they model the effect on changes in size of admission (adjusting for enrolments) rather than the effect of enrolment (adjusting for admissions). The correlation is significant but not perfect, which means the errors of closure cannot be zero. Regrettably, the authors plotted the number of admissions and the number still waiting but not the number of additions, so we are not able to construct a suitable table for ourselves.

We think this result is due – at least in part – to a mismatch between their model and the records. The KH07 census counted some people who were subsequently removed from the list without having been admitted. Street and Duckett (1996) recognized that the size of their waiting list diminished as a result of deletion from the list, and they counted other reasons for deletion alongside treatment, but Newton et al. (1995) did not supplement their counts of admissions with the counts of other removals though these were also available from the KH06 return.

If we modify formula (3) to allow for an outcome other than admission, we obtain



**Table 6** (continued)

Year	Month-end	Waiting for cataract extraction					Change in Stock	Counting dates of entry	Counting dates of exit	error of closure			
		No. of enrolments	No. of admissions	Size of list	Net in-Flow					difference	(%)		
		$E$	$A$	$C^{now}$	$E-A$	$C^{now}-C^{then}$				$C^{then}+E$	$A+C^{now}$	[13]	[14]
		[3]	[4]	[6]	[9]	[10]				[11]	[12]		
2005	31-Mar	19,061	20,739	26,476	-1,678	-1,678	47,215	47,215	0	0.00			
	30-Jun	18,658	21,244	23,890	-2,586	-2,586	45,134	45,134	0	0.00			
	30-Sep	13,640	13,866	23,664	-226	-226	37,530	37,530	0	0.00			
	31-Dec	18,106	20,877	20,893	-2,771	-2,771	41,770	41,770	0	0.00			
2006	31-Mar	18,435	21,213	18,115	-2,778	-2,778	39,328	39,328	0	0.00			
	30-Jun	16,486	18,559	16,042	-2,073	-2,073	34,601	34,601	0	0.00			
	30-Sep	13,858	13,110	16,790	748	748	29,900	29,900	0	0.00			
	31-Dec	20,026	19,164	17,652	862	862	36,816	36,816	0	0.00			
2007	31-Mar	19,855	20,370	17,137	-515	-515	37,507	37,507	0	0.00			
	30-Jun	18,094	18,749	16,482	-655	-655	35,231	35,231	0	0.00			
	30-Sep	14,693	13,231	17,944	1,462	1,462	31,175	31,175	0	0.00			
	31-Dec	20,490	19,699	18,735	791	791	38,434	38,434	0	0.00			

Source: Armstrong 2010

$$E - (A + R) = C^{now} - C^{then}, \tag{3.1}$$

which shows the relationship between the change in size and the balance of enrolments and admissions (plus other removals).

If we add  $A + R$  to both sides, we obtain

$$E = A + R + C^{now} - C^{then}. \tag{4.1}$$

Formula (4.1) has been used to estimate the number of enrolments when the relevant count is not available (Naylor et al. 1997).

If we add  $C^{then}$  to both sides of formula (4.1), we obtain

$$C^{then} + E = (A + R) + C^{now}, \tag{1.1}$$

which allows us to compare the dates of entry and the dates of exit of those on the list at any point between the two censuses (Armstrong 2000).

Nevertheless, the data used by Newton et al. (1995) is compatible with formula (3) and the primary hypothesis – an increase in “admissions” (net of ‘additions’) may accompany a reduction in the size of the list.

**One Period of 3 months Duration**

The National Audit Office (NAO 2001a) reports a study of all elective inpatient and day-case activity combined across the NHS in England. It uses

counts of decisions to admit and of the number admitted or removed to describe activity over one period of 3 months duration. (It is therefore not possible to assess the strength of association between change in stock and net in-flow.)

The NAO (2001a, 21) “was unable to reconcile aggregated changes in [the size of] the waiting list.” It found 24,312<sup>†</sup> more patients still on the list at the close of the quarter than were accounted for by additions and “admissions” plus “removals” (Table 3c). “The Department of Health explain the discrepancy by acknowledging that they do not measure every flow onto and off of the waiting list, but focus on the major ones such as hospital admissions and suspensions” (NAO 2001a, 21). It is noteworthy that the patients removed from the list are a substantial flow but are not mentioned, and the patients suspended are mentioned but are neither substantial, accounting for a reduction in size of another 74 cases<sup>†</sup>, nor a flow – as recorded in the available returns.

$E - (A + R)$  must exactly equal  $C^{now} - C^{then}$ , if  $E - (A + R)$  accounts for all of those who joined the list or who left it in the interval between  $C^{then}$  and  $C^{now}$ ; if enrolments, admissions, removals, and size enumerate the same lifelines (whether these are episodes of investigation or treatment, the conditions which prompted those, or the patient in possession of one or more of these); and if all four counts are accurate. This is why the National Audit Office (2001a) was not happy with

**Table 3c** Did the balance of decisions to admit and of “admissions” and “removals” adequately account for the change in size in England?

Year	Month-end	Waiting for Admission (in-patient or day case)						error of closure					
		No. of 'decisions-to-admit'	No. of elective 'admissions'	No. of other 'removals'	Size of waiting list	Net in-Flow	Change in Stock	Counting dates of entry	Counting dates of exit	difference	(%)		
		<i>E</i>	<i>A</i>	<i>R</i>	<i>C<sup>now</sup></i>	<i>E-(A+R)</i>	<i>C<sup>now</sup>-C<sup>then</sup></i>	<i>C<sup>then</sup>+E</i>	<i>A+R+C<sup>now</sup></i>				
		[3]	[4]	[5]	[6]	[9]	[10]	[11]	[12]	[13]	[14]		
2000	31-Dec				1,034,381								
2001	31-Mar	992,918	872,188	172,696	1,006,727	-51,966	-27,654	2,027,299	2,051,611	-24,312	-1.19		

Source: NAO 2001a

**Table 3d** Did the balance of decisions to admit and of admissions and removals adequately account for the change in size in England?

Year-end	Waiting for Admission (in-patient or day case)						error of closure						
	No. of "decisions to admit"	No. of elective "admissions"	No. of other "removals"	Size of waiting list	Net in-Flow	Change in Stock	Counting dates of entry	Counting dates of exit	difference	(%)			
	<i>E</i>	<i>A</i>	<i>R</i>	<i>C<sup>now</sup></i>	<i>E-(A+R)</i>	<i>C<sup>now</sup>-C<sup>then</sup></i>	<i>C<sup>then</sup>+E</i>	<i>A+R+C<sup>now</sup></i>					
	[3]	[4]	[5]	[6]	[9]	[10]	[11]	[12]	[13]	[14]			
31-Mar-89	2,783,298	2,632,085	200,677	922,676	-49,464			3,755,438					
31-Mar-90	2,943,658	2,768,482	260,503	958,976	-85,327	36,300	3,866,334	3,987,961	-121,627	-3.10			
31-Mar-91	2,964,836	2,761,005	306,899	948,243	-103,068	-10,733	3,923,812	4,016,147	-92,335	-2.33			
31-Mar-92	3,257,615	2,993,532	387,980	917,717	-123,897	-30,526	4,205,858	4,299,229	-93,371	-2.20			
31-Mar-93	3,480,268	3,111,627	412,299	994,974	-43,658	77,257	4,397,985	4,518,900	-120,915	-2.71			
31-Mar-94	3,501,715	3,110,477	451,559	1,065,369	-60,321	70,395	4,496,689	4,627,405	-130,716	-2.87			
31-Mar-95	3,765,407	3,376,016	521,320	1,044,051	-131,929	-21,318	4,830,776	4,941,387	-110,611	-2.26			
31-Mar-96	3,968,825	3,500,353	547,863	1,048,029	-79,391	3,978	5,012,876	5,096,245	-83,369	-1.65			
31-Mar-97	4,111,511	3,549,074	551,999	1,158,004	10,438	109,975	5,159,540	5,259,077	-99,537	-1.91			
31-Mar-98	4,192,037	3,543,634	558,242	1,297,662	90,161	139,658	5,350,041	5,399,538	-49,497	-0.92			
31-Mar-99	4,189,323	3,826,507	672,432	1,072,860	-309,616	-224,802	5,486,985	5,571,799	-84,814	-1.53			
31-Mar-00	4,159,078	3,682,180	622,787	1,037,066	-145,889	-35,794	5,231,938	5,342,033	-110,095	-2.08			
31-Mar-01	3,935,930	3,467,338	613,931	1,006,727	-145,339	-30,339	4,972,996	5,087,996	-115,000	-2.29			
31-Mar-02	3,781,437	3,244,185	581,534	1,035,365	-44,282	28,638	4,788,164	4,861,084	-72,920	-1.51			
31-Mar-03	3,778,390	3,330,981	601,353	992,075	-153,944	-43,290	4,813,755	4,924,409	-110,654	-2.27			
31-Mar-04	3,802,744	3,391,644	621,345	905,753	-210,245	-86,322	4,794,819	4,918,742	-123,923	-2.55			
31-Mar-05	3,787,713	3,390,694	612,004	821,722	-214,985	-84,031	4,693,466	4,824,420	-130,954	-2.75			
31-Mar-06	4,031,519	3,577,104	613,626	784,572	-159,211	-37,150	4,853,241	4,975,302	-122,061	-2.48			
31-Mar-07	4,154,486	3,746,666	613,886	700,624	-206,066	-83,948	4,939,058	5,061,176	-122,118	-2.44			
31-Mar-08	4,355,950	4,043,307	646,394	531,520	-333,751	-169,104	5,056,574	5,221,221	-164,647	-3.20			
31-Mar-09	4,979,682	4,418,090	647,550	565,954	-85,958	34,434	5,511,202	5,631,594	-120,392	-2.16			

Source: House of Commons Health Committee 2010

any discrepancy between the two figures and why the Department of Health concurred (CRIR 1997).

The NAO’s data is compatible with formula (3) and the primary hypothesis – an increase in ‘admission’ (plus “removal” net of enrolment) may accompany a reduction in the size of the list.

**Twenty Periods each of 12-months Duration**

The House of Commons Health Committee (2010) published an extended series of counts obtained from the Department of Health. It uses

counts of decisions to admit and of the number admitted or removed from the list to describe activity over 20 periods each of 12-months duration (Table 3d). These counts were obtained from the same returns used by the NAO (2001a).

The correlation between  $E - (A + R)$  and  $C^{now} - C^{then}$  was positive, strong, and statistically significant (Spearman’s  $\rho = +0.97$ ,  $n = 20$ ,  $p < 0.01$ ). But the number of dates of entry ( $C^{then} + E$ ) did not equal the number of dates of exit ( $A + R + C^{now}$ ): the discrepancy ranges from -164,647 (-3.20%) to

**Table 7** Does the balance of enrolments and admissions (plus other removals) correctly predict the direction of any change in the size of the list?

a) Street & Duckett, 1996				b) Health Committee, 2009				c) Armstrong, 2000						
		$E - A$				$E - A$				$E - A$				
		+	-			+	-			+	-			
$C^{now} - C^{then}$	+	11	2	13	$C^{now} - C^{then}$	+	2	6	8	$C^{now} - C^{then}$	+	2	5	7
	-	0	19	19		-	0	12	12		-	0	2	2
$L_B = 84.62$ (95% C.I. = 45–100)				$L_B = 25.00$				$L_B = 0.00$						

–49,497 (–0.92%) patients. The counts systematically overestimate the number of exits from the English waiting list (or systematically underestimate the number of entries on it).

Had we predicted that the size of the list would shrink, we would have been mistaken only eight times out of 20 (Table 7). Had we used net in-flow to predict the direction of change in stock, we would have predicted an increase on two occasions and a decrease on 18, i.e., we would have been mistaken on six out of 20 occasions. This reduction in the error of prediction of 25% ( $L_B = 0.25$ ) is not significant. So the direction of any change in size appears to have had little to do with the efforts made during the course of the year. Results such as this might go some way to explaining the frustration of at least one former Minister of Health (Powell 1966).

The six exceptions in this data might be thought consistent with hypotheses of self-regulation and of supplier-induced demand – the size of the list showed an increase when it ought to have shown a decrease. But it should be noted that the exceptions in the data provided by Street and Duckett (1996) occur only when  $E = A$ , i.e., when  $E - A = 0$ , and that there are no exceptions in the data presented by other researchers (White 1980; Moral and de Pancorbo 2001; Kreindler and Bapuji 2010; Armstrong 2010), i.e., the direction of net in-flow ( $E - A$ ) perfectly predicts the direction of any change in size ( $C^{now} - C^{then}$ ). More importantly, exceptions (Street and Duckett 1996; House of Commons Health Committee 2010) are observed only because the number of dates of entry does not equal the number of dates of exit in the KH06 and KH07 returns.

The Health Committee’s data is compatible with formula (3) and the primary hypothesis –

an increase in admission (plus removal) net of enrolment may accompany a reduction in the size of the list.

**Nine Periods each of 6-months Duration**

Armstrong (2000) reports a study of elective inpatient and day-case activity combined across NHS hospitals in England. He describes nine periods each of 6-months duration using counts of decisions to admit and of the number “admitted” or “removed”, who “self-deferred”, “failed to attend”, or were “suspended”. These counts were obtained from the same returns used by the NAO (2001a) and by the House of Commons Health Committee (2010).

In Table 3e, the change in size is always more positive than the net in-flow by between 68,237 and 32,115 patients, so the error of closure ranged between –2.27% and –1.15%. Armstrong asserts that “[t]he number of patients waiting at the start of a calendar period of interest or who counted as new ‘decisions-to-admit’ or as those ‘reset-to-zero’ or ‘reinstated’ during it, must be reconciled with the numbers admitted, removed, self-deferred, failed, medically deferred or suspended during the calendar period of interest or still awaiting admission at its close” (Armstrong 2000, 2043). But he was unable to account for this discrepancy by allowing for other flows “onto and off of the waiting list” for which there were data, i.e., those who were suspended from the list, those who canceled arrangements for their own admission or who simply did not attend, those who were reinstated to the list, and those whose start date was reset to zero (Armstrong 2000, 2043–2045).

The correlation between  $E - (A + R)$  and  $C^{now} - C^{then}$  was positive, strong, and statistically

**Table 3e** Did the balance of “decisions to admit” and of “admissions” and “removals” adequately account for the change in size in England?

Year	Censused		'Decisions-Reset-to-'		'Admitted'		'Removed'		'Self-deferred'		'Failed'		'Medically suspended'		'Deferred'		'Censused'		Net in-Flow	Change in Stock	Counting dates of entry	Counting dates of exit	error of closure difference (%)	
	$C^{then}$	$E$	to-admit'	zero'	$A$	$R$	$A$	$R$	$A$	$R$	$A$	$R$	$A$	$R$	$A$	$R$	$A$	$R$					$E-(A+R)$	$C^{now}-C^{then}$
1988	878,306	1,389,133	298,687	*	†	‡	§	¶											7,615	53,189	2,566,126	2,611,700	-45,574	-1.8
1989	922,877	1,446,243	307,945	-	1,286,087	95,431	95,508	203,179	-	-	-	-	-	-	-	-	-	-	647	48,968	2,677,065	2,725,386	-48,321	-1.8
1990	955,786	1,485,021	210,352	-	1,323,492	122,104	99,189	208,756	-	-	-	-	-	-	-	-	-	-	-43,111	9,734	2,651,159	2,704,004	-52,845	-2.0
1991	964,050	1,614,328	190,474	-	1,373,394	154,738	101,028	109,324	-	-	-	-	-	-	-	-	-	-	-46,067	-13,952	2,768,852	2,800,967	-32,115	-1.2
1992	937,054	1,748,716	204,380	-	1,463,869	196,526	93,065	97,409	-	-	-	-	-	-	-	-	-	-	-6,879	40,135	2,890,150	2,937,164	-47,014	-1.6
1993	1,019,341	1,731,690	225,203	-	1,553,237	202,358	111,373	93,007	-	-	-	-	-	-	-	-	-	-	-21,793	46,444	2,976,234	3,044,471	-68,237	-2.3
1994	1,077,497	1,861,754	257,577	50,008	1,665,747	251,393	160,343	97,234	-	50,008	-	-	-	-	-	-	-	-	-55,386	-7,005	3,246,836	3,295,217	-48,381	-1.5
1995	1,052,958	1,972,067	288,143	92,966	1,739,917	273,491	182,723	105,420	-	92,966	-	-	-	-	-	-	-	-	-41,341	1,990	3,406,134	3,449,465	-43,331	-1.3
1996	1,056,122	2,067,520	306,572	123,383	1,799,013	273,861	193,345	113,227	-	123,383	-	-	-	-	-	-	-	-	-5,354	48,862	3,553,597	3,607,813	-54,216	-1.5
1997	1,207,515	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	†	54,400					

Note: The numbers in italics contribute nothing to the difference between those becoming eligible for admission, and those becoming ineligible, so the error of closure is really a comparison of  $C^{then} + E$  and  $A + R + C^{now}$

Adapted from Armstrong (2000)

<sup>a</sup>Estimated as the number who self-deferred or failed-to-attend for admission to hospital that quarter

<sup>b</sup>Estimated as the number temporarily suspended or deferred on medical grounds that quarter

<sup>c</sup>The quarterly counts were not collected in 1997/98

significant (Spearman's  $\rho = +0.95$ ,  $n = 9$ ,  $p < 0.01$ ). The data (Armstrong 2000) is compatible with formula (3.1) and the primary hypothesis – an increase in admission plus other removals (net of enrolment plus other additions) may accompany a reduction in the size of the list.

Armstrong (2000) concludes that “[a]lthough the NHS is obliged to produce complete and accurate reports of how it used public monies, the same standard has yet to be applied to accounts of what became of patients enrolled on the waiting list for England” (Armstrong 2000, 2045). If we combine the numbers who self-deferred, failed to attend, or were suspended from the list, we find that they accounted for between 10.3% and 17.8% of flows off the list. The Department of Health acknowledges that “they do not measure every flow onto and off of the waiting list, but focus on the major ones” (NAO 2001a, 21).

No counts were collected of the numbers ‘reset to zero’ having previously deferred admission or having failed to attend on the date in question, and no counts were collected of the numbers ‘reinstated’ to the list having previously been suspended from it, i.e., the data model was more complex than the central returns allowed. So we do not know the size of the error of closure or whether it is systematic; and we do not know whether  $E - (A + R)$  predicts  $C^{\text{now}} - C^{\text{then}}$  or not.

### A Problem of Our own Making

Under the primary hypothesis, we expect the size of the list to change from one census to another by any difference in enrolments and admissions during the interval. But researchers do not appear to have had sufficient confidence in the hypothesis to subject it to rigorous testing.

- Many researchers omit enrolments (Feldstein 1967; Culyer and Cullis 1976; Snaith 1979; Frost 1980; Buttery and Snaith 1980; Fowkes et al. 1983; Goldacre et al. 1987; Niinimäki 1991; Harvey et al. 1993; Nordberg et al. 1994; Kenis 2006). So White (1980, 273–274) examines the effect of admissions, i.e., of “discharges and deaths,” on the size of the inpatient waiting list without considering the effect of enrolments. This is more than usually

incongruous because White expects changes in the number of enrolments, i.e., “GP referrals” (White 1980, 271), to affect the size of the outpatient waiting list as well as changes in the number of admissions, e.g., “new outpatients booked” (White 1980, 271–272).

- Other researchers mismatch the timing of the counts. Street and Duckett (1996) present the number of additions and deletions each month for use with annual censuses of the list, and Kreindler and Bapuji (2010) present the number of arrivals and departures each quarter for use with censuses taken 1 month apart.
- A few researchers draw conclusions so reluctant as to falsify what the data otherwise verifies. So Hamblin et al. (1998) present counts which seem to confirm the existence of a perfect mathematical relationship between the changes in size and the balance of enrolments and admissions (but for the error discussed above in connection with Tables 3a and 3b). Yet rather than evaluating the primary hypothesis, in which the variation in enrolment confounds the effect of variation in admission on the size of the list and for which they appear to have data, they advocate “the acceptable wait hypothesis” in which the variation in enrolments duplicates variation in admissions (Hamblin et al. 1998, 37, 42, 59, 64) in order to maintain the length of wait for which they do not have data.

Instead, rigorous testing has been left to auditors untroubled by secondary hypotheses of, for example, supplier-induced demand. So the National Audit Office for England (2001a) expects to “reconcile” the two counts of lifelines,  $C^{\text{then}} + E$  and  $(A + R) + C^{\text{now}}$ , because it appreciates that – as in double-entry bookkeeping – the relationship ought to be exact.

It is difficult to obtain consistent counts of stock and flow if the population (or waiting list) is narrowly defined. These difficulties are exaggerated if members move from one population to another and if the methods of data capture are felt to be unduly onerous. So the error of closure allows demographers to assess whether the registration of vital events (births and deaths) and of migration (in and out) has yielded counts

consistent with periodic censuses of population. It ought to be relatively easy to obtain counts of enrolments, admissions, and removals which are consistent with periodic censuses of the list. We can cross-examine the paper, or digital, records rather than the individuals they represent: the records are retrieved and dismissed at the researchers' convenience, and their details are always available for inspection and analysis. It ought therefore to have been possible to report an error of closure of 0% under the *Körner Reporting System* (CRIR 1997) which counted relevant records from a hospital's *Patient Administration System*. One part of the error observed was due to the use of inconsistent definitions (Newton et al. 1995), another was due to incomplete flows (NAO 2001a), and still another was, we think, the result of allowing the data model to become too elaborate (Armstrong 2000). If our systems do not allow us to identify who was eligible for admission in the period between two censuses, and if they do not allow us to demonstrate that there are as many dates of exit for this set of records as there are dates of entry, the apparent complexity of the waiting list is a problem of our own making.

---

### **The Balance of Enrolments and Admissions (Plus Other Removals) Equals the Change in Size. Why?**

We attribute our success (Armstrong 2010) in demonstrating this relationship to two things.

#### **If the Model Is Not Complicated, the Data Must Be Simple!**

The first is our assumption that each wait started and ended on the start and end dates of the record. This implies (a) that the dataset is complete, i.e., that no record was omitted, and (b) that both dates were entirely accurate. It also implies (c) that everyone, having once enrolled, was eventually admitted and (d) that no wait was ever broken. Items (c) and (d) are implied by the data definitions and tables of Working Group A (DHSS

1981b). In other words, we assumed that the waiting list had all of the attributes implied by our use of these two variables. (We did not modify the size of the list, deducting any patient who was suspended or deferred at that point; and we did not modify the length of wait, deducting any period when a patient was considered to be unfit or thought to be unavailable (Armstrong 2010)).

Like ourselves, other researchers are obliged to assume that the data are complete (or are at least representative) and that the data are accurate (or are at least not distorted) if they wish to proceed with their enquiries. Our success seems to suggest that the difficulties experienced by others (Armstrong 2000; NAO 2001a) may be due to a mismatch between the model and the data. The dataset is simple (IMG 1992); the data model is elaborate.

So when the Steering Group on Health Services Information (1984) proposed what became the KH06 and KH07 returns, they envisaged that patients would join the list as the result of a 'decision to admit' authorized by a clinician (Steering Group 1984, 85; IMG 1992, 5/3 & 5/8) and that patients would leave the list either as the result of "hav[ing] been admitted" (Steering Group 1984, 85) or as the result of "no longer needing to be admitted" (Steering Group 1984, 86). The only complication which seems to have been envisaged relates to those patients whose arrangements for admission miscarry. These fall into four categories: (a) those who did not attend, i.e., who neither declined the arrangement in advance nor presented themselves on the day, (b) those who deferred admission by contacting the hospital in advance, (c) those whose admission was canceled by the hospital, and (d) those who were admitted but subsequently discharged without having undergone investigation or treatment.

The Working Group recommended that information be collected about the "[n]umber of patients for whom arrangements to admit were made but [who] were not admitted" (DHSS 1981b, 125), i.e., it did not distinguish between the first, second, and third categories. The Steering Group recommended that information be collected about the "[n]umber of patients ...



who were not admitted because they failed to attend” (Steering Group 1984, 87), i.e., it did not distinguish between the first and second categories. But the Steering Group also recommended that information be collected about the “[n]umber of patients for whom . . . admission did not take place because of cancellation by the hospital” (Steering Group 1984, 87), i.e., about the third category. But the KH07A return, developed in the 6 months prior to implementation of the system (DHSS 1986), asked for counts of the number of patients who deferred their own admission (the second category) rather than counts of the number whose admission was canceled by the hospital.

The earliest version of the KH06 reported four “events occurring during the quarter” (DHSS 1986, 4) namely, the “decisions to admit” which marked addition to the list and three mutually exclusive outcomes which marked subtraction from it. It was anticipated that a patient might be admitted from the waiting list to undergo investigation or treatment on an elective basis prior to discharge, that a patient might not be admitted although arrangements for this had been made, or else that a patient might be removed from the waiting list as no longer requiring the elective admission intended.

The three outcomes were subsequently defined by the *Data Manual* (version 1.0) so as to subsume other possibilities:

1. Patients who were admitted as emergencies were not to be counted as having been admitted from the waiting list as arranged (IMG 1992). Rather, they were to be counted as having been removed from the waiting list as no longer requiring elective admission (CRIR 1997).
2. Patients, who were admitted from the waiting list but were then discharged from hospital without undergoing the investigation or treatment planned, were not to be counted as having been admitted from the waiting list as arranged (IMG 1992).
3. Patients who were not admitted from the waiting list because the arrangement had been canceled by the hospital were not to be counted as “not admitted” (IMG 1992, para. 41; CRIR 1997).
4. Patients who were not admitted from the waiting list because they declined an offer or canceled an arrangement were also not to be counted as “not admitted” (IMG 1992, para. 41; CRIR 1997).

The instructions assert that “patients should only be taken off the elective admission list when they have been treated – unless the treatment is no longer required” (IMG 1992, para. 41), as though this had always been self-evident. But the examples given seem to suggest that practice was in need of correction. “Patients should not be removed from the waiting list, because of self-deferrals or deferral by the hospital. For example, a patient admitted but sent home because treatment has been deferred . . . should not be removed from the elective admission list” (IMG 1992, para. 41). Those who “failed to arrive” (IMG 1992, para. 48) are carefully distinguished from “self-deferred admissions . . . or admissions cancelled by the hospital” (IMG 1992, para. 87). They have neither been admitted from the waiting list as arranged nor removed from the waiting list as no longer requiring elective admission. So they appear to constitute a third class of event in the earliest version of the return in addition to the two expressly authorized.

The waiting list envisaged by the Steering Group on Health Services Information (1984) appears to have been one in which the arrangement of admission fulfilled the hospital’s entire responsibility to the patient. Such a view seems scarcely credible and therefore needs to be substantiated:

- Some of the instructions in the *Data Manual* (version 1.0) seem to confirm such an attitude toward the patient. So if a patient “failed to arrive” without giving notice of her intentions, her details are to be returned to the GPFH who will determine whether she requires a fresh referral, another consultation, and a new decision to admit (IMG 1992, para. 71; CRIR 1997). But the patient who declines an offer or cancels an arrangement in good time receives a degree of consideration. She is counted as waiting “with [a] date” until the intended

admission has passed, and she is then given a start date the same as that on which she ought to have left the list (IMG 1992). In other words, the hospital authorizes the patient's return to the list without forwarding her details to the GPFH, waiting for another letter of referral, and organizing a fresh consultation in due course. The consideration extended to the exception – the patient who self-deferred admission – seems to confirm the rule about the patient who gave no warning but failed to attend.

- This attitude also seems to be confirmed by instructions in the *Data Manual* (version 4.0) about patients discharged without having been investigated or treated. “Patients are taken off the elective admission list once they are admitted into hospital. If treatment is then deferred because of lack of facilities or for medical reasons ... the patient is discharged ... A new decision to admit and a new elective admission list entry will then be made for the patient” (CRIR 1997, 16). So the wait is considered to be completed upon admission regardless of what happens next (CRIR 1998), and the patient who has not received the elective investigation or treatment promised will need “[a] new decision to admit and a new [entry on the] elective admission list” if she wishes to try again. The size of such a list shrinks not only as a result of admissions which are followed by investigation and treatment but also as a result of admissions which are not.
- The Working Group recommended that “waiting lists [be] regularly reviewed to remove patients no longer needing or wishing to be admitted” (DHSS 1981b, 127), acknowledging that some would never be admitted from the waiting list. But it did not recommend counting the “[n]umber of patients ... removed from a list for reasons other than elective admission” (Steering Group 1984, 87). This suggests that the Working Group felt no responsibility toward those removed. One of the members of the group expressed an appropriate concern that the number of those still waiting should not be exaggerated by including anyone no longer eligible for elective

admission. He noted that they “no longer need ... or wish ... to be admitted” at the time of the review. But he thought their eventual removal from the list implied that they were never really available for admission. He infers that they were not eligible at the time of any census in which they appeared and that the decision to admit ought never to have been authorized (Lee et al. 1987). He recommends deducting their contribution to counts of decisions to admit and of the numbers still waiting. We think this view seriously flawed. He rejects the possibility that these patients could have received investigation or treatment had it been made available more promptly.

There are grounds therefore for imagining that the balance of decisions to admit less the three outcomes (KH06) ought to have accounted for differences between the number waiting (KH07) at the close of this quarter and the number waiting at the close of the last quarter in the earliest days of the *Körner Reporting System*. If this were the case, the simplest model would require the insertion of an additional variable in formula (3.1) so that  $E - (A + N + R) = C^{\text{now}} - C^{\text{then}}$ , where  $N$  represents the number “not admitted” during the interval between  $C^{\text{now}}$  and  $C^{\text{then}}$ .

Table 3f allows us to assess the consistency of these counts. The correlation between  $E - (A + N + R)$  and  $C^{\text{now}} - C^{\text{then}}$  was strong, but it was not statistically significant (Spearman's  $\rho = -0.96$ ,  $n = 4$ ,  $p = 0.20$ ), and it did not have the direction desired: the net in-flow indicates that the size of the list was getting smaller, while the change in stock indicates that the size of the list was getting bigger. (The counts of stock (KH07) and flow (KH06) do not appear to describe the same waiting list.) There was a substantial error of closure ranging from  $-10.40\%$  to  $-4.90\%$  of those eligible for admission at any point over the relevant 6 months.

The discrepancy in Table 3f might be explained in a number of ways. Apart from simple underreporting of the number of patients added to the list or overreporting of the numbers admitted from the list or removed, this might occur where individuals are reported as contributing more than

**Table 3f** Did the balance of “decisions to admit” and of those “admitted,” “not admitted,” or “removed” adequately account for the change in size in England?

Year	Waiting for Admission (in-patient or day case)								Net in-Flow	Change in Stock	Counting			
	No. of 'decisions- to-admit'				Size of list		E-(A+N+R)	C <sup>now</sup> -C <sup>then</sup>			Counting dates of entry	Counting dates of exit	error of closure	
	No. who were 'admitted'	No. who were 'admitted'	No. who were 'not admitted'	No. who were 'removed'	31-Dec	30-Jun							difference	(%)
	E	A	N	R	C <sup>now</sup>	C <sup>then</sup>							C <sup>then</sup> +E	A+N+R+C <sup>now</sup>
[3]	[4]	[5]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]				
1988	1,389,133	1,286,087	203,179	95,431	931,495	878,306	-195,564	53,189	2,267,439	2,516,192	-248,753	-10.40		
1989	1,446,243	1,323,492	208,756	122,104	971,845	922,877	-208,109	48,968	2,369,120	2,626,197	-257,077	-10.29		
1990	1,485,021	1,373,394	109,324	154,738	965,520	955,786	-152,435	9,734	2,440,807	2,602,976	-162,169	-6.43		
1991	1,614,328	1,463,869	97,409	196,526	950,098	964,050	-143,476	-13,952	2,578,378	2,707,902	-129,524	-4.90		

one outcome but no more than one decision to admit. For example, where a patient is transferred from a list at another hospital and is duly admitted or removed without a local decision to admit having been made (IMG 1992). Or where a patient is removed from the list as not medically fit for elective admission (CRIR 1997) and is subsequently reinstated without a fresh decision to admit having been made (IMG 1992).

Other possibilities are more complicated and appear to be capable of accounting only for a part of the problem. So if a patient is temporarily suspended from the list on medical grounds at the close of a quarter, he will either be omitted from the decisions to admit over that quarter or else be omitted from those still waiting at its close. In the first instance, there will appear to have been fewer dates of entry (column 11, Table 3f) to the period of interest and the reported difference (column 13) in counts of dates of entry and dates of exit and the error of closure (column 14) – being negative – will appear larger. In the second, there will appear to have been fewer dates of exit (column 12) from the period of interest and the reported difference (column 13) in counts of dates of entry and dates of exit and the error of closure (column 14) – being negative – will appear smaller.

The *Data Manual* presents a complicated series of rules about what parts of which records contribute data on the official wait for elective admission. But version 1.0 also asserts that “patients should only be taken off the elective admission list when they have been treated – unless the treatment is no longer required” (IMG 1992, para. 41). Version 4.0 claims that “[t]he . . . KH06 . . . relate[s] to elective admission list events – all the additions to the waiting list (i.e.,

the number of decisions to admit) and removals from the waiting list that have taken place during the quarter” and also asserts that “[t]he change in the total numbers waiting should reflect this activity” (CRIR 1997, para. 144). Despite the fact that “failed to attend” is classed as an event on the KH06 return (CRIR 1997, para. 148), the simplest explanation for the discrepancy within Table 3f is that there are two outcomes which end enrolment not three. We obtain a better account of the stock and flow of the English waiting list if we omit the “failed to attend” (Table 3g).

Table 3g shows the consistency of the counts if the relationship is, in practice, best described by formula (3.1). The correlation between  $E - (A + R)$  and  $C^{now} - C^{then}$  was perfect and had the direction desired, but it was not statistically significant (Spearman’s  $\rho = +1.00$ ,  $n = 4$ ,  $p = 0.20$ ). There was a small error of closure ranging from  $-2.14\%$  to  $-1.24\%$  of those eligible for admission at any point over the relevant 6 months.

This is a little disconcerting. The data model used in practice appears to be simpler (CRIR 1997) than the *Data Manual* would have us believe.

Within a short time of implementation, the Government Statistical Service began to modify the KH06, KH07, and KH07A returns. Now we sympathize with the performance analyst who wishes to restrict attention to that part of the list, and that portion of the wait, for which a manager (or a clinician) might reasonably be held responsible. But we think the returns were changed without considering the effect on the consistency of the counts.

Neither the DHSS (1981b), nor the Steering Group (1984), nor the authors of the first set of

**Table 3g** Did the balance of “decisions to admit” and of those “admitted” or “removed” adequately account for the change in size in England?

Year	Waiting for Admission (in-patient or day case)								Counting			
	No. of			Size of list		Net	Change	Counting	Counting	error of closure		
	'decisions-	No. who	No. who	31-Dec	30-Jun					in-Flow	in Stock	dates of entry
	to-admit'	'admitted'	'removed'	$C^{now}$	$C^{then}$	$E-(A+R)$	$C^{now}-C^{then}$	$C^{then}+E$	$A+R+C^{now}$	[13]	[14]	
[3]	[4]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]		
1988	1,389,133	1,286,087	95,431	931,495	878,306	7,615	53,189	2,267,439	2,313,013	-45,574	-1.99	
1989	1,446,243	1,323,492	122,104	971,845	922,877	647	48,968	2,369,120	2,417,441	-48,321	-2.02	
1990	1,485,021	1,373,394	154,738	965,520	955,786	-43,111	9,734	2,440,807	2,493,652	-52,845	-2.14	
1991	1,614,328	1,463,869	196,526	950,098	964,050	-46,067	-13,952	2,578,378	2,610,493	-32,115	-1.24	

returns (DHSS 1987) mention the possibility of suspension from the waiting list either on medical grounds or for social reasons. But version 1.0 of the *Data Manual* instructed the NHS to suspend from the list those “patients who are not medically ready for admission” (IMG 1992, 16), and version 4.0 of the *Data Manual* advised the NHS that this was consistent with the practice of not adding patients to the list until they are “likely to be fit for surgery when offered” (CRIR 1997, 17). Version 4.0 also advised the NHS that “[p]atients may also be suspended from [a] . . . list for social reasons such as holidays or family commitments which may be notified in advance” (CRIR 1997, 17).

The IMG (1992, 9 & 18) asserted that “[p]atients who are currently not medically ready should **not** be included in the national returns” and emphasized that “patients . . . who are not medically ready for admission are excluded from all waiting list central returns.” Now counts of enrolments, admissions, and size ought to be consistent if each of them exclude all of those removed from the list (Lee et al. 1987; Kreindler and Bapuji 2010). In the same way, counts of enrolments, admissions (plus removals), and size ought to be consistent if each of them exclude all of those ever temporarily suspended from the list.

But these are patients whose admission to the list was authorized because they were thought “likely to be fit for surgery when offered.” It is likely therefore that counts of decisions to admit enumerated some who were subsequently excluded from a census, so the size of the list is too small for the number enrolled. Moreover, (most of) those excluded from the census because they were not medically ready will be reinstated to the

list and will subsequently contribute to the relevant count of admissions or removals, so the size of the list is also too small for the number admitted or removed. The publication of well-worded definitions may have improved the consistency of meaning attached to the various items, and the suspension of some (IMG 1992) who were not medically ready may have improved the homogeneity of the group requiring investigation or treatment. But omitting those reinstated during the quarter, and those suspended at its close, did not improve the consistency of counts of enrolments and admissions (plus removals), with size.

Insistence on a model with more carefully specified outputs ought to have prompted the development of a dataset with more carefully defined classes and counts. The National Audit Office (2001a, 21) reports the Department of Health as “acknowledging that they do not measure every flow onto and off of the waiting list.” But insistence on a model which introduces a break anywhere between the beginning and end of the patient’s time on the list demands another level of complexity from the dataset.

- In some instances, the wait continues to accrue. The patient who is suspended from the list on medical grounds becomes invisible to enumeration in the census, but there is no outcome or end date before the census to account for the disappearance, and there is no start date or reinstatement after the census to account for the reappearance (IMG 1992). The effect on the *Körner Reporting System* is to make the counts of stock and flow less consistent. (The *Data Manual* (version 4.0) acknowledges the problem. The number of patients suspended

from the list – on social grounds – is to be added back to the number still waiting before assessing whether the counts are consistent (CRIR 1997).

- In other instances, the accumulated wait is discounted. The patient who declines an offer or cancels an arrangement accrues time on the list until the date offered or arranged. This then becomes the effective date of the patient's addition to the list and the wait accumulated to date is reset to zero. But no outcome marks the end of the first wait, and no decision to admit marks the beginning of the second, so there is no record of flows which can account for the changes within the relevant waiting time categories.

A more elaborate definition of the wait for investigation or treatment requires a more complicated dataset, with additional variables to provide a start date and an end date for the latest of those occasions on which the patient is classed as “not being medically ready” (IMG 1992, 9). A still more elaborate definition requires a still more complicated dataset, with variables to provide start dates and end dates for each occasion on which the patient is suspended (CRIR 1997) and for each occasion (first, second, etc.) when a patient deferred admission. But what has not been recognized is that the occurrence of a break between enrolment and admission (or removal) has to be accounted for by flows other than enrolment and admission (or removal). The definitions adopted under the *Körner Reporting System* soon became so complicated that there were not variables enough to represent all of those thought to be eligible, or ineligible, for admission over a period of interest (Armstrong 2000; NAO 2001a). Some patients who had been temporarily suspended as “not medically ready” (IMG 1992, 9) were subsequently removed from the list without having first been reinstated (CRIR 1997).

Data definitions have sometimes become so elaborate that it has proven impossible to reconstruct the state of the records as they stood on a particular date, even with the most up-to-date versions of the relevant software (Farquharson 2011). We think this reprehensible. The result is a list in which two counts – ostensibly of the same

thing – give different answers (Armstrong 2000; NAO 2001a) and in which a simple relationship has been made to appear complicated. If the dataset is to be used to develop insight as well as to manage performance, then it must satisfy the requirements of researchers as well as those of analysts.

### The Number of ‘Starts’ and ‘Stops’ Must Be the Same

The second reason for our success is that a simple relationship exists.

We identify all of those waiting – at a given moment – to be admitted for elective investigation or treatment, and we conduct a count. The only people on the list are those whose date (and time) of enrolment preceded the date (and time) of the census and whose date of admission (or removal) succeeded it. (If obtaining this count is complicated, it is because the list has been so narrowly defined that a great number of characteristics have to be evaluated in order to decide whether a particular record should be included or not.)

The count varies from one time to another. It is not difficult to apprehend that a unit increase in its size must follow each enrolment over the interval and that a unit decrease in its size must follow each admission (or removal), if no one contributes more than one record to the dataset. It follows that the balance of enrolments and admissions (plus removals),  $E - (A + R)$ , must exactly equal any change in the size of a list,  $C^{\text{now}} - C^{\text{then}}$ , and that the completeness, accuracy, and validity of the counts ought to be questioned whenever it fails to do so.

There is nothing original about the assertion that the balance of enrolments and admissions ought to equal any change in the size of the waiting list. Mason (1976) constructed a hypothetical example which – though it was incomplete – indicated that any difference in the numbers of enrolments and admissions was expected to account for any change in size, and Fordham (1987) provided a complete example which showed the behavior of two hypothetical lists over four quarters. The Department of Health

instructed those responsible for completing the KH06, KH07, and KH07A returns to check the consistency of their submissions for each provider each quarter. “Patients waiting at the end of the quarter should be equivalent to patients waiting at the end of the last quarter plus the number of additions . . . minus the number of patients admitted in the quarter or removed from the elective admission list for other reasons. For the figures to balance, suspended patients must also be taken into account” (CRIR 1997, 32). The National Audit Office (2001b) used the relationship to verify the purported reduction in the size of the list at Surrey and Sussex Healthcare NHS Trust (England), 1998–1999: they suspected a reduction of 1800 patients where the number of elective admissions was known to have reduced, and they found – among other things – 700 new patients and 300 transfers from other hospitals who had not been added to the list.

---

## Secondary Hypotheses

### Inexplicably Complicated

One of the secondary hypotheses offered by the literature is attributed to the field of organization science. Kenis (2006, 296) claims that “[e]mpirical studies carried out in The Netherlands and elsewhere show . . . that the input of extra resources does not automatically lead to a shortening of the waiting list,” and he declares that “[w]aiting lists seem . . . to be an . . . example of a problem . . . characterized by a high level of complexity.” Neither observation is new. We don’t know who first suggested that the size of the list is influenced by many factors. But Sanmartin et al. (1998) drew attention to a plethora of factors which appeared to account for a part of the variation in size (DHSS 1975; Newton et al. 1995; Hanning and Lundström 1998) and advocated the use of complex models to evaluate their interaction and combined effect (DeCoster et al. 2007; Kreindler and Bapuji 2010).

Kenis (2006) does not tell us whether the extra resources had the intended effect on the number of admissions, and he does not tell us whether

allowance was made for the effect of variation in the number of enrolments. In other words, he has neither established that the first hypothesis needs to be replaced nor has he justified the assertion that “decisions are often taken which are based on a simplified vision [sic] of the problem, which [are] inappropriate” (Kenis 2006, 296). Kenis (2006, 296) asserts that “[g]iven a certain level of complexity of a problem[,] it will become impossible to react in an equally complex way,” and he claims that this is properly the domain of organization science. But he does not substantiate the claim that waiting lists possess the requisite level of complexity, and he has not demonstrated that the paradigm fits. Instead, he classifies the first hypothesis as an example of “our modernist-rationalist way of thinking” (Kenis 2006, 296) and – perhaps as a consequence – anticipates its failure; he does not recognize the first hypothesis as an example of double-entry bookkeeping and – perhaps as a consequence – does not anticipate its success.

### Supplier-Induced Demand

Another of the secondary hypotheses offered by the literature comes from the field of health economics. It is unfortunate that ‘supplier-induced demand’ (Culyer and Cullis 1976) envisaged a direct association between the number of admissions (or its surrogate) and the size of the list (Culyer and Cullis 1976) because the notion lay ready to hand and provided what some would think a plausible explanation. But the first hypothesis anticipates a relationship between the number of enrolments, the number of admissions, and the size of the waiting list which is mathematically exact, so there is no room for a second hypothesis until the first has proven false. Moreover, it is still necessary – once the primary hypothesis has proven false – for the secondary hypothesis to prove true.

In a cross-sectional study, we might expect to see variations between one hospital and another that are the result of differences in size of the two populations they serve. Let us imagine that there are no differences that would invalidate a simple comparison, e.g., no differences in the mix of age,

sex, and other salient factors and no differences in the indications for treatment or in the thresholds at which a patient is added to, or admitted from, the list, etc. Let us imagine that comparison reveals no difference in the rates of diagnosis specified on a suitable cross-classification of salient factors. If the only difference between one hospital and another is one of scale, then large hospitals serving large populations would report large numbers of admissions and large numbers waiting, while small hospitals serving small populations would report small numbers of admissions and small numbers waiting, i.e., we would expect a direct association between the number of admissions and the size of the list. The same reasoning would also lead us to expect a direct association between the number of admissions and the number of enrolments (Newton et al. 1995).

It is not enough to show a direct association between the number of admissions and the size of the list and attribute it to supplier-induced demand. This does not allow us to distinguish the effect of supplier-induced demand from the effect of the flow of patients on the stock (when the number of enrolments is **not** fixed and unvarying). It is also not enough to show a direct correlation between the number of admissions and the number of enrolments. This does not allow us to distinguish between the effect of supplier-induced demand and the effect of scale.

The use of the term supplier-induced demand suggests the futility of making additional resources available for elective treatment and – despite assurances to the contrary – implies that clinicians have been complicit. The way had been prepared for the notion long before the term entered the literature. Commentators viewed the waiting list “as a kind of iceberg” (Powell 1966, 39), likened the waiting list to a “bottomless pit” (Haywood 1974, 38), and thought that “trying to ‘get the waiting lists down’ [was] an activity about as hopeful as filling a sieve” (Powell 1966, 40); and the conviction that a plentiful supply might prompt burgeoning demand is (we think) older than any of these (Culyer and Cullis 1976). But the hypothesis of supplier-induced demand will prove to have been counterproductive – a diversion of attempts to

understand the dynamics of the waiting list – if we find there is no need for a second hypothesis, whether as a result of empirical data or of mathematical proof. The same will be true if the second hypothesis is found not to fit: e.g., if the number of enrolments is found to determine the number of admissions rather than *vice versa* or if the financial transaction, which serves to authorize enrolment and underwrite admission, is found to occur at some other point in the market without any further exchange in the stock-cupboard.

---

### **Why has the Effect of Enrolment Confounded Analyses to Date?**

Commentators, analysts, and researchers have shown very little interest in the effect of enrolment on the size of the waiting list. We wonder how this important confounder came to be overlooked and what might provide a sufficient incentive to correct the fault.

We assert that it is the relationship between the balance of enrolments and admissions and changes in the size of the list which is of primary concern, although it is the relationship between admissions and size which dominates the relevant literature. Such a view seems to imply that commentators, analysts, and researchers were wrong-footed at the start of the debate and that the early error has been reproduced in most of the work conducted since. Neither the scope of this chapter nor the extent of our scholarship allows this standpoint close consideration at present, but a few waymarks may be enough to indicate the route proposed.

### **Some Assumed Enrolment Was Fixed and Unvarying**

In 1963, the then Ministry of Health (MoH) for the UK published what was only its fifth memorandum on the NHS waiting list (MoH 1963b, 1). The author claims that a stationary waiting list “normally represents not a deficiency of resources . . .” – there is no imbalance of enrolments and

admissions – “but a backlog of cases . . .,” a result of the accumulated imbalances of the past. The author also says that “[a] growing waiting list may often indicate a deficiency of resources,” i.e., there is an imbalance of enrolments and admissions. But he obscures matters by asserting that the “growing waiting list . . . will generally also include an element of backlog” (MoH 1963b, 1), insisting that “[a] continuous effort will be needed to prevent a backlog from arising again” (MoH 1963b, 3). His use of the words “generally,” also “normally,” and “often” implies doubt where there is, in fact, ground for none.

Whether an individual is on the list as the result of an historic backlog or as the result of its continuing growth, the additional case can only be cleared if additional means allow the number of admissions to exceed the number of enrolments however briefly. This is what the memorandum asserts. The term “backlog” is useful if it is confined to those who are awaiting admission from a list that is stationary: if any one of these is cleared, the reduction in size is permanent. The individual will never be replaced because the number of admissions equals the number of enrolments. But if we clear anyone from a waiting list that is growing, the reduction in size is momentary. This individual will shortly be replaced by another because the number of admissions does not equal the number of enrolments, and our efforts have to be never ending.

In an earlier memorandum, the Ministry expressed the view that “the hospital service is roughly keeping pace with demand but is not appreciably succeeding in reducing the very large waiting numbers” (MoH 1954, 1). (For the sake of the narrative, we shall assume that the same author wrote both memoranda.) He seems to have thought that the size of the list was approximately stationary, that is, fixed and unvarying. As a result, he sees the problem as one of clearing the backlog (DHSS 1981a; Naylor 1991). (According to Culyer and Cullis (1976), the waiting list for all specialties (excluding psychiatry), England and Wales, showed an increase in size of 4.3% over 7 years from 444.0 thousand on 31 December 1955 to 462.9 thousand on 31 December 1962.) This is

helpfully confirmed in the memoirs of the then Minister of Health, Enoch Powell, who refers to “the circulars enjoining such devices as the use of mental hospital beds and theatres, or of military hospitals” (MoH 1963b, 1 & 3), to “the ‘waiting list at 31st December’ in the Ministry of Health’s annual reports . . . [as] . . . a reliably stable feature in an otherwise changing scene” (Culyer and Cullis 1976), and to “the special operations to ‘strafe’ the waiting lists, urged on the . . . ground that a stationary waiting list is not evidence of deficient capacity – otherwise it would lengthen – but of a backlog which, once ‘cleared off’, ought not . . . to recur” (Powell 1966, 40). The Minister confirms the understanding of his staff but considers the ground of their reasoning to have been “fallacious.” He no longer views the stationary waiting list in the same light. We disagree. The Minister’s error was in thinking the list stationary when there had been substantial variation in one at least of the factors thought to determine size, i.e., in admissions.

Had the size of the list in fact been stationary, the number of enrolments ought to have equaled the number of admissions. So it is not clear to us why anyone would expect the number of enrolments to be stationary, that is, fixed and unvarying, when “the total annual number of in-patients treated in hospitals has increased by one-sixth [16.7%], . . . since the early days of the service” (MoH 1954, 1). (Culyer and Cullis (1976) report that throughput capacity, their surrogate for elective admissions, showed an increase of 24.2% – from 11,547 cases/day in 1955 to 14,336 cases/day in 1962.) Nevertheless, the author of the memoranda feels no need to discuss the effect of variation in the number of enrolments, but he expects there to be a decrease in the size of the list if there is any increase in the number of admissions. A subsequent Secretary of State for Health and Social Services, Barbara Castle, presents her analysis in very similar terms. She knows that the list has both shrunk and swelled since MoH (1963b), but she chooses to describe it as approximately stationary: “over the past 10 years the total surgical waiting list in England and Wales has hovered at the half million



mark, with little change from 1 year to another” (DHSS 1975, 2). She seems to think it incongruous that “the number of admissions nevertheless increased by more than 7%” (DHSS 1975, 2) but like her predecessor feels no need to discuss the possibility of underlying variation in the number of enrolments.

According to Culyer and Cullis (1976, 244), “HM(63)22 . . . emphasized that a long waiting list that was numerically stationary is not normally an indication of resource deficiency in any permanent sense but represents instead a ‘backlog’ of cases which could, and should, be removed by determined short-term efforts”. The “situation is one in which the system has settled down into a kind of long-run administrative equilibrium producing a constant addition to the waiting list . . . each time period which is just sufficient to offset the numbers called from the existing waiting list during the period” (Culyer and Cullis 1976, 245). They think the Ministry envisaged a situation in which the number of enrolments “is just sufficient to offset” the number of admissions.

Frost (1980) traces this to the Annual Report of the Chief Medical Officer for the year 1962, which asserts that “a long but steady waiting list is an indication only of a backlog of work remaining from the past” and that “[i]t is only if the waiting list is steadily increasing that one has any justification for deducing . . . from waiting list data alone . . . that there is a shortage of beds” (MoH 1963a, 205). We might conclude that the list was not “steadily increasing” (Culyer and Cullis 1976) in the absence of any data on the number of elective admissions. Indeed, we would think it stationary were we to compare the size of the list in 1964 with the size of the list in 1960 ( $475,863/475,643 = 1.000$ ) or the size of the list in 1965 with the size of the list in 1951 ( $498,972/496,131 = 1.006$ ) (Powell 1966). But according to Frost (1980), the waiting list for general surgery and related specialties, England and Wales, showed an increase in size of 23.0% from 126,000 on 31 December 1949 to 155,000 on 31 December 1962.

But while Culyer and Cullis (1976) and Frost (1980) agree with our reading of the Ministry’s

position, neither attributes the failure of initiatives to the correct cause. The number of enrolments was not stationary, so a brief excess of admissions was not capable of effecting a permanent reduction in size.

### Some Only Registered Discharge (and Death)

The first dataset, which was intended to inform the administration of the NHS across England and Wales, provided even less evidence of insight. When it was implemented across the two countries in 1958, the *Hospital In-Patient Enquiry* required the completion of a printed form (HIP 1A) for a one-in-ten sample of discharges from, and deaths in, hospitals (MoH and GRO 1961a). (Several categories of discharges (and deaths) were excluded such as those originating from maternity units and psychiatric wards.) The form allowed hospitals to record the dates on which the patient had been “put on the list or booked” for the condition and had been “first sent for” to come in to hospital (MoH and GRO 1961a). Successive iterations were intended to improve the coverage, completeness, and consistency of the data.

### Doubtful Definitions

The second version of the form, which was introduced in 1967 (DHSS and OPCS 1970), established the pattern of data capture for the 18 years that followed. It allowed hospitals to continue recording the date of admission, the date of first operation, and the date of discharge (or death), but it omitted the date “first sent for.” The original definition of the “waiting time” was “[t]he interval between the date a case is placed on the waiting list, or booked, and the date of admission (or the date first sent for if the patient did not come into hospital when first offered a bed)” (MoH and GRO 1961a, 264). This suggests that length was calculated using either the date of admission, or else the date “first sent for,” depending on which gave the shorter answer. If this is correct, then the definition of length and the method of calculation subsequently changed:

**Table 8** Whose origin is acknowledged when admissions are enumerated?

The first definition was used to collect data in the years 1959–1973.	The second definition was used to collect data in the years 1974–1975.	The third definition was used to collect data in 1976–1985.
“A patient for whom the hospital had previously agreed to arrange an admission in due course, it not being possible at that time to define in advance the exact day of admission, and who comes in when sent for by the hospital” (MoH and GRO 1961a, 262).	“A patient for whom the hospital had previously arranged an admission in due course. Booked cases (non-maternity) are <b>included</b> with those who come in when sent for by the hospital” (DHSS et al. 1978, ix).	“A patient for whom the hospital had previously agreed to arrange an admission in due course, and who comes in when sent for by the hospital. Booked cases, that is those for whom an admission date has been reserved, are excluded, as are patients whose admission has been deferred whether for medical or personal reasons” (DHSS and OPCS 1987, xi–xii).

the later definition of the “waiting time” was “[t]he interval in weeks between the date a case is placed on the waiting list and the date of admission” (DHSS and OPCS 1970, 1987, xii), so the length of wait reported in 1967–1985 was longer – by definition – than in 1955–1966. We do not know why it was thought necessary to discount a part of the completed wait in the early years of the dataset, if a patient declined a reasonable offer of admission, and we do not know why the practice was abandoned in the later years of the dataset.

The definition of a “waiting list case” used in the later tabulations also differed from that used in the earlier tabulations See Table 8.

Booked cases are included under the second definition but excluded explicitly under the third and implicitly under the first: a case cannot be booked, “it not being possible at that time to define in advance the exact day of admission.” “[P]atients whose admission has been deferred” are excluded under the third definition but are not excluded under the first or second. If this is correct, then there was a change in the mix of those included in official statistics over the 31 years of the *Hospital In-Patient Enquiry*: the discharges (and deaths) which follow elective admission were more narrowly defined and made to appear less numerous in 1967–1985 than in 1955–1966. We do not know why the entire waits, of each of those temporarily suspended at any point “for medical or personal reasons,” were included in the earlier version of the dataset but not in the later.

### Event-Based Data Capture Makes some Vanish

The number of admissions should exactly equal the number of discharges in every subset of records defined on geography, or demography, or diagnostic group if the lengths of stay were always zero, and the number of admissions should approximately equal the number of discharges if the lengths of stay were short compared with the period of data capture. But not everyone admitted to hospital was eventually discharged with an appropriate diagnosis, having completed the series of investigations or the relevant course of treatment. Death accounted for 5.67% of the records submitted for 1958 (MoH and GRO 1961a, 107). Fortunately, those responsible for designing the *Hospital In-Patient Enquiry* thought it important to record the frequency and distribution of fatalities among those admitted so there were no outcomes of admission not represented in the dataset. The authors were able to claim “[a]lthough strictly related to discharges, in the majority of cases the data will approximately correspond to admissions” (MoH and GRO 1961a, 3).

It was not possible to collect information on the length of wait for admission until the HIP 1A was implemented as the first revision of the transcription form in 1952 (Registrar General 1959). Regrettably, the item “date put on the list or booked” (MoH and GRO 1961a, 298) appears to have been added without fully appreciating its implications for the dataset (Douglas 1962). The authors warn “that the . . . data presented here only

give details of those patients who are admitted to hospital” (MoH and GRO 1961b, 12). Just as discharges underestimate admissions by the number of deaths, so booked admissions and admissions from the waiting list underestimate enrolments by the number removed. “Nothing is known of those patients who did not obtain admission” (MoH and GRO 1961b, 12). But whether it is the discharges (and deaths) of the *Hospital In-Patient Enquiry* (1952–1987) or the finished consultant episodes of *Hospital Episode Statistics* (1987 to date), using an end date associated with elective admission to define the set of records, does not allow us to establish the frequency of occurrence of other outcomes or the length of wait with which they are typically associated.

Had the designers chosen to accumulate lengths of wait by sampling all of the outcomes of enrolment, the dataset would have allowed other researchers to identify cohorts of additions to the list, e.g., in 1958, and would have allowed us to examine what happened to their members prospectively. But the designers chose instead to accumulate lengths of wait by sampling only those patients who had experienced the event of interest and only those records where this had occurred within a specified period. This has left subsequent analysts and researchers with very little choice. If they want to use the existing datasets, they must be ready to assume that removal from the list is infrequent, or that it has nothing to do with the length of wait, or that the experience of this group of patients doesn’t matter. If they want to use the latest accessions to the dataset and present timely analyses, they must be prepared to examine the prior wait of the quarter’s admissions instead of the subsequent wait of the quarter’s enrolments.

It is regrettable that the event-based and period-specific data capture modeled by the *Hospital In-Patient Enquiry* has been emulated so widely. It means there are few examples where the date of an event at the start, rather than at the end, of the wait is used to define the set of records, so there has been little opportunity to demonstrate the consequences of the approach empirically. We think those responsible for funding enquiry in this area too suspicious of

novelty and too content with the existing state of affairs.

### **Period-Specific Cross-sections Estimate the Probability of Enrolment**

The dataset was constructed by combining samples from cross-sections of records where membership was defined by the date of discharge (or death), i.e., the dataset was period, rather than cohort, specific. But having used the end date to determine whether a record ought to be included or not, we are obliged to use the start date to discover the length of wait. In other words, the *Hospital In-Patient Enquiry* supplied measures which were retrospective rather than prospective – it calculated the length of wait backward. (The same is true of most of the datasets currently available to health services researchers.)

The technical terms fail to convey the incongruity of substituting one approach for the other: if we want to know how long a patient might expect to wait, the retrospective approach is akin to putting the cart in front of the horse. This is seldom appreciated because we seldom take sufficient care in defining what it is that we have calculated. Let us imagine that the dataset allows us to count all of those who were admitted as booked or waiting list cases during 1952, and to identify that proportion of these which had a prior wait of less than 3 months. Strictly speaking, it allows us to estimate the probability of being “put on the list” 0–2 months prior to being admitted. But we want to know the probability of being admitted 0–2 months after being “put on the list.” So we need to count all of those who were “put on the list” during 1952 and to identify what proportion of these had a subsequent wait of less than 3 months. Now the prior waits for the period will have the same distribution as the subsequent waits of the cohort if the waiting list happens to be stationary (and closed). But publication of the length of the prior wait for 32 out of 34 years would seem to imply very great confidence in the veracity of this assumption.

It is likely that the design of the first dataset owed something to the preferences, practices, and technologies of the day. Each form

represented a finished spell in hospital. The details of admission, investigation, diagnosis, treatment, and discharge ought to have been a matter of record. It should therefore have been possible to complete the transcription form by handling the case notes once. It should never have been necessary to submit a partially completed form with the rest of the details to follow on a second copy at a later date. This kept the work of completing the forms to a minimum. It avoided the problem of matching two (or more) forms which described the same spell; it simplified the sorting, selection, and counting of relevant discharges (and deaths); and it eliminated the possibility of double counting.

But the submission of electronic records in 1965 implies that some of the work could have been done by computer. The dataset could have been amended at this point to derive some of its inputs from those admitted and the remainder from those discharged (or dead). It would have required the submission of two records for each spell (Steering Group 1984) as a matter of routine. The first would have registered admission to hospital with all of the details known at that time, and the second would have supplemented these with the additional details established by the time of discharge. The computer would have then been used to find the appropriate admission for each discharge (or death), and to merge the two, creating a single record for each finished spell.

It would have been possible to restrict attention to the discharges (and deaths) in the dataset by selecting only those records which met the relevant criterion, e.g., a date of discharge (or death) during 1952. But it would also have been possible to restrict attention to the admissions by selecting only those records which met the relevant criterion, e.g., a date of admission during 1952 regardless of whether the spell was finished or not. Once, the submission of two records would have meant returning to the same case notes on a second occasion, with a commensurate increase in the clerical workload. But that need no longer be so. The production of an initial record about admission and a subsequent record about discharge (or death) reflects the sequence of data entry on the

*Patient Administration System*. As a result, it ought to have been possible to extend the usefulness of the *Hospital In-Patient Enquiry* with very little increase in labor once the submission of electronic records was sufficiently widespread. But the stakeholders who chose to compile records of discharges (and deaths) rather than of admissions continued to influence the design of the *Hospital In-Patient Enquiry* when it was no longer necessary to choose one rather than the other.

The usefulness (and coverage) of the dataset could have been extended had a further departure from the original design been allowed to provide information about enrolments as well as admissions and discharges (or deaths). This would have required the submission of a preliminary record, which would have registered the decision to admit and provided relevant details known at the time. Many of these patients were eventually admitted to hospital and subsequently discharged (or died), but some were removed from the list without having been admitted. In these instances, we would have wanted the second extract from the *Patient Administration System* to record the fact that the patient had been removed from the list and to record the date on which this occurred.

Had such a modification been introduced, we would now be able:

- To calculate the length of the subsequent wait (without needing 2 or more years follow-up of those most recently enrolled) (Armstrong 2010)
- To describe the characteristics and experience of a group of patients which is currently excluded from most of the available statistics

We would be able to do this without any loss of data about discharges (and deaths) and without any loss of data about admissions. We would also be able to identify all of those who were on the list and to calculate the length of each individual's wait to date (Armstrong 2010), at any specified date and time.

The construction of the *Hospital In-Patient Enquiry* changed very little between 1952 and

1985. In 1957, the Ministry invited non-participating hospitals to extend coverage by submitting forms for a one-in-ten sample of inpatients discharged (or dead). In 1974, hospitals were invited to extend coverage by submitting forms for a one-in-ten sample of all whose discharge (or death) followed treatment (or investigation) as a day case.

We do not know whether the waiting list was thought to be stationary, or not, and we do not know whether there was an understanding of the consequences of assuming that the list is stationary, when it is not. We have found no documentation which alerts users to the fact that the prior waits for a period do not have the same distribution as the subsequent waits of the cohort unless the list is stationary (and closed). There is therefore no evidence that the Government Statistical Service considered the published measures to be erroneous when the waiting list was not, in fact, stationary.

### **Design, Analysis, and Interpretation are Constrained**

The *Hospital In-Patient Enquiry* was compiled from period-specific cross-sections of those who had died in hospital, or been discharged, having been admitted electively. This method of data capture is analogous to drawing samples from each year's contribution to the filing cabinets. It is easy to understand and implement, and it is widely used and familiar. It may provide inexpensive data for the purposes of research if items are collected as a matter of routine for other purposes, but the advantage of this has always to be set against the disadvantage that records were not constructed and items not collected with the aim of this particular investigation clearly in mind. As a result, the dataset may not contain all of the necessary records, i.e., the representation it provides may be biased (Berkson 1946; MoH and GRO 1961b; Cornfield and Haenszel 1960). The dataset may not contain all of the necessary variables, i.e., the analyses it permits may not allow for confounding and effect modification. And, where the dataset seems to include the necessary variables, the data may prove insufficiently reliable, valid, sensitive, or complete.

Datasets have been constructed which make use of the inputs of hospital administration, under standard definitions and across many hospitals, in order to meet the needs of researchers as well as those of analysts. The investment which their development represents is sometimes justified in part by the benefit – unspecified and intangible – which the designers expect to accrue from subsequent investigations. But the usefulness of these datasets for the purposes of research depends upon the goals and design of investigations not yet envisaged and on the extent to which the designers have succeeded in anticipating their requirements.

The dates of compilation, the list of contributors, and the stated inclusion and exclusion criteria indicate some of the more obvious limitations of these datasets. But most also constrain researchers in a way that is not obvious. Although the datasets supply records of the wait for elective admission, researchers may not use these to conduct cohort analyses – prospective or retrospective – of all of those who were added to the waiting list. The event-based (and period-specific) method of data capture used to compile the dataset obliges researchers to examine the prior waits of those admitted and the probabilities of enrolment, e.g., 0–2 months, prior to admission when they might have preferred to examine the subsequent waits of those enrolled and the probabilities of admission, e.g., 0–2 months, after enrolment.

This constraint is an artifact of the method of data capture. The *Hospital In-Patient Enquiry* aimed to compile information about hospital morbidity. It opted to do this by collating records of discharges (and deaths) instead of records of admissions or enrolments because case notes were more likely to include diagnoses, investigations, and treatments at the later of the three events. By definition, those who were removed from the list were not admitted, and their omission from the dataset may have been quite unintentional. Their case notes contained little information about diagnoses, investigations, or procedures, no date of admission, and no date of discharge (or death). So it would have been easy to class them with incomplete records and other examples of missing

data and to assume that the error was random rather than systematic.

We do not think the designers of the *Hospital In-Patient Enquiry* fully appreciated the consequences of appending the “date put on the list or booked” (MoH and GRO 1961a, 298) to form HIP 1A (Douglas 1962). Nevertheless, they established a precedent which resulted in the popularization of a defective method and widespread publication of biased estimates. Existing methods of data capture should be amended to include outcomes of enrolment other than admission (Armstrong 2000), and new datasets should define the set of interest – wherever possible – by using the date of an event at the start of the record rather than the date of an event at the end.

### An Apparent Lack of Candor

The Ministry of Health (1963b) discussed the numbers waiting as reported in the SH3 return at the close of each year in its memorandum, HM(63)22, but it made no mention of the length of wait although the tables from the *Hospital In-Patient Enquiry* for 1955, 1956–1957, and 1958 were all available at the close of 1961. We think it unlikely that any data on the length of wait would have been ignored when the *Hospital In-Patient Enquiry* was intended to inform the administration of the NHS and the Ministry of Health was preparing to issue guidance (MoH 1963b). But the tables published during 1963 (for the 1959 and 1960 datasets) were the only ones in the series (1955–1985) which failed to report the length of wait despite collecting the dates needed to do so. The omission of appropriate statistics from the tables for 1959 (MoH and GRO 1963a) and 1960 (MoH and GRO 1963b) implies a lack of candor in the run-up to the British General Election of 1964 (Conservative, 1951–1964; Labour, 1964–1970).

The Government Statistical Service said nothing about the length of wait in 1959 and 1960 when it published its collection of historical tables in 1972. But it drew attention to an increase in “the median waiting time” and to an increase in “the proportion of those admitted who had been waiting six months or more,” when it examined the data for 1957–1960 as part of a longer

series after the British General Election of 1979 (Labor: 1974–1979; Conservative: 1979–1997). It inferred “that hospitals were losing ground, . . . between 1957 and 1967, against increasing pressure on their resources” (DHSS et al. 1979, 266).

This observation in 1979 is consistent with the views expressed in HM(63)22. Had the number of enrolments been stationary in the early 1960s, the Government Statistical Service expected a decrease in the length of wait to accompany an increase in the number of admissions. But “the proportion of those admitted who had been waiting six months or more” and “the median waiting time” was observed to increase despite an increase in the number of admissions, which suggests “increasing pressure on resources,” i.e., that the number of enrolments increased.

### Some Compiled Returns

A judgment was passed on the set of discharges (and deaths), which resulted in abolition of the *Hospital In-Patient Enquiry* after 31 December 1985 and in implementation of the *Körner Reporting System* on 1 April 1987. It was asserted that “[t]his survey is being replaced by the Körner data system” (DH and OPCS 1989, 1), i.e., that the *Körner Reporting System* replaced records of discharges (and deaths) with aggregate counts, sometimes of those admitted (or removed) from the list, sometimes of those still awaiting admission, and sometimes of those enrolled on the list. This might seem to suggest that the work of compiling the records of discharges (and deaths) had become too burdensome, even on the basis of a one-in-ten sample (MoH and GRO 1961a), or else that the English NHS had decided that a series of aggregate counts could better meet its needs and had identified those it thought necessary. But this is not the whole story. The *Körner Reporting System* replaced a number of returns in addition to the *Hospital In-Patient Enquiry*, e.g., the SBH 203 and the EDP4 and EDP5 of the SH3 (Steering Group 1984); and, even as the assertion was being published, the first records of inpatient episodes were being compiled into *Hospital Episode Statistics*. It appears that none of the criticisms made

by Working Group A on hospital clinical activity have to do with items supplied by the *Hospital In-Patient Enquiry* (DHSS 1981b).

Nevertheless, it was the *Körner Reporting System* which introduced the count of decisions to admit each quarter, the first data on the number of enrolments, additions, or accessions to be collected in almost 39 years of the UK NHS (DHSS 1986, 4; Newton et al. 1995). Counts were also proposed of the number of patients admitted, and of the number of patients removed, from the list each quarter and of the number of patients awaiting admission at the quarter's end (Steering Group 1984). The four counts seem to imply that the stock-flow model, or some version of the basic demographic equation (Newell 1988; Pressat 1985), may have informed the design of the relevant returns. But this is doubtful. Working Group A used a different model to justify its proposals to the NHS in 1981, one which claimed to provide information about demand (expressed, met, and unmet) and about attempts to supply demand (DHSS 1981b; Steering Group 1984).

We think that this is why its recommendations were presented under the heading "Information about demand for hospital facilities" (DHSS 1981b, 120) and why 'demand' was mentioned 42 times in the relevant chapter while 'stock' and 'flow' were not mentioned at all (DHSS 1981b). We think that this is why the forms were first implemented as returns about the "demand for elective admission" (DHSS 1987, 1) and why 'demand' is mentioned 13 times (and 'stock' and 'flow' are not mentioned at all) in the penultimate "DataSet Change Notice (DSCN)" of the series. We think that this interest in supply and demand is why Working Group A proposed the counting of "admission decisions" (DHSS 1981b, 129) despite the confusion of these with "admissions arranged" (DHSS 1987, 1) and why it coined the term "decision to admit" (DHSS 1981b, 123, 125–6 & 130) instead of "patients added to the list" (CRIR 1997, 2–5 of 7). We think that this is why Working Group A proposed a count of patients who were not admitted (despite arrangements having been made) as well as a count of patients who were (DHSS 1981b), and we think

that this is why the Steering Group proposed counts of those who failed to attend, counts of admissions canceled by the hospital, and counts of patients removed from a list for any reason other than elective admission (Steering Group 1984).

We know that the design of the relevant returns was not solely dependent upon the members of Working Group A. So the Steering Group added the count of patients removed from the list to the KH06 return on "events occurring during [the] period" (1984, 90) and published its recommendations before it was realized that the additional counts of the KH07A return would be required. Later versions of the KH06 return (CRIR 1997; CRIR 1998) instructed NHS Trusts to check that the counts on the KH06, KH07, and KH07A returns were consistent, although the possibility of doing this was not mentioned by Working Group A, the Steering Group, or those responsible for the development of the earliest versions of the returns (DHSS 1981b; Steering Group 1984; DHSS 1986).

Despite the addition to the KH06 return of an instruction to evaluate the consistency of the data, we have found little evidence (in 40 sets of returns submitted by each provider) that the stock-flow model, or any version of the basic demographic equation, has been used to do this. (The instruction was added no later than 1 April 1996 (CRIR 1997) and remained in force until the return was abolished on 1 April 2006 (ISB 2006).)

- The version of the KH06 return, which was issued for use from 1 April 1998 (CRIR 1998, 7 of KH06), added "[e]xplanations may be given in the box below" to the second paragraph of instructions about checking consistency, and it also added a box with the invitation [t]his area can be used for your notes and maybe [sic] used to explain any special features which have affected this return. These changes might imply that the eight previous sets of submissions contained inconsistencies large enough to warrant explanation. But there were numerous changes in this version of the return – most having to do

with format and layout and very few having any effect on the counts. (The addition of pain management to the list of main specialty functions will have generated an additional series of counts, and the counts against one (or more) of the existing categories might have diminished as a consequence.) Given that previous versions of the return invited comment on counts of ordinary (or inpatient) admissions and counts of day-case admissions, the invitation to explain any inconsistency may reflect a desire for consistent presentation rather than grounds for concern.

- The National Audit Office (2001a, 21) “was unable to reconcile” the counts. It found 24,312<sup>†</sup> more patients on the list at the close of the quarter than were accounted for by enrolments less admissions and removals (Table 3c), and the Department of Health was unable to explain the discrepancy when asked to do so. The National Audit Office (2001b) also queried an inconsistent reduction in the size of the list at Surrey and Sussex Healthcare NHS Trust (England), 1998–1989. It is not likely that this Trust had checked the consistency of its returns.
- We have been informed that “[t]he NHS Data Model and Dictionary team are not aware of any reviews or audits that [were] commissioned by the Department of Health into the internal consistency of the KH06 and KH07 returns” (personal communication, Mayet M, 24 January 2016.).

Discussing attempts “to tackle waiting-list problems,” Yates (1987, 71) claimed “there is no tradition of writing up managerial work of this type in medical, or even in management journals.” (Copyright © John Yates 1987.) The paper by White (1980) appears to be the only example of its type which survived peer review and made it into print, but it is scarcely possible that he was the only analyst in England and Wales who was interested in the relationship between inputs, outputs, and the size of outpatient and inpatient waiting lists. So while the lack of documentary evidence suggests that NHS Trusts and District Health Authorities did not

check the internal consistency of the KH06 and KH07 returns, this is a conclusion we are not yet ready to draw.

The four counts used to describe the inpatient waiting list might have been consistent when first proposed (Steering Group 1984; DHSS 1986; IMG 1992). The Steering Group (1984, 87) recommended counting the “[n]umber of patients for whom a decision-to-admit has been made,” the “[n]umber of patients admitted electively,” the “[n]umber of patients . . . removed from a list,” and the “[n]umber of patients still awaiting admission.”

It appears to have discounted – at least for the purposes of the narrative – the possibility that an individual might require elective investigation or treatment more than once a quarter. Instead, it claims that “a cohort of all the patients for whom a decision to admit has been made during a specified time period can be followed up at regular intervals and the number in the cohort admitted at different times recorded” (Steering Group 1984, 86). The members of the cohort are “patients for whom a decision to admit has been made,” which seems to imply a single decision to admit per patient. Moreover, the cohort is “followed up at regular intervals” to identify those no longer awaiting the outcome of interest, i.e., “the number . . . admitted,” which indicates that a member either has, or has not, been admitted “at different times” and seems to imply a single outcome per patient. The narrative does not mention removal from the list for reasons other than admission.

We do not think the Steering Group ignorant of the possibilities. It understood that while the counts describing the outpatient waiting list might be correlated, they were not consistent. Alluding to the decision to admit to the list, the Steering Group claims that “[p]iloting and consultation have shown the practical difficulty of capturing and recording any requests other than those made in writing. It is however feasible to record the number of written requests made by general practitioners and changes in this statistic should reflect changes in the total number of requests” (Steering Group 1984, 87).

But the Steering Group (1984, 87) also recommended regular reports of the “[n]umber of patients for whom arrangements to admit were



made but who were not admitted because they failed to attend” and of the “[n]umber of patients for whom arrangements were made but admission did not take place because of cancellation by the hospital.” If these are understood to be alternative outcomes of enrolment, then admission and removal by definition cannot provide a consistent account for the change in the size of the list.

And the definitions of the four counts used to describe the inpatient waiting list were not wholly consistent in subsequent iterations of the *Körner Reporting System* (CRIR 1997; CRIR 1998). The CRIR Secretariat (1997, 32) asserts that “[p]atients waiting at the end of the quarter should be equivalent to patients waiting at the end of the last quarter plus the number of additions and minus the number of patients admitted in the quarter or removed from the elective admission list for other reasons.” This is what we would expect if (a) the date of addition marked the start of each wait, (b) the date of admission (or of removal) marked the end of each wait, and (c) if everyone waiting was eligible for admission on any and all of the intervening dates. But not everyone was considered eligible for admission on any and all of the dates separating their addition to the list from their removal.

“For the figures to balance,” providers were told, “suspended patients must also be taken into account” (CRIR 1997, para. 164). There are two ways of doing this.

The first of these handles the count of those suspended as though it was a flow. The number suspended that quarter is added to decisions to admit this quarter as though that number were reinstated this quarter (Armstrong 2000), and the number suspended this quarter is added to the number removed. So we expect

$$(E + S^{\text{then}}) - (A + R + S^{\text{now}}) = C^{\text{now}} - C^{\text{then}}, \quad (3.2)$$

where  $S^{\text{then}}$  represents those reinstated to the list, and  $S^{\text{now}}$  represents those removed from the list, this quarter. The *Körner Reporting System* does not tell us how many were suspended over the course of the quarter.  $S^{\text{now}}$  estimates the count in question by assuming that each suspension lasts

one quarter on average. Moreover, the *Körner Reporting System* does not tell us how many were reinstated over the course of the quarter.  $S^{\text{then}}$  estimates the count in question by assuming that each suspension lasts one quarter on average and that everyone suspended is duly reinstated (CRIR 1997).

The second handles the count of those suspended as though it was a stock. The number suspended at the end of that quarter is added to the count of those awaiting admission at that date, and the number suspended at the end of this quarter is added to the count of those awaiting admission at this date. So we expect

$$E^{\text{now}} - (A^{\text{now}} + R^{\text{now}}) = (C^{\text{now}} + S^{\text{now}}) - (C^{\text{then}} + S^{\text{then}}), \quad (3.3)$$

where  $S^{\text{now}}$  represents those suspended from the list at the time of this census, and  $S^{\text{then}}$  represents those suspended from the list at the time of that census. We do not need to make any assumptions about the length of suspension or the frequency of reinstatement under this approach. Instead, we expect the balance of enrolments less admissions (and removals) to account for the difference between the censuses once we have corrected those counts by adding back the suspended.

Formulae (3.2) and (3.3) are equivalent. But formula (3.2) tells us that enrolment *and reinstatement* cause the official list ( $C^{\text{now}}$ ,  $C^{\text{then}}$ ) to swell and that admission, removal, *and suspension* cause it to shrink, whereas formula (3.3) provides a simpler account – the number waiting increases as a result of enrolment and decreases as a result of admission and removal – but the list ( $C^{\text{now}} + S^{\text{now}}$ ,  $C^{\text{then}} + S^{\text{then}}$ ) is not the one reported in the Press. With a little rearrangement, both formulae yield the relationship which providers were to use to check the consistency of their counts of inpatients and of day cases (CRIR 1997, para. 164 & p. 6 of KH06), namely,

$$C^{\text{now}} = (C^{\text{then}} + S^{\text{then}}) + E^{\text{now}} - (A^{\text{now}} + R^{\text{now}}) - S^{\text{now}}, \quad (2.1)$$

so the two approaches give identical results. The CRIR Secretariat claims that “[t]he change in the total numbers waiting should reflect this activity,” that is, “all the additions to the waiting list (i.e., the number of decisions to admit) and removals from the waiting list that have taken place during the quarter” (CRIR 1997, para. 144). If we understand “the total numbers waiting” to include those suspended, i.e.,  $C^{\text{now}} + S^{\text{now}}$  and  $C^{\text{then}} + S^{\text{then}}$ , this statement would seem to imply the relationships of formula (3.3). But if we understand the “total . . . of all patients waiting for admission” to exclude those suspended (CRIR 1997, para. 155), the statement would seem to imply the relationships of formula (3.2). Given that the data about suspensions ( $S^{\text{now}}$ ,  $S^{\text{then}}$ ) were obtained by taking a census (CRIR 1997), formula (3.3) is the model which ought to be used.

Having demonstrated the consistency of the data by adjusting for suspensions (CRIR 1997), we ought to be willing to acknowledge – in the first instance – that it is “the total numbers waiting” and not the official numbers which reflect the balance of enrolments less removals (and admissions) and, in the second instance, that it is the balance of enrolments and reinstatements less admissions (and removals and suspensions) which changes the official numbers and not “the total numbers waiting.”

Now some patients will require elective treatment (or investigation) on more than one occasion (IMG 1992). Some will require treatment (or investigation) for the same condition, will undergo the same procedure, and will appear on the same list, on two (or more) occasions. The NHS accepts that the manager ought not to be held responsible for that part of any wait over which she can be expected to exercise no control. So if a patient is admitted to the same waiting list twice (CRIR 1997), e.g., for extraction of two cataracts, she is not considered as waiting for the second operation until she has been discharged from hospital after the first. But the data model implied by this is more complicated than that in which each patient is (assumed) to require just one admission or in which we count, for example, the number of decisions to admit – rather than the number of individuals added – to the list. The instructions

for the KH07 return are simple: by definition, no patient can wait for more than one procedure at a time, so no patient may be counted more than once in the census at the end of the quarter. But the instructions for the KH06 return are not simple: if the dates of the decisions to admit fall in the same quarter for both procedures, the count of decisions to admit must not include the second of them; and if the dates of the admissions do not fall in the same quarter for both procedures, the count of admissions must include the second of them. We think that the date of admission (or removal) for the subsequent procedure will be counted more often than the date of the decision to admit which preceded it. So the consistency of the four counts was impaired when the KH07 was modified to exclude all of those ‘awaiting’ an additional procedure and the KH06 was modified to exclude those ‘awaiting’ a second procedure only when the first procedure had not yet been completed.

While the terms stock and flow have not been used in any document about the KH06, KH07, and KH07A returns or in any of the official commentary, they were introduced as labels for the datasets which took their place. DSCN 09/2006, which announced the “data flow” intended to replace the tabulated content of the returns (ISB 2006, 1), mentioned ‘stock’ 29 times and did not mention demand once. (It also mentioned ‘flow’ 41 times, but not all of these were to do with the events previously recorded by the KH06.) Despite this, there seems to be little understanding of the relationships implied by the stock-flow model even when the terms are used extensively. The definitions of the four counts represented either as a ‘stock’ or as a “flow” are not perfectly consistent (ISB 2006, 44 & 46).

Dr. A. Mason, who had previously demonstrated an excellent understanding of the relationship between stock and flow (Mason 1976), was a member of the Secretariat and therefore party to the deliberations both of Working Group A and the Steering Group. Now Working Group A claimed that “information is required about the balance between referrals and the number seen . . . [t]o identify whether the number of patients waiting for an out-patient appointment is increasing or

decreasing” (DHSS 1981b, 122). It also claimed that “information is required about the balance between expressed and met demand” (DHSS 1981b, 123), presumably in order to determine whether the number of patients waiting for an inpatient admission is increasing or decreasing. Nevertheless, we fear that neither the stock-flow model nor the basic demographic equation had much influence on the analysis of the data. The English NHS appears to have collected relevant counts for 24 years (1 April 1987–31 March 2010) without ever testing its convictions about the effect of enrolment on the size of the list, and it appears to have done so for 10 of these despite instructions to check the consistency of the counts (1 April 1996–1 April 2006).

The KH06, KH07, and KH07A returns were abolished on 31 March 2010, on the grounds that the suite of 18-week referral to treatment times adequately met the needs of users. But this dataset has failed to provide any information about the number of enrolments, additions, or accessions for 5½ years (31 March 2010–1 October 2015 (Analytical Services 2015)). The deficiency has now been rectified, ostensibly to allow the reintroduction of a check on the consistency of the four counts, i.e., of “new RTT clock starts” (*E*), “completed RTT pathways” (*A*), “validation removals” (*R*), and changes in the size of the list ( $C^{\text{now}} - C^{\text{then}}$ ). But Analytical Services did not explain why we expect start dates and end dates to yield exactly the same count of those eligible for admission at any point during the month of interest (Analytical Services 2015, 8). It is perhaps not surprising that it permits “a reasonable tolerance” for the consistency check as did the CRIR Secretariat before it (CRIR 1997, 6 of KH06).

### Some Made Hay

Culyer and Cullis (1976) note that the size of the waiting list for England and Wales has not decreased as a result of increases in the number of admissions. They claim “that no one has to date succeeded in formulating a systematic and testable model to explain the phenomena ... satisfactorily” (Culyer

and Cullis 1976, 251), and they advocate “[a]n alternative approach, likely to appeal to those who prefer not to reject the supply/demand approach entirely” (Culyer and Cullis 1976, 247). But “despite very diligent searching” (Culyer and Cullis 1976, 264), and despite emphasizing the “one behavioural law that has never been refuted” (Culyer and Cullis 1976, 244), they are obliged to confess that “we have been unable to uncover any systematic and reliable empirical relationships among the relevant variables, nor have we been able to devise a plausible ‘behavioural’ model that has led to the specification of such a set of relationships” (Culyer and Cullis 1976, 264). Culyer and Cullis (1976) claim that the first hypothesis has failed without realizing that it has not been subject to a fair trial. They attempted to construct a model without considering the effect of variation in the number of enrolments.

Researchers continue to find fresh evidence of the direct association between the number of admissions (or an appropriate surrogate) and the size of the list (Buttery and Snaith 1980; Frost 1980), which Culyer and Cullis (1976) viewed as indicating the failure of the first hypothesis. Moreover, there appears to have been little diminution in the popularity of the “one behavioural law that has never been refuted” as a result of Culyer and Cullis’s inability to implement it satisfactorily. The direct relationship continues to be explained by the appetites of those who enter the marketplace to sell (supplier-induced demand) rather than the appetites of those who enter the marketplace to buy.

The Institute of Social and Economic Research received support from the Department of Health and Social Security “for ... research into the economics of waiting lists.” It received a grant, and Culyer and Cullis (1976, 239) “benefited enormously from discussions with DHSS officials,” which may be why the DHSS turned to the Institute for advice. But it is likely that the enquiry was also prompted by prevailing opinion, e.g., by “Parkinson’s Law of Hospital Beds” (Powell 1966, 43) and “Say’s Law of Hospitals” (Culyer and Cullis 1976, 244), and by the expressions of other economists (Feldstein 1967) earlier on the scene. Whatever the reason, the DHSS

chose to consult economists rather than the members of any other school of social science. It is perhaps no surprise that supplier-induced demand has become the dominant paradigm in the literature from the UK.

---

### The Primary Hypothesis Has Not Been Falsified

The Ministry of Health (1954, 5) recommended “the careful and regular study of such figures as . . . size of waiting list in proportion to number . . . of patients treated, degree of urgency of need of patients on the waiting list, numbers waiting for defined periods and such other indices as are available in published documents.” In other words, the Ministry expected the compilation of information about the numbers waiting and the numbers admitted, but it did not expect the compilation of information about the numbers enrolled. It is therefore not surprising that the *Hospital In-Patient Enquiry* did not provide counts of the numbers enrolled in England and Wales.

The omission was hallowed by successive datasets, first by those that relied on printed forms and second by those that relied on electronic media for their inputs. The national dataset compiled records after investigation or treatment had been completed, and these records were collated by the period in which the event was registered, i.e., by the date of discharge (or death) (Registrar General 1959). This architecture facilitated the counting and cross-classification of discharges (and deaths), and it reflected our need for data on morbidity (Registrar General 1959).

But we are also interested in the use made of the costlier resources. This has expressed itself in an interest in the length of stay and therefore in the occurrence of admission as well as discharge (or death). Extracting information about admissions from a collection of discharges (and deaths) is a little involved. It is not difficult to obtain the information we require when we have both the date of admission and the date of discharge (or death), but we face the problem of our choices while we await the date of the second event. If we have chosen to register discharges and deaths, we

can determine how many of these were admitted during the period of interest, and we can calculate the length of their completed stay. But we have no information about those who have yet to be discharged (or to die): we cannot determine how many of them were admitted during the period of interest, and we cannot calculate the length of their incomplete stay. And if we have chosen to register admissions, we can count them and calculate the length of stay with ease, i.e., we know which of those admitted during the period of interest have yet to (die or) be discharged, but we have little information about the outcome of their admission, e.g., diagnosis, treatment, and destination of discharge.

Extracting information about enrolments from a collection of discharges (and deaths) is more involved. If we are to obtain a complete set of enrolments, we must:

- Identify those who have been discharged (or who died) following admission from the waiting list.
- Identify those who have not been discharged (or have not died) following admission from the list.
- Identify those who have not been admitted from the list.

This third group includes (i) some who will be admitted from the list and who will, in due course, be numbered among the discharged (or dead), and it includes (ii) others who – having been removed from the list – will never be admitted and will therefore never be numbered among the discharged (or dead).

We face the problem of our choices. If we had chosen to register patients immediately after their enrolment on the list, instead of after their discharge from hospital, it would be easy to determine the size of a cohort and to cross-classify its members. But the architecture of successive datasets in England prized economy of effort: it set about capturing the requisite variables, and relevant records, in a single pass. This can only be done using discharges (and deaths). If we attempt to compile our records on admission, some data about the outcome of admission will be missing.

These details could be supplied by taking a second pass at a later date and replacing any record which was incomplete with the now completed version.

There has been no attempt to construct a national dataset from enrolments in England using repeated passes to upload the latest details from the most recent accessions. And there has been no attempt to construct an equivalent dataset out of discharges (and deaths) for the purposes of longitudinal research, where timeliness is much less of an issue. But the relationship between enrolments, admissions (and removals), and the size of the list cannot be assessed empirically using the dataset available (*Hospital Episode Statistics*). It would not be reasonable however to attribute a lack of interest in the effect of enrolment to the lack of relevant data. The Department of Health and Social Security instructed hospitals to report the number of enrolments as aggregate counts between 1 April 1987 and 31 March 2010, by completing the KH06 return on a quarterly basis. Nonetheless, there is little evidence that this data has been used to check the reliability of the counts or the validity of the relationship hypothesized.

## References

- Analytical Services. Aligning the publication of performance data – statistics consultation. Leeds: NHS England; 2015. p. 8. <https://www.engage.england.nhs.uk/consultation/aligning-publication-performance-data>. Accessed 11 July 2016. Contains public sector information licensed under the Open Government Licence v3.0.
- Armstrong PW. First steps in analysing NHS waiting times: avoiding the 'stationary and closed population' fallacy. *Stat Med*. 2000;19:2037–2051. By permission of John Wiley and Sons. [https://doi.org/10.1002/1097-0258\(20000815\)19:15<2037::AID-SIM606>3.0.CO;2-R/pdf](https://doi.org/10.1002/1097-0258(20000815)19:15<2037::AID-SIM606>3.0.CO;2-R/pdf).
- Armstrong PW. Spotting the pantomime villain: do the usual approaches correctly indicate when waiting times got shorter? *Health Serv Manag Res*. 2010;23:103–115. By permission of SAGE. <https://doi.org/10.1258/hsmr.2009.009021>.
- Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics*. 1946;2:47–53.
- Buttery RB, Snaith AH. Surgical provision, waiting times and waiting lists. *Health Trends*. 1980;12:57–61.
- Carvel J. Tories doubt fall in hospital waits. *Guardian*, 10 Jan 2004, p. 6.
- Committee for Regulating Information Requirements (CRIR) Secretariat. Central returns: waiting times. DSCN: 10/98/P10. Birmingham: NHS Executive; 1998. p. 3, 7 of KH06. Contains public sector information licensed under the Open Government Licence v3.0.
- Committee for Regulating Information Requirements (CRIR) Secretariat. Patients awaiting elective admission. In: *The Data Manual*. Hospital services module, version 4.0. Birmingham: Information Management Group, NHS Executive; 1997. p. 7, 12–4, 16–7, 29–32, 2–6 of 7, 3 of 4. Contains public sector information licensed under the Open Government Licence v3.0.
- Cornfield J, Haenszel W. Some aspects of retrospective studies. *J Chronic Dis*. 1960;11:523–34.
- Culyer AJ, Cullis JG. Some economics of hospital waiting lists in the NHS. *J Soc Policy*. 1976;5(3):239–64. By permission of Cambridge University Press.
- DeCoster C, Chateau D, Dahl M, et al. Waiting times for surgery, Manitoba 1999/2000 to 2003/04. Winnipeg: Manitoba Centre for Health Policy; 2007. p. 6, 37–8, 53, 59. [http://mchp-appserv.cpe.umanitoba.ca/reference/swt\\_3web.pdf](http://mchp-appserv.cpe.umanitoba.ca/reference/swt_3web.pdf). Accessed 11 July 2016.
- Department of Health and Social Security (DHSS). A report of the working groups A to the steering group on health services information. London: NHS/DHSS Steering Group on health services information; 1981b. p. 120–30. Contains public sector information licensed under the Open Government Licence v3.0.
- Department of Health and Social Security (DHSS). Management services. Demand for elective admission: statistical returns KH06, KH07 and KH07A. SM(87)2/8. Blackpool: Statistics and Research Division 2A; 1987. p. 1. Contains public sector information licensed under the Open Government Licence v3.0.
- Department of Health and Social Security (DHSS). Management services. Post Korner aggregate statistical returns. SM(86)2/11. Blackpool: Statistics and Research Division 2A Fylde; 1986. p. 4. Contains public sector information licensed under the Open Government Licence v3.0.
- Department of Health and Social Security (DHSS). Orthopaedic services: waiting time for out-patient appointments and in-patient treatment. Report of a working party to the Secretary of State for Social Services. London: DHSS; 1981a. p. 11, 24, 33, 42, 76, 80–1. <http://nhsreality.wordpress.com/2015/01/>. Accessed 11 July 2016.
- Department of Health and Social Security (DHSS). Reduction of waiting times for in-patients admission: management arrangements. HSC(IS)181. London: DHSS; 1975. p. 1–4. Contains public sector information licensed under the Open Government Licence v3.0.
- Department of Health and Social Security, Office of Population Censuses and Surveys (DHSS & OPCS). Hospital in-patient enquiry, summary tables. Based on a one in ten sample of NHS patients in hospitals in England, 1985. MB4 no. 26. London: HMSO; 1987.

- p. xi–xii. Contains public sector information licensed under the Open Government Licence v3.0.
- Department of Health and Social Security, Office of Population Censuses and Surveys (DHSS & OPCS). Report on hospital in-patient enquiry for the year 1967. Part I. Tables. London: HMSO; 1970. p. 298–9. Contains public sector information licensed under the Open Government Licence v3.0.
- Department of Health and Social Security, Office of Population Censuses and Surveys, Welsh Office. Hospital in-patient enquiry. Main tables. Based on a one in ten sample of NHS patients in hospitals in England and Wales, 1974, Series MB4, no. 2. London: HMSO; 1978. p. ix. Contains public sector information licensed under the Open Government Licence v3.0.
- Department of Health and Social Security, Office of Population Censuses and Surveys, Welsh Office. Hospital in-patient enquiry. Patterns of morbidity. Based on a one in ten sample of NHS patients in hospitals in England and Wales, 1962–67, Series MB4, no. 3. London: HMSO; 1979. p. 266. Contains public sector information licensed under the Open Government Licence v3.0.
- Department of Health, Office of Population Censuses and Surveys (DH & OPCS). Hospital in-patient enquiry in-patient and day case trends. Based on a nominal one in ten sample of NHS patients in hospitals in England 1979–1985, Series MB4, no. 29. London: HMSO; 1989. p. 1. Contains public sector information licensed under the Open Government Licence v3.0.
- Douglas JWB. Ministry of Health and General Register Office: report on hospital in-patient enquiry for the year 1958: Part II. London: HMSO; 1961. p. 301. 17s. 6d. *Popul Stud (Camb)* 1962; 16(2):196.
- Farquharson D. Waiting times management in Lothian. Edinburgh: NHS Lothian; 2011. p. 7. [http://www.scot.nhs.uk/parliament/parliamentary\\_documents/2012.01.09\\_to\\_DM\\_-\\_report\\_from\\_NHS\\_Lothian\\_on\\_waiting\\_times\\_management.pdf](http://www.scot.nhs.uk/parliament/parliamentary_documents/2012.01.09_to_DM_-_report_from_NHS_Lothian_on_waiting_times_management.pdf). Accessed 11 July 2016.
- Faulkner A, Frankel S. Delayed access to non-emergency NHS services. A review of NHS waiting times and waiting list research issues. Bristol: Health Care Evaluation Unit, University of Bristol; 1993. p. 23, 84.
- Feldstein MS. Economic analysis for health service efficiency. Amsterdam: North-Holland; 1967. p. 152, 200.
- Finn C. The management, collection and publication of acute day and inpatient waiting lists. Dublin: Institute for the Study of Social Change, University College Dublin; 2004. p. 12–7.
- Fordham R. Managing orthopaedic waiting lists. Discussion paper no. 27. York: Centre for Health Economics, University of York; 1987. p. 9. <http://www.york.ac.uk/che/pdf/dp27.pdf>. Accessed 11 July 2016.
- Fowkes FGR, Page SM, Phillips-Miles D. Surgical manpower, beds and output in the NHS: 1967–1977. *Br J Surg*. 1983;70:114–6.
- Frost CEB. How permanent are NHS waiting lists? *Soc Sci Med*. 1980;14C:1–11.
- Goldacre MJ, Lee A, Don B. Waiting list statistics. I: relation between admissions from waiting list and length of waiting list. *Br Med J (Clin Res Ed)*. 1987;295:1105–8.
- Hamblin R, Harrison A, Boyle S. Access to elective care: why waiting lists grow? London: King's Fund; 1998. p. 12–5, 26, 58. <http://kingsfund.koha-ptfs.eu/cgi-bin/koha/opac-detail.pl?biblionumber=20657>. Accessed 17 Aug 2016. By permission of The King's Fund.
- Hanning M, Lundström M. Assessment of the maximum waiting time guarantee for cataract surgery. The case of a Swedish policy. *Int J Technol Assess*. 1998;14: 180–93.
- Harvey I, Webb M, Dowse J. Can a surgical treatment centre reduce waiting lists? Results of a natural experiment. *J Epidemiol Community Health*. 1993;47:373–6.
- Haywood SC. Managing the health service. London: Allen & Unwin; 1974. p. 38.
- Hinde A. The lexis chart. In: Demographic methods. London: Arnold; 1998. p. 12–3.
- House of Commons Health Committee. Public expenditure on Health and Personal Social Services 2009. Memorandum received from the Department of Health containing replies to a written questionnaire from the Committee. London: The Stationery Office; 2010. p. 132–4. <http://www.publications.parliament.uk/pa/cm200910/cmselect/cmhealth/269/269i.pdf>. Accessed 11 July 2016. Contains public sector information licensed under the Open Government License v3.0.
- Hurst J, Siciliani L. Tackling excessive waiting times for elective surgery: a comparison of policies in twelve OECD countries. Paris: OECD; 2003. <https://doi.org/10.1787/108471127058>. Accessed 11 July 2016.
- Information Management Group (IMG). Patients awaiting elective admission. In: The Data Manual. Hospital services module, version 1.0. Birmingham: NHS Management Executive, Department of Health; 1992. p. 5, 8–10, 14–8. Contains public sector information licensed under the Open Government License v3.0.
- Kenis P. Waiting lists in Dutch health care. An analysis from an organization theoretical perspective. *J Health Organ Manag*. 2006;20(4):294–308. By permission of Emerald. <https://doi.org/10.1108/1477260610680104>.
- Kreindler SA. Policy strategies to reduce waits for elective care: a synthesis of international evidence. *Br Med Bull*. 2010;95:7–32.
- Kreindler SA, Bapuji SB. Evaluation of the WRHA prehabilitation program. Winnipeg: Winnipeg Regional Health Authority; 2010. p. 73–9.
- Lee A, Don B, Goldacre MJ. Waiting list statistics. II: an estimate of inflation of waiting list length. *Br Med J (Clin Res Ed)*. 1987;295:1197–8.
- Mason A. An epidemiological approach to the monitoring of hospital waiting list statistics. *Proc R Soc Med*. 1976;69:939–42.

- Ministry of Health (MoH). National Health Service. The more effective use of hospital beds. HM(54)89. London: Ministry of Health; 1954. p. 1.
- Ministry of Health (MoH). On the state of the public health. The annual report of the Chief Medical Officer of the Ministry of Health for the year 1962. London: HMSO; 1963a. p. 205–7.
- Ministry of Health (MoH). Reduction of waiting lists, surgical and general. HM(63)22. London: Ministry of Health; 1963b.
- Ministry of Health (MoH). Report of the Ministry of Health for the year ended 31st December 1963. The health and welfare services. 1963–64 Cmnd. 2389. London: HMSO; 1964. p. 44.
- Ministry of Health, General Register Office (MoH & GRO). Report on hospital in-patient enquiry for the two years 1956–1957. London: HMSO; 1961b. p. 12.
- Ministry of Health, General Register Office (MoH & GRO). Report on hospital in-patient enquiry for the year 1958. Part II. Detailed tables and commentary. London: HMSO; 1961a. p. 107, 262, 264, 298–99.
- Ministry of Health, General Register Office (MoH & GRO). Report on hospital in-patient enquiry for the year 1959. Part II. Detailed tables and commentary. London: HMSO; 1963a.
- Ministry of Health, General Register Office (MoH & GRO). Report on hospital in-patient enquiry for the year 1960. Part II. Detailed tables and commentary. London: HMSO; 1963b.
- Moral L, de Pancorbo CM. Surgical waiting list reduction programme. The Spanish experience. In: HOPE sub-committee on coordination, editor. Waiting lists and waiting times in health care. Managing demand and supply. Leuven: European Hospital and Healthcare Federation (HOPE); 2001. p. 7, 10–7, 48–9. <http://www.hope.be/documents-library/>. Accessed 11 July 2016.
- National Audit Office. Inappropriate adjustments to NHS waiting lists. Report by the Comptroller and Auditor General. HC452. Session 2001–2002: 19 December 2001. London: The Stationery Office; 2001b. p. 27. <https://www.nao.org.uk/report/inappropriate-adjustments-to-nhs-waiting-lists/>. Accessed 11 July 2016.
- National Audit Office. Inpatient and outpatient waiting in the NHS. Report by the Comptroller and Auditor General. HC 221. Session 2001–2002: 26 July 2001. London: The Stationery Office; 2001a. p. 21. <https://www.nao.org.uk/report/inpatient-and-outpatient-waiting-in-the-nhs/>. Accessed 11 July 2016. By permission of the National Audit Office.
- National Audit Office Wales. NHS waiting times in Wales. Volume 1 – the scale of the problem. Cardiff: The Stationery Office; 2005. p. 7, 11, 18–9, 32–4, 43.
- National Waiting Times Unit (NWTU). Managing waiting times. A good practice guide. Edinburgh: Scottish Executive; 2003. p. 4–5. <http://www.gov.scot/Publications/2003/09/18035/25483>. Accessed 11 July 2016.
- Naylor CD. A different view of queues in Ontario. *Health Aff (Millwood)*. 1991;10(3):110–28.
- Naylor CD, Slaughter P, Sykora K, et al. Waits and rates: the 1997 ICES report on Coronary Surgical capacity for Ontario. Toronto: Institute for Clinical Evaluative Sciences; 1997. p. 14.
- Newell C. The basic demographic equation. In: *Methods and models in demography*. London: Belhaven Press; 1988. p. 8.
- Newton JN, Henderson J, Goldacre MJ. Waiting list dynamics and the impact of earmarked funding. *BMJ*. 1995;311:783–5.
- NHS Information Standards Board (ISB). Measuring and recording of waiting times. DSCN: 09/2006. Birmingham: NHS Management Executive; 2006. p. 1, 44, 46. Contains public sector information licensed under the Open Government License v3.0.
- NHS/DHSS Steering group on Health Services Information (Steering Group). A report on the collection and use of information about hospital clinical activity in the National Health Service. London: HMSO; 1984. p. 27–8, 86–90, 131. Contains public sector information licensed under the Open Government Licence v3.0.
- Niinimäki T. Increasing demands on orthopedic services. *Acta Orthop Scand*. 1991;62(S241):42–3.
- Nordberg M, Keskimäki I, Hemminki E. Is there a relation between waiting-list length and surgery rate? *Int J Health Plann Manage*. 1994;9:259–65.
- Powell JE. Supply and demand. In: *A new look at medicine and politics*. London: Pitman Medical; 1966. p. 39–40. <http://www.sochealth.co.uk/national-health-service/health-care-generally/history-of-healthcare/a-new-look-at-medicine-and-politics-4/>. Accessed 11 July 2016.
- Pressat R. Balancing equation. In: Wilson C, editor. *The dictionary of demography*. Oxford: Blackwell; 1985. p. 15.
- Purcell J. The waiting list initiative. Report on value for money examination. Dublin: Office of the Comptroller and Auditor General; 2003. p. 8, 17, 23, 26, 28. <http://www.audgen.gov.ie/ViewDoc.asp?DocId=-1&CatID=5>. Accessed 11 July 2016. By permission of the Office of the Comptroller and Auditor General.
- Registrar General. Statistical review of England and Wales for the year 1955. Supplement on hospital in-patient statistics. London: HMSO; 1959. p. 2.
- Sanmartin C, Barer ML, Sheps SB. Health care waiting lists and waiting times: a critical review of the literature. In: *Waiting lists and waiting times for health care in Canada: more management, more money*. Ottawa: Health Canada; 1998. p. 196, 198, 241–54, 270, 281. <http://publications.gc.ca/site/eng/9.6471111/publication.html>. Accessed 11 July 2016.
- Smethurst DP, Williams HC. Self-regulation in hospital waiting lists. *J R Soc Med*. 2002;95:287–9.
- Snaith AH. Supply and demand in the NHS. *Br Med J*. 1979;1(6171):1159–60.

- Street A, Duckett S. Are waiting lists inevitable? *Health Policy*. 1996;36:1–15. By permission of Elsevier. [https://doi.org/10.1016/0168-8510\(95\)00790-3](https://doi.org/10.1016/0168-8510(95)00790-3).
- Sykes PA. DHSS waiting list statistics – a major deception? *Br Med J (Clin Res Ed)*. 1986;293:1038–9.
- Torkki M, Linna M, Seitsalo S, et al. How to report and monitor the performance of waiting list management. *Int J Technol Assess*. 2002;18(3):611–8.
- White A. Waiting lists. A step towards representation, clarification and solving of information problems. *Hosp Health Serv Rev*. 1980;76(8):270–4.
- Worthington D. Hospital waiting list management models. *J Oper Res Soc*. 1991;42(10):833–43.
- Yates J. *Why are we waiting? An analysis of hospital waiting lists*. Oxford: Oxford University Press; 1987. p. 71. By permission of Oxford University Press.





# Waiting Times: Evidence of Social Inequalities in Access for Care

# 15

Luigi Siciliani

## Contents

<b>Introduction</b> .....	346
<b>Sources of Inequalities in Waiting Times</b> .....	347
<b>Data and Empirical Methods</b> .....	348
Data .....	348
Methods .....	349
<b>A Review of the Evidence</b> .....	354
International Studies: Evidence from SHARE and the Commonwealth Fund .....	355
United Kingdom .....	355
Australia .....	356
Norway .....	357
Sweden .....	358
Canada .....	358
Germany .....	358
Spain .....	359
Italy .....	359
<b>Conclusions and Implications for Policy</b> .....	359
<b>References</b> .....	360

## Abstract

Equity is a key policy objective in many publicly funded health systems across OECD countries. Policymakers aim at providing access based on need and not ability to pay. This chapter focuses on the use of waiting times for studying inequalities of access to care. Studies of inequalities in waiting times

by socioeconomic status are relatively rare, the traditional focus being on measurement of inequalities in healthcare utilization. Waiting time data are readily available for the analysis through administrative databases. They are commonly used for reporting on health system performance. Within publicly funded health systems, the duration of the wait is supposed to be the same for patients with different socioeconomic status for a given level of need. Patients with higher need or urgency are supposed to wait less based on implicit or explicit prioritization rules. A recent empirical literature

---

L. Siciliani (✉)  
Department of Economics and Related Studies, University  
of York, York, UK  
e-mail: [luigi.siciliani@york.ac.uk](mailto:luigi.siciliani@york.ac.uk)

seems however to suggest that within several publicly funded health systems, nonprice rationing does not guarantee equality of access by socioeconomic status. Individuals with higher socioeconomic status (as measured by income or educational attainment) tend to wait less for publicly funded hospital care than those with lower socioeconomic status. This negative gradient between waiting time and socioeconomic status may be interpreted as evidence of inequity within publicly funded systems which favors rich and more-educated patients over poorer and less-educated ones. The chapter provides an overview of methods and data to investigate the presence of social inequalities in waiting times and highlights key results.

---

## Introduction

Equity is a key policy objective in publicly funded health systems. In many OECD countries, this takes the form of payments towards health care funding being related to ability to pay, not the use of medical care; access to health care being based on patients' need, not patients' ability to pay; and overall reduction in health inequalities.

An extensive empirical literature has been devoted to document inequalities in healthcare financing, access, and health (see Wagstaff and Doorslaer 2000 for a review). This chapter focuses on one form of inequalities in *access*. The principle that "access should be based on need" seems both intuitive and desirable. However, the words "access" and "need" are subject to different interpretations. "Access" can simply refer to healthcare utilization, i.e., whether a patient has received treatment or not. But it could also refer to the opportunity to receive treatment, when monetary and nonmonetary costs that people incur have been taken into account. Money costs involve any copayment the patient has to pay or monetary expenses to reach a healthcare provider (a patient from a rural area may, for example, face significant travel costs). Nonmonetary costs can take the form of waiting times, if the patient has to wait several weeks for

an elective treatment (e.g., a hip replacement) or a few hours in the emergency room. "Need" can be interpreted as ill health or severity, but also as the ability (or capacity) to benefit. The two concepts differ since ill patients may have low capacity to benefit from treatment (as for some cancer patients).

An extensive empirical literature has been devoted to test whether, controlling for need, individuals with different socioeconomic status differ in healthcare utilization (Wagstaff and Doorslaer 2000 for a review). In most studies, the level of healthcare utilization is measured by the number of visits to a specialist or a family doctor, while need is measured by self-reported health. Comparative international studies suggest that in many OECD countries there is generally pro-rich inequity for physician contacts, in particular in relation to specialist visit and to a lower extent family-doctors consultation (where in some instances pro-poor inequities may be present) (see van Doorslaer et al. 2000, 2004; Devaux 2015 for a recent analysis).

This chapter focuses on inequalities of access as measured by waiting times for nonemergency treatments. Studies of inequalities in waiting times by socioeconomic status are relatively infrequent. This is perhaps surprising given that waiting times are a major health policy issue in many OECD countries. Average waiting times can reach several months for common procedures like cataract and hip replacement (Siciliani et al. 2014). In the absence or limited use of prices in combination with constraints on the supply, publicly funded systems are often characterized by excess demand. Since the number of patients demanding treatment exceeds supply, patients are added to a waiting list and have to wait before receiving treatment (Martin and Smith 1999).

Waiting times generate dissatisfaction for patients since they postpone benefits from treatment, may induce a deterioration of the health status of the patient, prolong suffering, and generate uncertainty. A number of policies have been introduced across the globe to reduce or tackle waiting times (see Siciliani et al. 2013a for a review).

From an equity perspective, one possible advantage of rationing by waiting times is that

within publicly funded health systems, access to services is not supposed to depend on the ability to pay, unlike any form of price rationing where access is dependent on income. For a given level of need, the duration of the wait is supposed to be the same for patients with different income. Patients with higher need or urgency are supposed to wait less based on implicit or explicit prioritization rules.

A recent empirical literature, reviewed in this chapter, seems however to suggest that within several publicly funded health systems, nonprice rationing does not guarantee equality of access by socioeconomic status. Individuals with higher socioeconomic status (as measured by income or educational attainment) tend to wait less for publicly funded hospital care than those with lower socioeconomic status.

This negative gradient between waiting time and socioeconomic status may be interpreted as evidence of inequity within publicly funded systems which favors rich and more-educated patients over poorer and less-educated ones. Therefore, rationing by waiting times may be less equitable than it appears.

The chapter focuses on studies employing large samples either from administrative or survey data. The study is organized as follows. Possible sources of inequalities in waiting times are first discussed. Second, appropriate data and empirical methods are presented which can be usefully employed to investigate inequalities in waiting times. Third, the existing evidence is reviewed. Fourth, possible policy implications are drawn.

---

## Sources of Inequalities in Waiting Times

This section describes different mechanisms that generate inequalities in waiting times. Several health systems are publicly funded and characterized by universal health coverage (e.g., Australia, Italy, New Zealand, Spain, and the United Kingdom). These often coexist with a parallel private sector for patients who are willing to pay out of pocket or who are covered by private health insurance. A key feature of this private sector is that

waiting times are negligible: the promise of low wait is indeed the main way to attract patients from the public to the private sector. For several elective treatments, patients therefore can wait and obtain treatment in the public sector for free (or by paying a small copayment) or opt for the private sector and obtain care more swiftly if they are willing to pay the price (or prospectively insure themselves privately).

Since it is individuals with higher income that are more likely to be able affording private care, this generates inequalities in waiting times by socioeconomic status within a country. The extent of such inequalities due to the presence of the private sector is likely to depend on its relative size. For example, about 50 % of treatments are private in Australia, but these tend to be negligible in the Nordic countries where the option of going private is much more limited.

Within publicly funded systems, access to care should be based exclusively on need, not on ability to pay (in contrast to contributions to funding of health systems instead based on ability to pay, not need). Therefore, waiting times for patients on the list should reflect need and not on socioeconomic status. Indeed, patients on the list are prioritized by doctors. Patients with higher severity and urgency are supposed to wait less than less severe and urgent patients.

In practice, it is possible that variations in waiting times for publicly funded patients reflect also non-need factors. Waiting-time inequalities may be due to hospital geography and therefore arise “across” hospitals. This could be due to some hospitals having more capacity (number of beds and doctors) and being able to attract a more skilled workforce. This may be the case for hospitals located in an urban as opposed to a rural area. Also, some geographical areas may be underfunded compared to others. If individuals with higher socioeconomic status live in areas where hospitals are better funded or have higher capacity, then this may contribute to inequalities in waiting times by socioeconomic status.

Inequalities in waiting times may also arise “within” the hospital as opposed to across “hospitals.” Individuals with higher socioeconomic status may engage more actively with the health

system and exercise pressure when they experience long delays. They may be able to express better their needs. They may also have better social networks (know someone) and use them to have priority over other patients (attempt to jump the queue). They may have a lower probability of missing scheduled appointments (which would increase the waiting time). They may search more actively for hospitals with lower waiting times and willing to travel further.

---

## Data and Empirical Methods

### Data

#### Administrative and Survey Data

Two main data sources have been employed in the existing literature: administrative data and survey data. Each of these has relative merits. Registry data have also been employed but to a lower extent.

The key advantage of (mainly hospital) administrative data is that they cover the whole population of patients admitted to hospitals for treatment. Moreover, waiting times can be measured at disaggregated level, i.e., for specific conditions, treatments or surgeries (such as cataract or hip replacement) with large sample size. Administrative data contain detailed control variables on patients' severity (as proxied by number of comorbidities, number of secondary diagnoses or the Charlson index) and information on the hospital which provided the treatment.

The key disadvantage of administrative data is the difficulty in linking patients' wait with detailed and precise information on patients' socioeconomic status. Ideally, the researcher would like to access measures of income and educational attainment for each patient who was admitted to hospital. This would involve linking health administrative data with fiscal ones (generally for tax purposes). Except for Nordic countries, this link is not easily available in most countries. Researchers therefore have to use proxies. Since patient's postcode is usually available, the waiting time experienced by the patient can be linked with socioeconomic variables measured at small-area level (income deprivation, individuals living on benefits, proportion of

individuals with primary, secondary, and tertiary educational attainments, etc.).

The key advantage of survey data is that socioeconomic status (such as income and highest educational attainment) is routinely recorded at individual level. However, the sample tends to be smaller and more heterogeneous: patients' treatment can range from less urgent ones (e.g., a cataract surgery) to more urgent ones (e.g., cancer treatment). Detailed measures of severity are generally also missing. A measure of self-reported health tends to be used as a proxy of health needs which in line with previous literature on measuring social inequalities in healthcare utilization (Wagstaff and Van Doorslaer 2000).

#### Measures of Waiting Times

Waiting times in health care can be measured in different ways (Siciliani et al. 2013a, Chap. 2). For many elective surgical procedures and medical treatments, the most common measure is the *inpatient* waiting time. This measures the time elapsed from the specialist addition to the list to treatment for all publicly funded patients treated in a given year (Siciliani et al. 2014). Collecting data for all publicly funded patients implies that patients can receive treatment either by publicly or privately (nonprofit and for-profit) owned providers. Waiting times on privately funded patients are generally not collected on a routine basis.

The definition of inpatient waiting time does not include the *outpatient* waiting time, i.e., the time elapsed from the date of referral of the general practitioner to the date of specialist assessment.

Some countries (like Denmark and England) have started to collect a third measure known as *referral-to-treatment* waiting time. This measures the time elapsed between family doctor referral and treatment for patients treated in a given year. This measure therefore includes also the time elapsed from the family doctor referral to the specialist visit. It is approximately the sum of outpatient and inpatient waiting times, though it allows for gaps between the specialist visit and the addition to the list which could be significant.

An alternative measure to the inpatient waiting time of patients *treated* is the inpatient waiting

times (from specialist addition to the list) of *patients on the list* at a census date. This measure is analogous to the definition provided above but refers to the patients on the list at a given census date (as opposed to patients treated in a given year). Similarly, the referral-to-treatment waiting time of patients *on the list* can be defined.

The distribution of waiting time of patients treated measures the full duration of the patient's waiting time experience (from entering to exiting the list). The distribution of the waiting times of patients on the list refers to an incomplete duration since, if on the list, patients are still in the process of waiting. The waiting time of patients treated has the advantage of capturing the full duration of a patient's journey, but it is retrospective in nature. However, it does not capture the wait of the patients who never received treatment since they died while waiting, changed their mind, received a treatment in the private sector, etc. The two distributions of waiting times are different but related. Both distributions can be used to compute the probability of being treated (i.e., of waiting time ending) as time passes, i.e., the hazard rate in terms of survival analysis. The hazard rate derived under the two distributions will be the same if the system is in steady state and if each patient on the list is ultimately treated. Both conditions are unlikely to hold in reality. This emphasizes some of the differences between the two distributions (but see Armstrong 2000, 2002; Dixon and Siciliani 2009 for a fuller discussion of these issues).

Table 1 below provides comparative figures of median and mean waiting times across OECD countries in 2011. It illustrates how some countries report inpatient waiting time from specialist addition to the list to treatment, some report inpatient waiting time for patients on the list, and some report both measures. Among the countries included, waiting times appear lowest in Denmark and the Netherlands. It is also evident that mean waiting times are longer than the median ones, and this is due to the skewed distribution of waiting times with a small proportion of patients having a very long wait. As an example, Fig. 1 provides the distribution of waiting times for hip replacement for several OECD countries

(Australia, New Zealand, Portugal, Finland, and the United Kingdom).

It is important to emphasize that such measures of waiting times refer to elective (nonemergency) conditions where the wait is generally long (in the order of weeks or months) though they can be shorter for more urgent elective care (e.g., cancer care). Emergency care is therefore often excluded from the empirical analyses.

Most empirical analyses making use of administrative data surveyed in this chapter have employed data that measure the inpatient waiting time, which is computed retrospectively once the patient has received treatment. Those with survey data have included both the inpatient and the outpatient waiting time (for a specialist visit). Waiting-time measures from survey data are typically self-reported. Surveyed individuals are asked questions of the type "if you had an inpatient (outpatient) care in the last year, how long did you wait to be treated (to see a specialist)?" Answers may therefore suffer from recall bias.

## Methods

The empirical analyses are interested in testing whether patients with higher socioeconomic status wait less than patients with lower socioeconomic status when admitted to hospital. This section first presents a simple model specification which can be estimated with the Ordinary Least Square (OLS) method and then proceeds to more sophisticated models such as duration analysis.

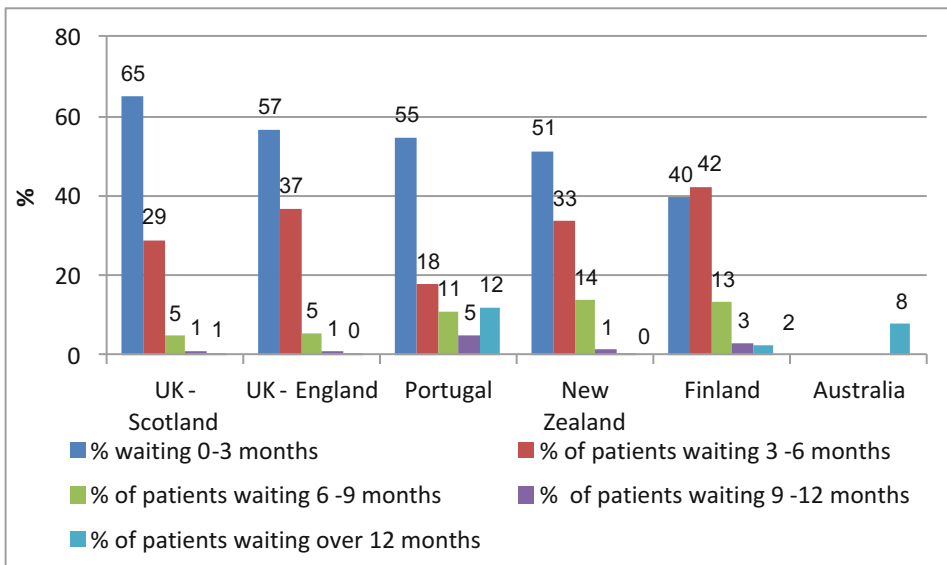
### Model Specification with Administrative Data

Suppose that the researcher has at her disposal a sample of  $I$  patients receiving treatment in  $J$  hospitals. The sample includes all patients who received a specific treatment (e.g., hip and knee replacement, cataract surgery, coronary bypass, varicose veins). Each patient receives treatment only in one hospital. Each hospital in the sample treats at least one patient. Define  $w$  as the inpatient waiting time for patients receiving treatment in a public hospital for treatment. It is assumed that waiting times are measured in days and that

**Table 1 Median (mean) waiting times for common surgical procedures: 2011**

<b>Patients treated – inpatient (time from specialist addition to list to treatment)</b>											
	Hip replacement	Knee replacement	Cataract	Hysterectomy	Prostatectomy	Cholecystectomy	Hernia	CABG	PTCA		
<b>Australia</b>	108	173	90	49	47	54	57	17			
<b>Canada</b>	89	107	49					7			
<b>Finland</b>	108 (125)	127 (141)	122 (125)	84 (98)	49 (72)	69 (90)	76 (96)	35 (45)	21 (31)		
<b>Netherlands</b>	(46)	(43)	(32)	(34)	(32)	(35)	(36)	(26)	(16)		
<b>New Zealand</b>	90 (104)	96 (112)	84 (94)	98 (109)	63 (86)	62 (86)	57 (82)	28 (37)	51 (66)		
<b>Portugal</b>	92 (149)	192 (231)	49 (67)	90 (125)	61 (115)	80 (134)	82 (120)	2 (29)			
<b>Spain</b>	(127)		(89)		(91)	(89)	(87)				
<b>UK-England</b>	81 (91)	85 (96)	57 (65)	61 (70)	31 (41)	70 (81)	60 (71)	52 (62)	35 (40)		
<b>UK-Scotland</b>	75 (90)	80 (94)	62 (70)	48 (53)	51 (55)	61 (77)	63 (82)	35 (47)	29 (33)		
<b>Patients treated – Referral to treatment (time from family doctor referral to treatment)</b>											
<b>Denmark</b>	39 (51)	46 (59)	70 (99)	35 (49)	36 (56)	38 (46)	45 (56)	13 (19)			
<b>Patients on the list – Inpatient</b>											
<b>Ireland</b>	103 (130)	119 (153)	118 (144)	96 (131)	81 (127)	93 (132)	98 (128)	77 (102)	54 (78)		
<b>New Zealand</b>	60 (78)	65 (84)	51 (63)	65 (73)	51 (66)	58 (75)	54 (69)	46 (60)	38 (51)		
<b>Portugal</b>	129 (189)	156 (200)	67 (100)	82 (111)	103 (185)	117 (178)	95 (147)	93 (118)			
<b>Spain</b>	(93)		(71)		(74)	(74)	(71)				
<b>Sweden</b>	43	45	40					25			
<b>Slovenia</b>	340 (354)	495 (512)	58 (63)			90 (122)	90 (132)		240 (275)		

Source: Siciliani et al. (2013b)



**Fig. 1** Distribution of waiting times of patients treated (Source: Siciliani et al. (2013b))

waiting time is a continuous variable. The following linear model can be specified:

$$w_{ij} = d_j\beta_j + y'_{ij}\beta_y + s'_{ij}\beta_s + e_{ij} \quad (1)$$

where  $w_{ij}$  is the waiting time of patient  $i$  in public hospital  $j$ . Waiting times are a function of (and additively separable in) the determinants outlined on the Right Hand Side of Eq. 1.

$s_{ij}$  is a vector of patients' characteristics capturing patients' severity. These could include age, gender, and number of comorbidities. These factors control for the severity of patient's health condition. In many countries, patients on the list are prioritized on the basis of their severity and more severe patients wait less relative to nonsevere ones. The coefficients  $\beta_s$  are therefore expected to be negative. They provide a measure of the extent to which patients with higher severity wait less.

$y_{ij}$  is a variable (or a vector of variables) which captures socioeconomic status, as measured by the income in the area where the patient lives. Inequalities in waiting time across patients with different socioeconomic status arise if  $\beta_y \neq 0$ . If  $\beta_y$  is negative then individuals with higher (lower) socioeconomic status wait less (more), keeping other variables (including severity) constant.

$d_j$  is a vector of hospital dummy variables (fixed effects), one for each hospital. These are included to control for systematic differences in waiting times across hospitals which arise from differences in supply (beds, doctors, efficiency) or in demand (e.g., proportion of the elderly). Hospitals with higher  $\beta_j$  have longer waiting times on average.

$e_{ij}$  is the idiosyncratic error term. This can be interpreted as any variation in waiting time which is not captured by the other variables (this includes coding and measurement error, or unobserved – to the researcher – dimensions of severity).

The simplest way to estimate Eq. 1 is with ordinary least squares (OLS). OLS minimizes the sum of the squared distances between the observed data and the predicted ones based on linear approximation, i.e., the sum of the squared of the errors (Cameron and Trivedi 2010, Chap. 3). OLS relies on a number of assumptions, including the exogeneity of the regressors, the error terms having the same variance (homoscedasticity) and conditionally uncorrelated observations. Under the assumption that the error terms are normally distributed, the hypothesis can be tested on whether the estimated coefficients are statistically different from zero.

For the coefficients  $\beta_y$  to provide an unbiased (correct) estimate of whether patients with higher socioeconomic status wait more or less than other patients, either socioeconomic status has to be uncorrelated with other determinants of waiting times (which seems implausible) or, if it is correlated, it has to be controlled for all possible determinants of waiting times. Otherwise, the estimates of  $\beta_y$  will be prone to so-called omitted variable bias.

For example, more severe patients are more likely to have lower socioeconomic status (Wagstaff and van Doorslaer 2000). Patients' severity may therefore be correlated negatively with both waiting time and socioeconomic status. Failure to control for patient severity might generate biased results. Without controlling for severity, a positive correlation between waiting time and income may be observed, while such correlation may disappear once controls for severity are added.

Similarly, hospitals with high supply (and lower waiting times) are likely to be located in urban areas where high-income patients are concentrated leading to a correlation between hospital characteristics and socioeconomic characteristics of patient's area of residence. Omitting hospital dummies (fixed effects) might overestimate inequalities. Including hospital fixed effects allows interpreting socioeconomic inequalities in waiting times "within" a hospital, rather than across hospitals. If researchers are interested in explaining waiting times inequalities across hospitals, a range of supply variables (e.g., number of beds and doctors, length of stay) can be employed instead of hospital fixed effects.

In summary, inequalities in waiting time across patients with different socioeconomic status arise if  $\beta_y \neq 0$ , i.e., when differences in waits are statistically significant even after controlling for patients' severity and hospital fixed effects.

Hypothesis testing requires the error terms to be normally distributed. Given that waiting times have a skewed distribution, the error terms in Eq. 1 are unlikely to be normal. To address this issue, the dependent variable  $w_{ij}$  is typically transformed by the logarithmic function, so that the dependent variable becomes  $\log(w_{ij})$ . If the

covariates (regressors) on the RHS of Eq. 1 are also in transformed in log, then each estimated OLS coefficient can be interpreted as elasticity. For example, if socioeconomic status is measured with income and  $\beta_y = -0.5$ , then a 10 % increase in income reduces waiting times by 5 %. If instead the covariates are dummy variables, then the estimated coefficient can be interpreted (approximately) as the proportionate change in waiting times (semielasticity). For example, suppose that socioeconomic status is measured through the highest level of education attained by the patient and patients either went to university or not. Suppose further that the estimated coefficient associated to the dummy variable (equal to one if the patient has a university degree) is equal to  $\beta_y = -0.1$ . Then, patients with a university degree wait 10 % less.

Estimating Eq. 1 by OLS treats hospital effects as fixed. This approach generates unbiased but inefficient estimates due to the inclusion of a large number of regressors (therefore introducing the possibility of not identifying a gradient when there is one). An alternative approach is to assume that hospital effects are random. Under the assumption that hospital effects are uncorrelated with other covariates, the coefficients in Eq. 1 will be estimated more efficiently. However, a random effect model will generate biased coefficients if hospital effects are correlated with other covariates. Whether the random effects generate different estimated coefficients compared to the fixed effects, can be tested through a Hausman test (Cameron and Trivedi 2010, Chap. 8).

### Model Specification with Survey Data

Studies that employ survey data have typically smaller samples. Investigating waiting times by treatment or procedure is often precluded. An analysis can still be conducted by pooling the sample across different treatments and conditions. In such studies, additional dummy variables have to be introduced to control for systematic differences in waiting times across conditions (e.g., waiting for a cataract surgery tends to be longer than for coronary bypass). Moreover, survey data rarely have information on the provider (e.g., the hospital) where the patient received the treatment.



It is therefore not possible to control for hospital fixed effects.

The model in Eq. 1 can be modified in the following way. Define again  $w$  as the inpatient or outpatient waiting time for patients who received treatment in a given year. The model specification is:

$$\ln(w_{ik}) = d'_k\beta_k + y'_i\beta_y + s'_i\beta_s + e_{ik} \quad (2)$$

where  $w_{ik}$  is the waiting time of patient  $i$  receiving treatment  $k$ ;  $y_i$  is a variable (or a vector of variables) which captures socioeconomic status (income and/or educational attainment),  $s_i$  measures, in addition to age and gender, self-reported health or whether the patient has chronic conditions.  $d_k$  is a vector of dummy variables controlling for different types of treatment (e.g., cataract, coronary bypass etc.) or speciality (orthopedic, ophthalmology).  $e_{ik}$  is the error term. Again, inequalities, in waiting time across patients with different socioeconomic status, arise if  $\beta_y \neq 0$ .

Depending on the survey employed, waiting time can be measured separately for publicly funded and privately funded patients. The availability of this information is critical. If public and private patients are pooled together, then an obvious reason for patients with higher socioeconomic status to wait less is that they can afford to go private. If only publicly funded patients are included in the analysis, then other mechanisms are responsible for the estimated gradient.

If waiting times are long and measured in days, they may be treated as a continuous variable (like in Eq. 1). However, if waiting times are short and/or measured in weeks or months, then waiting times should be treated as a discrete variable. Given that waiting times' distributions are skewed, a negative binomial model (NBM) can be employed to investigate the determinants of waiting times. The NBM gives a useful generalization of the Poisson model (PM), allowing for heterogeneity in the mean function, thereby relaxing the restriction on the variance (Cameron and Trivedi 2005; Jones 2007). In the PM, the dependent variable follows a Poisson distribution and the variance is set equal to the mean. The

NBM reduces to the PM in the special case when there is no overdispersion in the data.

If measured in weeks or months, waiting times data are discretized: the variable is observed discretely, whereas the underlying process generating waiting times is intrinsically continuous. An alternative to the NGM is the interval regression model which is specifically designed for discretized continuous variables.

### Duration Analysis

Duration models are an alternative method to investigate the determinants of waiting times. They can be employed to test for differences in waits between socioeconomic groups over the whole distribution of time waited (see Laudicella et al. 2012; and Dimakou et al. 2009; Appleby et al. 2005, for other applications of duration analysis to waiting times).

A key concept in duration analysis is the hazard rate,  $h(t)$ . This measures the instantaneous probability of leaving the waiting list at time  $t$  (and therefore of being treated) conditional on having waited on the list until time  $t$ .

A popular duration model is the Cox regression model. This model is semiparametric since it does not require assumptions over the distribution of the time waited. The Cox model identifies the effect of each covariate on waiting time in terms of hazard ratios, i.e., the ratio between the hazard rates of different groups of patients. The Cox model calculates the conditional hazard rate,  $h(t; x)$ , as:

$$h(t; x) = h_0(t) \exp\left(\sum_k \beta_k x_k\right) \quad (3)$$

where  $x_k = y, s$  (with  $k$  being the total number of covariates in each vector) and  $h_0(t)$  is the baseline hazard rate, i.e., the probability of leaving the list when all covariates are zero. The estimated coefficients  $\beta_k$  provide the effect of an increase in socioeconomic status and severity on the probability of leaving the waiting list, and therefore of being admitted for treatment. Suppose that socioeconomic status is measured by education with a dummy variable equal to one if the patient does not have a university degree. Then a coefficient

which is less than one will imply that less-educated patients have a lower probability of exiting the list (and therefore of being treated within a given time).

The Cox model assumes the hazard ratio between two different groups, for example, those treated in hospital  $j$  and hospital  $j'$ ,  $\exp\left[\sum_k \beta_k (x_j - x_{j'})_k\right]$  is constant with time waited (Cameron and Trivedi 2005, Chap. 17.8). If this assumption is violated, then the *stratified* Cox model and the *extended* Cox model may be more appropriate. The former introduces group-specific baseline hazards,  $h_{0j}(t)$ . Therefore, the conditional hazard rate becomes:

$$h(t; x) = h_{0j}(t) \exp\left(\sum_k \beta_k x_k\right).$$

The main advantage of the stratified Cox model is that it relaxes the common baseline hazard assumption. The main disadvantage is that hazard ratios between the stratified groups cannot be identified. The *extended* Cox model introduces time dependency by interacting covariates with the time waited,  $g_k(t)$ , (Pettitt and Daud 1990; Fisher and Lin 1999):

$$h(t; x(t)) = h_0(t) \exp\left[\sum_k \beta_k x_k + \sum_k \delta_k x_k g_k(t)\right] \quad (4)$$

where  $\delta_k$  are the coefficients of the time interactions.

### Other Methods

Another useful regression method for investigating waiting times is *quantile regression* (Cameron and Trivedi 2010, Chap. 7). Estimating Eq. 1 by OLS allows estimating the effect of socioeconomic status at the sample mean. Since patients differ in the degree of urgency, it may be interesting to estimate whether such effect is persistent also when waiting times are high or low, i.e., across different cut-off points in the waiting time distributions (say the 20th and 80th percentile, and at the median) through a quantile regression model. Doctor prioritize

patients based on their degree of urgency. Some dimensions of urgency may however remain unobservable to the researcher. Whether a larger socioeconomic gradient should be expected at low or high waiting times is in principle indeterminate. Since waiting times are short when the condition is more urgent, richer and more-educated people may be keener to obtain reductions in waiting times when they perceive delays to affect their health more critically. On the other hand precisely because waiting times are short, there may be less scope for influencing them.

Finally, a concern may be raised that estimates in Eq. 1 are contaminated by what is known as *sample selection* based on unobserved factors (to the researcher). For example, patients with higher income who expect to wait a long time are more likely to afford and opt for the private sector. It may therefore arise that public hospitals treat poor patients with expected high and low waiting times but only rich patients with low waiting times. In turn, this may generate an apparent negative gradient between income and waiting time for patients receiving treatment within publicly funded hospitals. If the researcher observes whether patients went for public and private treatment, then a Heckman Selection model can be performed to adjust for sample-selection bias (Heckman 1979). Such model involves estimating a selection equation for the choice of the patient between opting for private care versus public care, which can include socioeconomic status among its determinants. For the model to perform well, an identification variable is recommended, i.e., a variable which predicts the choice of going public versus private but does not directly affect waiting times (distance to the hospital may be such an identifying variable; see Sharma et al. 2013).

## A Review of the Evidence

This section first reviews key results from international studies and then on studies that focus on individual countries.

## International Studies: Evidence from SHARE and the Commonwealth Fund

Using survey data from Survey of Health, Ageing and Retirement in Europe (SHARE), Siciliani and Verzulli (2009) test whether waiting times for specialist consultation and nonemergency surgery differ by socioeconomic status. The sample includes nine European countries: Austria, Denmark, France, Germany, Greece, Italy, the Netherlands, Spain and Sweden. The survey covers 22,000 respondents across these European countries. The analysis controls for severity as proxied by age, gender, and self-reported health (and type of specialist care and treatment). Privately funded patients are excluded from the analysis (a minority of the sample). Therefore, the analysis can be interpreted in terms of inequalities among publicly funded patients. Since waiting times are measured in weeks and months, a negative binomial model is employed.

For specialist consultation, they find that individuals with high education experience a reduction in waiting times of 68 % in Spain, 67 % in Italy and 34 % in France (compared with individuals with low education). Individuals with intermediate education report a waiting-time reduction of 74 % in Greece (compared with individuals with low education). There is also evidence of a negative and significant association between education and waiting times for nonemergency surgery in Denmark, the Netherlands, and Sweden. High education reduces waits by 66 %, 32 %, and 48 %, respectively. There is some evidence of income effects, although generally modest. An increase in income of 10,000 Euro reduces waiting times for specialist consultation by 8 % in Germany and waiting times for nonemergency surgery by 26 % in Greece. Surprisingly, an increase in income of 10,000 Euro increases waits by 11 % in Sweden.

Schoen et al. (2010) use data from the 13th annual health policy survey conducted in 2010 by the Commonwealth Fund in eleven countries (Australia, Canada, France, Germany, New Zealand, the Netherlands, Norway, Sweden, Switzerland, the United Kingdom, and the

United States). Waiting times are measured for a specialist visit and for elective surgery. Socioeconomic status is proxied by a dummy variable equal to one if income is above average. Control variables include age, health status, and for the USA for private insurance status.

Employing logistic regression, the study shows that individuals with above-average income have a lower probability of waiting more than 2 months for a specialist visit in Australia, New Zealand, and the Netherlands. They also have a higher probability of waiting less than 4 weeks for a specialist visit in Australia, Canada, New Zealand, and the United States. No marked differences in waiting times by socioeconomic status are found for elective surgery. Since no control variable is included for patients going to private provider, differences in waiting times by socioeconomic status could to some extent be explained by richer patients opting for the private sector when waiting times are high.

## United Kingdom

Using administrative data, Cooper et al. (2009) investigate for the presence of inpatient waiting-time inequalities in England for the following elective procedures: hip and knee replacement and cataract surgery. They also compare whether such inequalities varied during the Labor government between 1997 and 2007. Waiting time was much higher in the early years but then gradually fell. The analysis refers to publicly funded patients only, i.e., patients treated by the National Health Service. Patients who do not want to wait can opt for treatment in the private sector, but they will have to pay or hold a private health insurance.

The regression analysis (similar to Eq. 1) controls for patients' age, gender, area type (e.g., city, town and fringe, isolated village), but not for hospital fixed effects. The regressions are run for three periods corresponding to different government policy (1997–2000, 2001–2004, and 2005–2007). Socioeconomic status was measured through an index of income deprivation (the 2001 Carstairs index at the output area level then

transformed in to five income deprivation quintiles). The Carstairs index is based on car ownership, unemployment, overcrowding, and social class within output areas, calculated by the Office of National Statistics.

The study finds that compared to patients with lowest income deprivation (highest socioeconomic status) patients in other groups tend to wait longer, up to about 2 weeks longer. For some procedures and years, the effect is not-monotonic with patients with middle-income deprivation waiting longest. Inequalities in waiting times tend to decrease over time. This is probably due to waiting times falling over the considered period. The authors conclude that equity improved over time. In the period 2005–2007, very little differences existed in waiting times across patients with differing deprivation.

The analysis by Cooper et al. (2009) does not account for hospital fixed effects. Therefore, inequalities in waiting times may reflect variations “across hospitals” due, for example, to different resources or variations “within the hospital” due, for example, to some patients being able to get ahead of the queue. Laudicella et al. (2012) extends the analysis by introducing hospital fixed effects but focuses on hip replacement only in 2001. They split the deprivation index between “income” deprivation (based on individuals on benefits) and “education” deprivation. They provide evidence of inequalities in waiting times favoring more-educated and richer individuals. More precisely, a patient who is least skill deprived in education wait 9–14 % less than other patients; patients in the fourth and fifth most income-deprived quintile wait about 7 % longer than other patients. The analysis provides evidence that most inequalities occur within hospitals rather than across hospitals (failure to control for hospital fixed effects results in underestimation of the income gradient). The key insights are similar when the Cox nonparametric model is employed. More educated patients have a higher probability of leaving the list (the inverse of hazard ration) by 2–6 %. Richer patients have a higher probability of leaving the list by 4–9 %.

Pell et al. (2000) investigate inequalities in waiting times for cardiac surgery in Scotland. They employ administrative data measuring the inpatient waiting time. Similarly to Cooper et al. (2009), socioeconomic status is proxied through the Carstairs deprivation index. They find that the most deprived patients waited 24 days longer than least deprived ones. This was in part due to less deprived patients more likely to be classified as urgent.

## Australia

Sharma et al. (2013) investigate the presence of inequalities in waiting times in the State of Victoria (which accounts for 25 % of Australian population) in 2005. The study employs administrative data on inpatient waiting time for publicly funded patients. Several surgical procedures are employed (including eye, hip and knee procedures, hysterectomy, and prostatectomy).

A key institutional feature of the Australia system is that although everyone has public insurance, about half of the population has private health insurance and about half of the care is provided by private hospitals. More precisely, patients who seek treatment in a public hospital receive treatment for free under Medicare (Australia’s universal public health insurance scheme) but have to wait. Patients who seek treatment in a private hospital incur the full cost of treatment, which is paid by the patient either directly or through her private health insurer.

Given such institutional feature, one explanation for a potential observed gradient between waiting time and socioeconomic status for publicly funded patients is the possibility of sample selection: rich patients who expect to wait are more likely to afford and opt for the private sector generating a negative gradient between income and waiting time in the public system. In other words, public hospitals treat poor patients with expected high and low waiting times, but only rich patients with low waiting times are treated in public hospitals. This is of potential importance for policy. If the gradient is explained by sample

selection, then it should not be interpreted as evidence of inequity.

Since private hospitals have to report the same data than public hospitals, detailed administrative data are available for both public and private sector (unlike many other countries). These data are therefore suitable for testing for sample selection generated by the private sector through a Heckman sample-selection model (the distance to the nearest public and private hospitals are used as identifying variable).

Like the English studies, socioeconomic status is measured through an index which captures economic resources at small-area level (suburbs), known as the SEIFA (Socio-Economic Indexes for Areas) for economic resources. Examples of variables which generate the SEIFA index for are the proportion of: people with low-income, single-parent families, occupied private housing with no car, households renting from community organization, unemployed, and households owning a house they occupy.

The analysis suggests that individuals who live in richer areas wait less. Compared to patients living in areas with lowest income, patients living in areas with highest income wait 13 % less. With an average waiting of 89 days, this implies an average reduction of 11 days. Patients in almost every decile of income have a progressively lower waiting time than the one below. Once selection is taken into account, the gradient between waiting times and socioeconomic status reduces significantly in size but does not disappear. Compared to patients in the lowest income decile, patients whose income falls between the 2nd and 7th deciles wait 3–4 % less, and patients whose income falls between the 8th and 10th deciles wait 5–7 % less. Therefore, the analysis still suggests evidence of inequity though a reduced one compared to the case when selection is not taken into account. The results from quantile regression models confirm that inequalities persist at different points of the waiting time distribution.

Johar et al. (2013) use administrative data from New South Wales in Australia to decompose variations in waiting times that are due to clinical need, supply factors, and nonclinical factors such as socioeconomic status. They measure

inpatient waiting times for publicly funded patients in public hospitals in 2004–2005 and include all acute illnesses. Socioeconomic status is measured through the SEIFA index (mentioned above) split into five groups. Without controlling for supply factors, they find that more deprived patients wait 30 % longer than those in the least deprived group (they wait about a month more with an average wait of about 3 months across all patients included in the sample). These differences reflect inequalities both within and across hospitals.

Once the authors control for supply factors (such as bed occupancy rate, length of stay, ratio of clinical staff to beds, proportion of emergency admissions), then patients wait 16–24 % longer compared to patients in the highest socioeconomic group. This implies that richer patients live in areas with better supply of hospital services. However, inequalities within the hospital persist after controlling for supply factors. Quantile regression results confirm that inequalities are present at all quantiles of the waiting time distribution.

## Norway

Monstad et al. (2014) use data from the Norwegian Arthroplasty Register for patients in need of hip replacement in Norway in 2002–2003 to test whether patients with higher socioeconomic status wait less. Income and education are measured at individual level. The sample covers 98 % of all hip replacements. Since every patient has a unique personal identification code, then the registry data can be perfectly matched with other registers at Statistics Norway.

The healthcare system in Norway is largely publicly funded with a negligible private sector (therefore, the possibility to opt out is limited). Waiting times for hip replacement were on average 170 days. The analysis is presented separately for men and women. All specifications control for hospital fixed effects. Therefore, results can be interpreted as inequalities arising “within the hospital.” The study finds that richer men and more-educated women tend to wait less: a 10 %

increase in income reduces waiting times by 8 %; women with 3 years of upper secondary education wait 7 % less compared to those with compulsory schooling only.

Carlsen and Kaarboe (2014) use administrative data (the Norwegian patient registry) from all elective inpatient and outpatient hospital stays in Norway for 2004–2005. The waiting time is measured from the referral (from family doctor) until the patient meets with a hospital specialist. Socio-economic status is measured at small-area level (about 31,000 cells). Since the register contains information about hospital stay, gender, year of birth, and resident municipality, patients can be uniquely assigned to population cells that combine gender, age, and municipality. For each population cell, Statistics Norway computed a set of variables that describe the income and educational levels of the cell population in 2004.

The study finds that men with tertiary education wait about 15 % less than men with primary education only. Women in the lowest income quintile wait 11 % longer than women with highest income quintile. However, once controls are added for hospital-specific factors (whether they went to the local hospital, travel time, and choice of hospital), most of inequalities disappear. Whether the patient goes to the “local hospitals” and travel distance are key factors explaining the gradient. Since hospitals in low-income regions have longer waiting time than hospitals located in high-income and middle-income regions, controlling for local hospitals makes the income gradient flatter. Travel distance also weakens the association between income and waiting time. Patients’ income decreases in traveling distance, whereas waiting time increases with distance.

## Sweden

Tinghög et al. (2014) use administrative hospital data on all elective surgeries performed in Östergötland in Sweden in 2007. These data were linked to national registers containing variables on socioeconomic variables. The study finds that patients with low disposable household income have 27 % longer waiting times for

orthopedic surgery and 34 % longer waiting times for general surgery. No differences on the basis of ethnicity and gender were found. Income mattered more at the upper tail of the waiting time distribution.

## Canada

Alter et al. (1999) employ a large administrative dataset to investigate whether publicly funded waiting times for patients in need of a coronary angiography in 1993–1997 in Ontario (Canada) differ for by socioeconomic status. The latter is proxied by neighborhood income as determined by the Canadian census. The study controlled for a number of supply factors such as the hospital volume, distance from hospital, type of hospital, in addition to clinical ones capturing patients’ severity. The study finds that patients in the highest income quintile wait 45 % less compared to patients in the lowest income quintile.

Carrière and Sanmartin (2007) investigate determinants of waiting times for specialist consultation using the 2007 Canadian Community Health Survey. Like other surveys, the analysis does not control for hospital variables. On the other hand, socioeconomic status (household income and educational attainment) is measured at individual level. The key finding is that compared with men in the top income quintile, those in the lowest were less likely to see a specialist within a month (after controlling for possible confounders). This was not the case for women.

## Germany

Using survey data between 2007 and 2009, Roll et al. (2012) investigate the impact of income and type of insurance on waiting times to see a family doctor and a specialist. Type of insurance is a critical control variable since Germany has a multipayer health system divided into two main components: statutory health insurance and private health insurance. While the first is financed by income-related contribution rates, private insurance is financed by risk-based rates. The

vast majority of the population is covered by statutory insurance. However, individuals with an income of approximately 50,000 Euro in 2011 can opt out to take private insurance which covers about 11 % of the population.

After controlling for insurance type, mild or severe severity, chronic conditions, and type of care needed, the study finds that income reduces waiting time for both an appointment with the GP and the specialist. Individuals with a household income with more than 2,000 Euro per month were associated with a reduction in waiting time for a GP appointment by 1 day or 28 % compared to respondents with an income of less than 500 Euro (with a sample mean of about 3 days). For the waiting time of an appointment with the specialist, a household income of more than 5,000 Euro per month was associated with significantly lower waiting time (28 % or 5 days less; sample mean of 30 days). Individuals with private insurance also obtain faster access to health services.

## Spain

Abasolo et al. (2014) use the 2006 Spanish National Health Survey to test for waiting time inequalities. The Spanish health system is characterized by universal coverage and tax funding. Waiting time is measured for the last specialist visit and is measured separately for a first visit and for a review visit. Like other studies employing survey data, household income and education are measured at individual level. Only public patients are included in the analysis. Public patients have no or limited copayments for specialist services. Average waiting time was about 2 months.

The analysis controls for type of speciality, self-assessed health, existing conditions (such as hypertension and heart problems), whether the patient has private insurance, employment status, living in a rural area, different regions, in addition to demographic variables. The study finds that an increase of 10 % of the income reduces waiting times for diagnosis visits in 2.6 %. Individuals with primary education wait 28 % longer than individuals with university studies.

## Italy

Petrelli et al. (2012) employ administrative data in Piedmont (a large Italian Region) in 2006–2008 to investigate inequalities in waiting times for selected surgical procedures, such as coronary bypass, angioplasty, coronarography, endarterectomy, hip replacement, and cholecystectomy. Waiting time is measured for publicly funded patients. It refers to the inpatient wait, from the specialist addition to the list to admission for treatment. Socioeconomic status was measured by education only, not income.

The Italian health system has universal coverage with limited or no copayments for inpatient hospital care. The analysis controls for severity (as proxied by the Charlson index) in addition to demographic variables. The results from Cox regression suggest that more-educated patients are more likely to wait less for all procedures except for coronary bypass (where the difference is not statistically significant).

---

## Conclusions and Implications for Policy

Within publicly funded systems, access to services is supposed to depend on need and not ability to pay (or, more broadly, socioeconomic status). The recent empirical literature reviewed in this chapter seems however to suggest that this is not necessarily the case. The chapter focuses on elective (i.e., nonemergency) services and does not cover the literature on waiting times in the emergency room. There is empirical evidence from several countries, suggesting that individuals with higher socioeconomic status (as measured by income or educational attainment) tend to wait less for publicly funded hospital elective services than those with lower socioeconomic status. Combined with the empirical literature reviewed in the Introduction, it suggests that not only individuals with higher socioeconomic status tend to see doctors more frequently, but also more swiftly.

Waiting-time inequalities within public systems may be due to a number of different reasons. They may be due to hospital geography with some

hospitals having more capacity and being located in more affluent areas. Inequalities in waiting times may also arise “within” the hospital if individuals with higher socioeconomic status engage more actively with the health system, exercise pressure when they experience long delays, are able to express better their needs, have better social networks (attempt to jump the queue), miss scheduled appointments less frequently, and are willing to travel further in the search of lower waits.

Although there is significant evidence on social inequalities in waiting times, it is still not known which of its possible determinants are the most critical. The methods and data outlined in this chapter could be usefully employed in future research to further uncover evidence on the presence of such inequalities in a number of countries, and perhaps most importantly, its key determinants. The degree to which these inequalities are unjust depends on its exact mechanisms, for example, whether richer patients exercise more active choice among public providers (a policy which is encouraged in many countries) or whether through more deliberate attempts to jump the queue. Therefore, rationing by waiting times may be less equitable than it appears.

In some countries, universal health coverage coexists with a parallel private sector for patients who are willing to pay out of pocket or who are covered by private health insurance. Individuals with higher income are more likely to be able affording private care, generating inequalities in waiting times by socioeconomic status within a country. In such circumstances, it is much less surprising that such inequalities exist.

Uncovering the exact mechanisms that explain the socioeconomic gradient in waiting times is also critical for policy design. For example, if the gradient is due to hospitals having access to different resources, then policymakers may want to improve allocation formulas that appropriately reflect need. If instead the gradient arises within the hospital with some patients attempting to jump the queue, more robust mechanisms to regulate the waiting list management may be required. If poorer people are struggling to keep up with the health booking systems, then simplifications and greater transparency could be considered.

## References

- Abasolo I, Negrin-Hernandez MA, Pinilla J. Equity in specialist waiting times by socioeconomic groups: evidence from Spain. *Eur J Health Econ*. 2014;15:323–34.
- Alter DA, Naylor CD, Austin P, Tu JV. Effects of socioeconomic status on access to invasive cardiac procedures and on mortality after acute myocardial infarction. *N Engl J Med*. 1999;348(18):1359–67.
- Appleby J, Boyle S, Devlin N, Harley M, Harrison A, Thorlby R. Do English NHS waiting time targets distort treatment priorities in orthopaedic surgery? *J Health Serv Res Policy*. 2005;10(3):167–72.
- Armstrong PW. First steps in analysing NHS waiting times: avoiding the ‘stationary and closed population’ fallacy. *Stat Med*. 2000;19(15):2037–51.
- Armstrong PW. The ebb and flow of the NHS waiting list: how do recruitment and admission affect event-based measures of the length of ‘time-to admission’? *Stat Med*. 2002;21:2991–3009.
- Cameron CA, Trivedi PK. *Microeconometrics: methods and applications*. Cambridge: Cambridge University Press; 2005.
- Cameron CA, Trivedi PK. *Microeconometrics using Stata*. Rev. ed. College Station: Stata Press. 2010.
- Carlsen F, Kaarboe O. Waiting times and socioeconomic status. Evidence from Norway. *Health Econ*. 2014;23: 93–107.
- Carrière G, Sanmartin C. Waiting time for medical specialist consultations in Canada. 2007. Statistics Canada, Catalogue no. 82-003-XPE. *Health Rep*. 2010;21 (2):7–14.
- Cooper ZN, McGuire A, Jones S, Le Grand J. Equity, waiting times, and NHS reforms: retrospective study. *Br Med J*. 2009;339:b3264.
- Devaux M. Income-related inequalities and inequities in health care services utilisation in 18 selected OECD countries. *Eur J Health Econ*. 2015;16(1):21–33.
- Dimakou S, Parkin D, Devlin N, Appleby J. Identifying the impact of government targets on waiting times in the NHS. *Health Care Manag Sci*. 2009;12(1):1–10.
- Dixon H, Siciliani L. Waiting-time targets in the healthcare sector. How long are we waiting? *J Health Econ*. 2009;28:1081–98.
- Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional hazards regression model. *Annu Rev Public Health*. 1999;20:145–57.
- Heckman JJ. Sample selection bias as a specification error. *Econometrica*. 1979;47(1):153–61.
- Johar M, Jones G, Keane M, Savage E, Stavrunova O. Differences in waiting times for elective admissions in NSW public hospitals: a decomposition analysis by non-clinical factors. *J Health Econ*. 2013;32:181–94.
- Jones AM. *Applied econometrics for health economists: a practical guide*. Oxford: Radcliffe Medical Publishing; 2007.
- Laudicella M, Siciliani L, Cookson R. Waiting times and socioeconomic status: evidence from England. *Soc Sci Med*. 2012;74(9):1331–41.



- Martin S, Smith PC. Rationing by waiting lists: an empirical investigation. *J Public Econ*. 1999;71:141–64.
- Monstad K, Engeaeter LB, Espehaug B. Waiting time socioeconomic status – an individual level analysis. *Health Econ*. 2014;23:446–61.
- Pell J, Pell A, Norrie J, Ford I, Cobbe S. Effect of socioeconomic deprivation on waiting time for cardiac surgery: retrospective cohort study. *Br Med J*. 2000;321:15–8.
- Petrelli A, De Luca G, Landriscina T, Costa G. Socioeconomic differences in waiting times for elective surgery: a population-based retrospective study. *BMC Health Serv Res*. 2012;12:268.
- Pettitt AN, Daud IB. Investigating time dependence in Cox's proportional hazards model. *J R Stat Soc. Ser C (Appl Stat)*. 1990;39(3):313–29.
- Roll K, Stargardt T, Schreyogg J. Effect of type of insurance and income on waiting time for outpatient care, the Geneva papers. *Int Assoc Study Insur Econ*. 2012;37:609–32.
- Schoen C, Osborn R, Squires D, Doty MM, Pierson R, Applebaum S. How health insurance design affects access to care and costs, by income, in eleven countries. *Health Aff*. 2010;29(12):2323–34.
- Sharma A, Siciliani L, Harris A. Waiting times and socioeconomic status: does sample selection matter? *Econ Model*. 2013;33:659–67.
- Siciliani L, Verzulli R. Waiting times and socioeconomic status among elderly Europeans: evidence from SHARE. *Health Econ*. 2009;18(11):1295–306.
- Siciliani L, Borowitz M, Moran V, editors. *Waiting time policies in the health sector. What works?* Paris: OECD Book; 2013a.
- Siciliani L, Moran V, Borowitz M. Measuring and comparing health care waiting times in OECD countries. *OECD health working papers*, 67. OECD Publishing; 2013b. <https://doi.org/10.1787/5k3w9t84b2kf-en>.
- Siciliani L, Moran V, Borowitz M. Measuring and comparing health care waiting times in OECD countries. *Health Policy*. 2014;118(3):292–303.
- Tinghög G, Andersson D, Tinghög P, Lyttkens CH. Horizontal inequality when rationing by waiting lists. *Int J Health Serv*. 2014;44(1):169–84.
- van Doorslaer E, Wagstaff A, et al. Equity in the delivery of health care in Europe and the US. *J Health Econ*. 2000;19(5):553–83.
- Van Doorslaer E, Koolman X, Jones AM. Explaining income-related inequalities in doctor utilization in Europe. *Health Econ*. 2004;13(7):629–47.
- Wagstaff A, van Doorslaer E. Equity in health care financing and delivery. Chapter 34. In: Culyer AJ, Newhouse JP, editors. *Handbook of health economics*, vol. 1. 1st ed. Amsterdam: Elsevier Science/North-Holland; 2000. p. 1803–62.



# Health Services Data: The Ontario Cancer Registry (a Unique, Linked, and Automated Population-Based Registry)

# 16

Sujohn Prodhan, Mary Jane King, Prithwish De, and Julie Gilbert

## Contents

<b>Introduction</b> .....	365
History of Cancer Registration in Ontario .....	366
Automation and OCRIS .....	366
EDW Reconstruction .....	367
<b>Who Uses OCR Data and for What Purpose?</b> .....	367
Examples of Provincial Stakeholders .....	367
Examples of National Stakeholders .....	367
Examples of International Stakeholders .....	368
<b>Data Sources</b> .....	369
Pathology .....	369
Activity Level Reporting .....	371
DAD and NACRS .....	371
Death Certificates .....	371
<b>Data Systems and Consolidation</b> .....	371
OCRIS and the EDW Successor .....	371
Patient Linkage .....	371
Case Resolution .....	372
<b>Data Elements</b> .....	373
<b>Data Quality</b> .....	374
Other Factors Affecting Quality .....	375
<b>The OCR Adopts a New Approach to Counting Cancers</b> .....	376
Topography and Morphology .....	376
Laterality .....	377
Timing .....	377

S. Prodhan (✉) · M. J. King · P. De (✉)  
Surveillance and Ontario Cancer Registry, Cancer Care  
Ontario, Toronto, ON, Canada  
e-mail: [prithwish.de@cancercare.on.ca](mailto:prithwish.de@cancercare.on.ca)

J. Gilbert  
Planning and Regional Programs, Cancer Care Ontario,  
Toronto, ON, Canada

Implications of Counting Rules on Data and Analysis ..... 378  
 Best Practices for Analysis ..... 378

**Cancer Stage at Diagnosis** ..... 379  
 CS Automation and Integration ..... 379  
 Source of Staging Data ..... 380  
 Stage Capture Rates ..... 380

**Linkage of the OCR to Other Datasets** ..... 381  
 Other Linkage Processes ..... 381  
 CCO’s Other Data Holdings ..... 382

**Health Services Research Using the OCR** ..... 382  
 Examples of Health Services Research Using the OCR ..... 382  
 Patient Contact ..... 386

**Data Privacy and Access** ..... 386  
 Privacy ..... 386  
 Data Request Process ..... 387

**Technical Appendix** ..... 387  
 ePath, eMaRC, and ASTAIRE ..... 387

**References** ..... 389

**Abstract**

Since its creation in 1964, the Ontario Cancer Registry (OCR) has been an important source of high-quality information on cancer incidence and mortality. As a population-based registry, the OCR can be used to assess the provincial burden of cancer, track the progress of cancer control programs, identify health disparities among subpopulations, plan and improve healthcare, perform health services research, verify clinical guideline adherence, evaluate screening effectiveness, and much more. With over one third of Canadians residing in Ontario, the OCR is the nation’s largest provincial cancer registry and a major contributor to the Canadian Cancer Registry. In 2015 alone, the OCR collected data on an estimated 83,000 malignant cases.

Through its active participation in Canadian, North American, and international standard setting bodies, the OCR adopts the latest methods for registry data collection and reporting. The OCR is created entirely from records generated for purposes other than cancer registration. These records include pathology reports, treatment-level activity, hospital discharges, surgery data, and death certificates. Electronic records are

linked at the person level and then “resolved” into incident cases of cancer using a unique computerized medical logic. Recent technological updates to the OCR have further modernized the registry and prepared it for future developments in the field of cancer registration.

This chapter describes the evolution of the OCR, its basic processes and components of automation, data elements, data quality measures, linkage processes, and other aspects of the registry that make it of particular interest to health services researchers and more broadly to the healthcare and public health community.

**List of Abbreviations**

AJCC	American Joint Committee on Cancer
ALR	Activity Level Reporting
CCO	Cancer Care Ontario
CIHI	Canadian Institute for Health Information
CS	Collaborative Stage
DAD	CIHI’s Discharge Abstract Database
DCO	Death certificate only
DSA	Data sharing agreement
eCC	Electronic Cancer Checklist

EDW	Enterprise Data Warehouse
EDW-OCR	Enterprise Data Warehouse based OCR
eMaRC	Electronic Mapping, Reporting, and Coding Plus
ePath	Electronic pathology data collection system
IACR	International Association of Cancer Registries
IARC	International Agency for Research on Cancer
ICBP	International Cancer Benchmarking Partnership
ICD	International Classification of Diseases
ICD-O	International Classification of Diseases for Oncology
MPH	Multiple Primary and Histology
NAACCR	North American Association of Central Cancer Registries
NACRS	CIHI's National Ambulatory Care Reporting System
OCR	Ontario Cancer Registry
OCRIS	Ontario Cancer Registry Information System
OCTRF	Ontario Cancer Treatment and Research Foundation
RCC	Regional Cancer Center
SEER	Surveillance, Epidemiology, and End Results program
SSF	Site-specific factors
TNM	Tumor Node Metastasis staging

(the Ministry) and its advisor on the cancer and renal systems, as well as on access to care for key health services. CCO strives for continuous improvement in disease prevention, screening, the delivery of care, and the patient experience. CCO works with Regional Cancer Programs across the province, cancer experts, community advisory committees, hospitals, provincial agencies and government, public health units, the Ontario Hospital Association, the not for profit sector, as well as with cancer agencies in other provinces and the federal government, among others, in order to achieve its mandate. Authority for CCO's programs and functions are provided in the provincial Cancer Act, the Personal Health Information Protection Act (PHIPA 2016), and a Memorandum of Understanding between the Ministry and CCO (Cancer Act 2006).

In accordance with Ontario's PHIPA legislation, CCO is defined as a "prescribed entity" for certain functions. This designation authorizes CCO to collect, use, and disclose personal health information for the purposes of cancer management and planning. The OCR is a prescribed entity to support this goal. The OCR team at CCO is comprised of pathology coders, standards advisors, stage abstractors, quality assurance and data analysts, and a management team. The OCR team's responsibilities include:

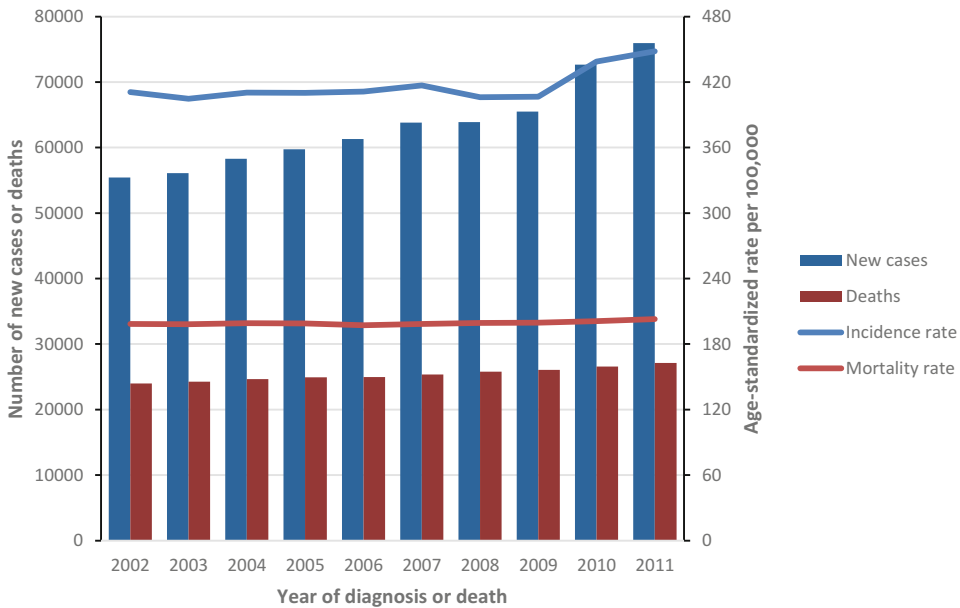
- Curating and coding source data to identify incident cancer cases
- Deriving population-level cancer staging values
- Working with standard setting bodies to establish the best practices for the registry
- Setting direction for the management of the registry
- Collaborating with partners and stakeholders to enable use of OCR data for surveillance and research, and more generally for cancer prevention and control

## Introduction

The purpose of this chapter is to describe the evolution of the Ontario Cancer Registry (OCR), and explore its many purposes, processes and applications that make it of particular interest to researchers. The chapter also emphasizes how the registry has established itself as an effective population-based surveillance and research tool.

The OCR is the official provincial cancer incidence registry for Ontario and is managed by Cancer Care Ontario (CCO). CCO is an agency of the Ontario Ministry of Health and Long-Term Care

The goal of the registry is to collect and disseminate timely and high-quality information describing all cases of cancer diagnosed among Ontario residents using measures of cancer



**Fig. 1** Cancer incidence and mortality counts and age-standardized rates per 100,000 (adjusted to the 1991 Canadian Standard population) from 2002 to 2011. The

sharp rise in incidence from 2010 onward is attributed to the adoption by the OCR of new counting rules for multiple primary cancers (*Source: Ontario Cancer Registry, 2014*)

burden such as incidence and mortality. The OCR is the largest provincial cancer registry in Canada, covering a population that comprises almost 40 % of the Canadian population. With Ontario's growing and aging population, the OCR is expected to have collected information on a projected 83,000 new cases of cancer in 2015 (Fig. 1).

### History of Cancer Registration in Ontario

Recognition of the importance of population-based cancer registration in Ontario goes back to 1943, with the passing of the provincial *Cancer Act* and the establishment of the Ontario Cancer Treatment and Research Foundation (OCTRF). The OCTRF, which became Cancer Care Ontario in 1997, was established to “conduct a program of research, diagnosis and treatment in cancer” (Cancer Act 2006). While the *Cancer Act* did not make the reporting of cancer diagnoses a legal obligation, it permitted

organizations and individuals in the healthcare system to provide such information to the OCTRF. This led to the formation of Ontario's cancer registry in 1964. Initially managed by the Ontario Department of Health, the cancer registry began tracking cancer incidence in 1964 and retrospectively collected cancer mortality data from as far back as 1950. In 1970, the OCTRF took ownership of the cancer registry. A more complete description of the historical milestones of the registry is described elsewhere (Clarke et al. 1991).

### Automation and OCRIS

One major transformation in the long history of the OCR was the adoption of the Ontario Cancer Registry Information System (OCRIS) in the early 1980s. Previously, the cancer registry was curated manually using records received from hospitals and cancer centers. This approach resulted in significant delays in the processing of case records. With the advent

of OCRIS and its automatic record delivery, the OCFRF was capable of receiving records almost instantly. OCRIS' improvements to data collection and the development of case resolution – the sophisticated system of computerized medical logic – further established the province's cancer registry as an important tool in cancer control. Enhancements to OCRIS were later made in the 1990s, and for the next 20 years, it continued to be an integral component of cancer registration in the province.

### EDW Reconstruction

In 2014, many years of work culminated in the first major reconstruction of the cancer registry since the adoption of OCRIS. The registry was rebuilt within the newly adopted technology of the Enterprise Data Warehouse (EDW). This change also coincided with the adoption of new standards for the registration of cancer cases, specifically the Multiple Primary and Histology (MPH) coding rules of the Surveillance, Epidemiology, and End Results Program (SEER). The need to modernize the OCR through technological improvements was prompted by greater demand for the registry's business intelligence capabilities. The new EDW-based OCR was officially launched in October 2014.

### Who Uses OCR Data and for What Purpose?

In recent years, the community of users of cancer registry data has expanded beyond the traditional audience of epidemiologists, cancer surveillance analysts, public health researchers, and policy analysts. Increasingly, the healthcare provider community, health services researchers, and cancer system planners are turning to population-based cancer registries like the OCR for foundational data to address questions related to clinical care and healthcare planning. The following sections highlight several examples of the OCR's stakeholders.

### Examples of Provincial Stakeholders

The Cancer Quality Council of Ontario is an arm's length agency of the Ontario government tasked with measuring the performance of the Ontario cancer system. The Council relies on OCR data to generate the Cancer System Quality Index, which reports on quality measures aimed at stimulating improvement in the cancer system.

Informing program delivery is another example of cancer registry data use. In partnership with CCO, Ontario's Regional Cancer Programs administer programs and services for cancer prevention and care in all 14 of the province's local health authorities and the Local Health Integration Networks (Fig. 2). OCR data are a source of information used by these networks in the planning, integration, and funding of local healthcare, as well as in improving access and the patient experience.

CCO also regularly shares OCR data with its provincial partners and collaborators, including:

- Pediatric Oncology Group of Ontario, Ontario's lead agency on childhood cancer surveillance, research, care, and support
- Institute for Clinical Evaluative Sciences, a research institute that performs many leading evaluative studies on healthcare delivery and outcomes, often by linking together health data such as physician billing claims and hospital discharge abstracts with cancer data
- Cancer Research Institute of Queen's University, which undertakes studies of cancer etiology, tumor biology, clinical trials, as well as outcomes and health services research
- Public Health Ontario, an agency dedicated to protecting and promoting the health of all Ontarians and reducing inequities in health through surveillance and research related to chronic and communicable diseases.

### Examples of National Stakeholders

The OCR is 1 of 13 provincial and territorial cancer registries that populate the Canadian Cancer Registry managed by Canada's statistical agency (Statistics Canada). The Canadian Cancer



**Fig. 2** Map of Ontario's Local Health Integration Networks. 1. Erie St. Clair, 2. South West, 3. Waterloo Wellington, 4. Hamilton Niagara Haldimand Brant, 5. Central

West, 6. Mississauga Halton, 7. Toronto Central, 8. Central, 9. Central East, 10. South East, 11. Champlain, 12. North Simcoe Muskoka, 13. North East, 14. North West

Registry is the main source of cancer statistics used in cancer health planning and decision-making at the national level. The OCR represents the Canadian Cancer Registry's largest provincial source of cancer data and, as a result, greatly influences national cancer statistics. The provincial and territorial cancer registries work with the Canadian Cancer Registry program to establish national standards for registry operations and data collection.

CCO also collaborates with the Canadian Partnership Against Cancer, a national agency that leads the performance measurement of Canada's cancer system. The partnership uses OCR and other data from CCO and other provincial cancer agencies to identify disparities in cancer care and management at the national and provincial levels.

### Examples of International Stakeholders

CCO actively shares OCR data with international organizations such as the North American Association of Central Cancer Registries (NAACCR) and the International Agency for Research on Cancer (IARC). The registry data are also used

in numerous international research initiatives, including but not limited to the International Cancer Benchmarking Partnership (ICBP) and the CONCORD studies on cancer survival.

Established in 1987, NAACCR is an umbrella organization for North American cancer registries, governmental agencies, professional associations, and private groups interested in the dissemination of cancer data. NAACCR achieves its mission through the active participation of selected US state cancer registries and Canadian provincial and territorial cancer registries. As with other member registries, the OCR shares its data with NAACCR annually. The compiled data are used to present North American cancer statistics in NAACCR's annual publication (*Cancer Incidence in North America*).

The OCR is one of several provincial cancer registries that submits its data every 5 years to IARC for inclusion in a compendium of cancer incidence data from internationally recognized cancer registries called *Cancer Incidence in Five Continents*. Data on childhood cancer incidence are also submitted by the OCR for inclusion in IARC's *International Incidence of Childhood Cancer* report.

**Table 1** The OCR's four main data sources for incident record creation

Source	Type(s) of information	Relative rank of importance in record creation	Load frequency into EDW-OCR
<b>Pathology (from public and private laboratories)</b>	Pathology reports and diagnostic test results	1	Weekly
<b>ALR (from Regional Cancer Centers)</b>	Treatment, past medical history and out-of-province records	2	Monthly
<b>DAD and NACRS (from CIHI)</b>	Admissions, discharge and surgery data	3/4	Monthly
<b>Death certificates (from the Registrar General of Ontario)</b>	Cause of death; Fact of death	3	Typically every 18–24 months; Every quarter

*ALR* Activity Level Reporting, *DAD* Discharge Abstract Database, *NACRS* National Ambulatory Care Reporting System, *CIHI* Canadian Institute for Health Information

The ICBP is a global initiative that combines the OCR with 12 comparable population-based cancer registries. The ICBP's registry data spans six countries across three continents. Open only to registry jurisdictions with universal access to healthcare and similar levels of healthcare spending, the ICBP aims to optimize the cancer policies and services of its partners. To date, the OCR has participated in three of five of the ICBP's research modules, exploring the topics of cancer survival, delays between treatment and diagnosis, and short-term survival (ICBP booklet 2014).

The CONCORD study was the first worldwide analysis of its kind to systematically compare cancer survival across five continents, involving 101 cancer registries from 31 countries (Coleman et al. 2008). Canadian data in the study was composed of the OCR and four other provincial and territorial cancer registries. The OCR was used again in the follow-up CONCORD-2 study, which assessed survival across 279 population-based cancer registries from 67 countries (Allemani et al. 2015).

databases, laboratory reports, and clinical records, including:

- Pathology reports
- Activity Level Reporting from Regional Cancer Centers (RCCs)
- Surgery and discharge data from the Canadian Institute for Health Information (CIHI)
- Death certificates
- Notification of out of province diagnosis or treatment of Ontario residents

Each data source is managed differently by the OCR and serves a unique purpose in record creation (Table 1).

It is uncommon to have a single data source for any given cancer case (Fig. 3), but certain sources are more commonly available than others. For example, of the 233,020 incident cases recorded between 2010 and 2012, 84 % included a pathology report. In 7 % of all cases, pathology reports were the only given source record. By comparison, 60 % of all cases had a corresponding NACRS record. However, in less than 0.1 % of all cases, NACRS was the only provided source record.

## Data Sources

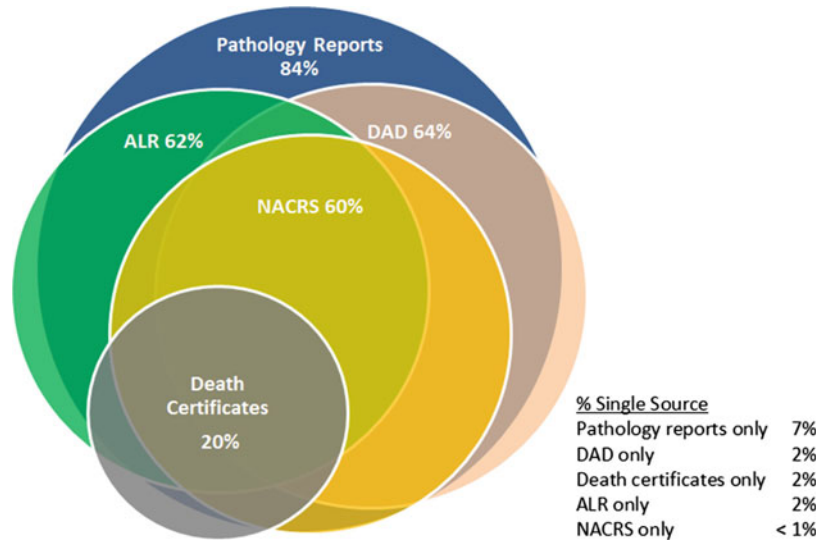
OCR records are created using data collected for purposes other than cancer registration. The data come from various administrative

## Pathology

Pathology reports are the main diagnostic source for new case record creation (Table 1). Through the



**Fig. 3** OCR data sources for 2010–2012 incident cases, showing the proportion of case records that included a particular data source. Also shown, percent of case records generated from a single source (Source: Ontario Cancer Registry, 2015)



ePath electronic pathology reporting system, CCO receives over one million pathology reports each year, sent in from 47 provincial facilities. In 2014, 237,834 of these reports were cancer relevant to 173,226 unique reports. To efficiently handle this large volume of information, pathology data is loaded into the EDW-OCR on a weekly basis.

Pathology reports are delivered to CCO in one of two forms – as narrative or as both narrative/synoptic reports. Narrative reports describing a patient’s pathology test results are those that have been written in sentence form or orally transcribed. While these types of reports can be submitted electronically, they cannot be handled automatically and are difficult to query. Coders must manually review narrative reports to derive relevant information and verify if there is indeed a cancer diagnosis.

Narrative reports currently account for approximately 70 % of all pathology reports received by CCO. The other 30 % of reports are received in synoptic form, a highly structured and standardized format of data submission submitted electronically. These reports improve overall completeness, ease of data exchange, treatment related decision-making, and turnaround time. First implemented in 2009, the synoptic pathology reporting system in Ontario is derived from the Electronic Cancer

Checklist (eCC) developed by the College of American Pathologists. Checklists and standard data fields in the eCCs eliminate the descriptive language found in narrative reports. Synoptic reports can be submitted in real time, making them a significantly more efficient method of pathology reporting.

One promising development is the inclusion of biomarkers in synoptic reporting. Biomarkers are laboratory indicators that can help identify abnormal processes, conditions, or disease. With respect to cancer care, biomarkers are of particular interest as they can provide information on cancer etiology, prognosis, and diagnosis. Examples of commonly used biomarkers include HER2 for breast cancer, KRAS for colorectal cancers, and ALK for lung cancer. In collaboration with the College of American Pathologists’ Pathology Electronic Reporting Committee, CCO is working to create biomarker templates for synoptic reporting. By September 2016, all 19 of Ontario’s genetic facilities are expected to implement eCC biomarker reporting. In preparation, Ontario has mandated 5 biomarker eCCs for lung, colorectal, breast, stomach cancers and melanoma. CCO is also equipped to handle optional for use biomarker eCCs for endometrial, gastrointestinal stromal tumor, myeloid, lymphoid, and CNS tumors.

## Activity Level Reporting

Data submitted by RCCs include Activity Level Reporting (ALR). ALR consists of patient records pertaining to radiation and systemic therapy services as well as oncology clinic visits. Sixty-two percent of new cancer cases in the OCR from 2010 to 2012 included ALR as a reporting source (Fig. 3). Some out-of-province data are collected for patients that access cancer services outside of Ontario (e.g., in neighbouring provinces). The loading of ALR data into the OCR occurs on a monthly cycle. ALR data can be reported in either ICD-10 or ICD-O-3 coding systems.

## DAD and NACRS

CIHI supplies data from the Discharge Abstract Database (DAD) and National Ambulatory Care Reporting System (NACRS). DAD includes administrative, clinical, and demographic data pertaining to all hospital in-patient discharges. NACRS reports all hospital- and community-based ambulatory care in day surgery, outpatient clinics, and emergency departments. As of 2002, all CIHI data are coded in ICD-10-CA.

## Death Certificates

Death certificates are obtained by the OCR from the Registrar General of Ontario. This information is used to track the vital status of patients in the registry and ensure that all incident cancer cases have been identified, particularly those that are only identified upon death. This process is known as death clearance.

Coded death certificates are received between 18 and 24 months after death. In lieu of death certificates, CCO also accepts fact of death for death clearance. CCO receives fact of death records approximately every quarter, describing deaths that have occurred approximately 6 months prior to the current quarter. Unlike death certificates, fact of death does not provide any insight into an individual's diagnosis of cancer and can only be used to close existing cases in the OCR.

## Data Systems and Consolidation

### OCRIS and the EDW Successor

OCRIS served as CCO's cancer registry information system since the 1980s. In an effort to modernize the registry and align it with current standards, OCRIS was formally decommissioned and replaced by the Enterprise Data Warehouse (EDW)-based OCR in late 2014. The EDW was initially designed to store ALR data for examining treatment and financial metrics, but in 2005 the decision was made to reconstruct the cancer registry within the EDW.

The EDW is composed of numerous data holdings, three of which are primarily related to cancer registration (see "[Technical Appendix](#)" for more details):

- Pathology/source data mart
- Ontario Cancer Registry (EDW-OCR)
- Collaborative Staging (CS) data mart

CCO's IT team is responsible for EDW support, data load, linkages, .net support and technical quality assurance.

## Patient Linkage

Through the key processes of patient linkage and case resolution, the OCR registrars are able to generate linkable records that combine all relevant data while eliminating redundant records. The EDW-OCR also permits any manual correction of cases at the record level, something not previously possible with OCRIS. Although the OCR relies on various automatic processes, manual review and input are still required to verify the completeness and accuracy of information for cancer registration.

Patient linkage is one of cancer registration's most fundamental processes and involves a combination of deterministic and probabilistic linkage routines to aggregate a person's source records into a "best" linked person record, which is a composite record representing the individual. This

entails the linking of new source records to existing person records. Source records that do not match to existing person records are consolidated and added to the OCR as new person records. However, there are several challenges with the linkage process. Aside from administrative errors like the misspelling of names or varying date formats, not all reports contain identical data elements. Unlike data from CIHI, ALR, or ePath, death certificates fail to provide patient Health Insurance Numbers. Because of the inconsistency in source data, deterministic linkage is ruled out as a major method for creating patient records and probabilistic linkage is used instead. Nonetheless, deterministic linkage is used to supply names to CIHI records (via health card number) and some other identifiers to other sources records, using the provincial client registry.

Probabilistic linkage allows matching of data where the completeness of matching variables is not 100 % and tolerates typing errors, transpositions, initials, nicknames, etc. Through probabilistic linkage matches are assigned a total match score (weight). Matches with the highest weights are automatically accepted, matches with low weights are rejected and links falling between the high/low thresholds are manually reviewed. The Master Patient Linkage links incoming CIHI, ALR and Pathology data to existing OCR persons. Incoming data that does not link to existing persons results in the addition of 'new' OCR persons. The Death Linkage links incoming death certificates to the OCR. Death certificates with a cancer cause of death that do not link to an existing OCR person result in the addition of a 'new' OCR person and a 'Death Certificate Only' cancer case.

Incorrect linkage would have several implications. For example, if multiple reports were not linked to their respective patient, redundant "persons" or cases would be generated. This would result in the over-reporting of cancer incidence. Similarly, if death certificates were not linked to the correct person record, the

existing case would not pass death clearance and the OCR would over-report cancer survival and prevalence.

## Case Resolution

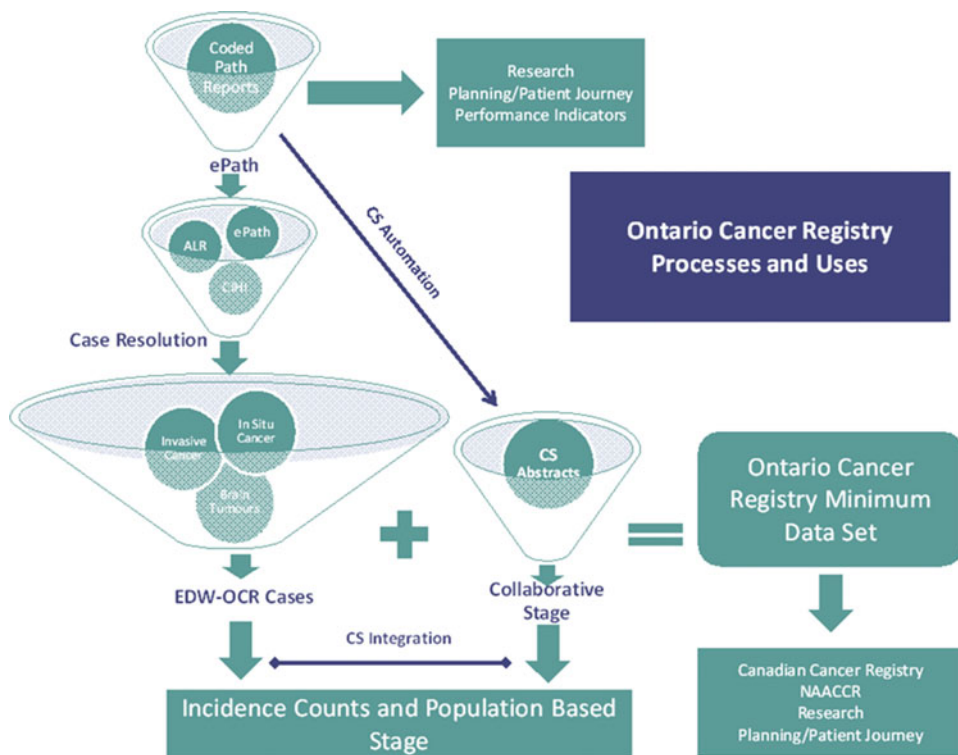
While the goal of patient linkage is to tie patient records together, case resolution works to consolidate these data into individual cancer cases. The immense volume of data received by CCO for the purpose of curating the OCR necessitates a highly competent system to handle information and pare it down into discrete cases. Case resolution does this by identifying individuals to process, reviewing their source data records, and identifying any primary cancers. A rigorous set of rules are then used to automatically produce a "best" diagnosis from the available data. At this point, only incident cases that have passed the various checks and filters remain.

Unlike patient linkage, case resolution is an automatic process without concurrent manual review. Automated logic processes all source records for a person, making cases for reportable neoplasms. Any case found to be non-incident, problematic, or outside the interest of CCO is appropriately flagged.

Non-incident cases are legitimate cases which do not qualify because the specific diagnosis is not covered by the OCR definition of "incident," which normally includes only invasive, reportable cancers. This includes in situ cases as well as benign and borderline brain and central nervous system tumors.

Problematic cases are those that either conflict with the system or do not meet the basic criteria for a proven case. An example of the latter is a case consisting only of hospital discharge records. Because discharge data alone is not indicative of a diagnosis or outcome, a definite case cannot be created. After a follow-up review, problematic cases can be identified as incident or non-incident or combined with already existing cases.

Some cases are deemed as "out of OCR range" or not of interest. These cases do not qualify as



**Fig. 4** Diagram of interrelated OCR processes (except for mortality data; death certificates and fact of death are processed separately). Patient linkage and case resolution processes are scheduled to run bimonthly. *ePath* electronic

pathology data collection system, *ALR* Activity Level Reporting, *CIHI* Canadian Institute for Health Information, *CS* Collaborative Stage, *NAACCR* North American Association of Central Cancer Registries

incident cases because they fall outside of CCO’s rules on geography and timing. The rules on geography exclude patients with a residence at diagnosis listed as outside of Ontario. However, patients without a listed residence are still treated as incident cases and considered as living in Ontario. Timing rules dictate that any cases diagnosed prior to 1964 be labeled as “out of OCR range” and ignored. Any cases that are not flagged by these rules are considered incident cases.

Resolved cases provide cancer-specific information such as the conditions of the diagnosis (ICD code, age, date of diagnosis, etc.), incidence status (in situ, invasive, etc.), cancer stage, and other data pertinent to oncologists and researchers.

All non-pathology source data come to CCO precoded. Because of the divergent coding

sometimes used by the sources, case resolution logic may mistakenly create multiple cases for a single person. Manual reviewers examine source records and merge any such cases. All cases are subject to manual review 6 months after their creation.

The successful completion of these processes allows for the creation of an OCR minimum dataset for a given year (Fig. 4).

### Data Elements

Information in the OCR spans several domains of data including demographic and vital statistics, tumor characteristics, treatment, and patient identifiers (Table 2).

**Table 2** Data domains and elements in the OCR

Data domain	Available data elements
Demographic and vital statistics	Date of birth Age at diagnosis Sex of patient Census tract, division, and subdivision Last known Place of residence Date of death
Tumor characteristics	Date of diagnosis Non-incident status Method of diagnosis/confirmation Type of pathology report Stage at diagnosis Stage (overall, clinical and pathological) Primary site (ICD-O-3 site code) Histology (ICD-O-3 histology code) Morphology Topography Site-specific factors Laterality SEER diagnosis group Clinical practice group Place of residence at diagnosis
Treatment	Local Health Integration Network Public health unit Treatment facility Treatment date Date of last contact Care site ID Discharge count (DAD) Surgery count (NACRS) ALR/RCC Number of pathology reports
Identifiable/linkable Information	Place of residence at diagnosis Patient name Ontario Health Insurance Plan number Health card number

*SEER* Surveillance, Epidemiology, and End Results Program; *DAD* Discharge Abstract Database, *NACRS* National Ambulatory Care Reporting System, *ALR* Activity Level Reporting, *RCC* Regional Cancer Center

Because numerous source records are often tied to a single case, some data elements such as date of diagnosis must be derived using algorithms. In this case, establishing the date of

diagnosis is an automated activity. First, all source records linked to a case are chronologically ordered. Then, the earliest date is selected as the date of diagnosis, regardless of record type. The complexity of the algorithms used varies depending on the nature of the element. For example, the methods used to generate stage data are considerably more complex (see section “[Cancer Stage at Diagnosis](#)”).

### Data Quality

The quality of cancer incidence data in the OCR compares favorably with that of other provincial and national cancer registries. The OCR adheres to four dimensions of data quality: comparability, completeness, accuracy, and timeliness (Parkin and Bray 2009).

Comparability is defined as the extent to which registry practices adhere to standard guidelines, which include the criteria for registration, coding systems such as ICD-O-3, multiple primary counting rules, and more. The standardization of OCR procedures ensures its comparability and compatibility with other cancer registries.

Completeness refers to how well incident cancer cases are registered. Specifically, how closely registry values for incidence and survival reflect the population’s true values. The OCR’s ability to draw upon multiple data sources to register cases, often with multiple sources per case, is conducive to a high level of completeness. OCR completeness is further verified through case-finding audits, record linkage with national and provincial databases, and comparisons with historic values.

Accuracy pertains to how well case records resemble their actual values. Just as with completeness, the OCR maintains a high level of accuracy thanks to its use of multiple data sources. Accuracy is further improved with re-abstraction studies and recoding audits, histological verification of cases, examining “death certificate only” cases, and analyses of missing information and internal inconsistencies.

**Table 3** Data quality indicators (NAACCR standard) for OCR 2008–2012 data years<sup>a</sup>

Indicator (% of all cases)	Year				
	2008	2009	2010	2011	2012
Completeness of case ascertainment	94.9	95.0	96.1	99.1	94.8
Missing age	0.0	0.0	0.0	0.0	0.0
Missing sex	0.0	0.0	0.0	0.0	0.0
Death certificate cases only (DCO)	1.0	1.3	1.5	1.4	1.8
Passing edits checks	100	100	100	100	100

<sup>a</sup>Current as of Nov 2015

Timeliness is the speed with which a registry can collect, process, and report complete and accurate cancer data. OCR timeliness is contingent upon two variables – the time until receipt and the time to process. The time until receipt refers to the time elapsed from diagnosis to delivery to CCO. With the exception of cause of death information through death certificates, which are typically received after 18–24 months, CCO receives and loads data into the EDW on a regular schedule (see section “Data Sources”).

Every year the OCR shares its data with NAACCR as part of its annual call for data, which is one of several calls for data by other organizations throughout the year to which the OCR responds. The measures of quality using the NAACCR data quality standard are shown in Table 3 for reference.

## Other Factors Affecting Quality

### Registration System

One concern regarding data quality pertains to the recent transition from OCRIS to the EDW-OCR. With each bimonthly case resolution cycle, EDW case data evolves. Existing cases expire and are replaced with a new case file. Cases tied to OCRIS, namely, all data from before 2010, remain unaffected and are listed as “frozen.” In order to mitigate variability, the data mart also tracks new versions of old case files. When new and old case files maintain a fixed degree of similarity, the two are linked in a process called case chaining. Case chaining ensures that current case files can be found once an older case is retired. Variability

can also arise in instances where registrars manually edit EDW-OCR data by merging cases together or modifying diagnosis codes and dates.

### Data Auditing

CCO practices routine data audits to verify the accuracy of its data holdings. One such audit is for inter-rater reliability aimed at assessing the level of agreement among coders or staging abstractors. These audits are necessary to minimize the loss of data integrity as a result of human error and establish consistency.

In a 2015 audit for stage quality, the inter-rater reliability between 16 CCO analysts was carried out. Each analyst independently staged an identical set of 96 randomly chosen cases diagnosed from 2012 to 2013. The “de-identified” set of cases included an equal amount of breast, colorectal, lung, and prostate primaries. Restrictions placed on the analysts prevented them from consulting each other or accessing full patient records, case histories, or pathology reports. Audits such as these allow CCO to discover any issues in data quality and promptly rectify them. Among the group of 16 analysts, an overall agreement rate of 93.5 % was found. In such audits, CCO strives to maintain a crude agreement rate of at least 90 %.

### Data Sources and Timing of Loads

As part of the transition from OCRIS to the EDW-OCR, some of the data source rankings have changed. In particular, pathology reports have replaced ALR data for the highest rank. This can be attributed to the more reliable and efficient nature of some sources, which makes them more valuable. Case data quality can be

further examined by performing NAACCR Edit Checks (Table 3). These checks identify cases that warrant further review. Often times such cases are coded incorrectly, with invalid topography and morphology combinations, unconfirmed multiple primaries, and other errors which are easily rectified.

Delays in the delivery and handling of source data mean that case resolution and registration cycles can on occasion be out of sync. As previously mentioned, DAD, NACRS, and ALR data are loaded on a monthly basis. In comparison, pathology (ePath) data is loaded weekly. However, as the case resolution and registration cycles operate on a bimonthly schedule, any misrepresentations of data become negligible over time.

### **Ontario Patients Treated Outside Ontario - Removal of Duplicates**

Statistics Canada conducts a national duplicate resolution process with the provincial and territorial cancer registries each year to account for multiple reporting of cases (e.g., due to patients moving between jurisdictions). The exchange of data between provincial and territory cancer registries is necessary to resolve duplicate cases and identify cases that may be missed, such as among individuals who access cancer services outside of their home province. For example, residents of northwestern Ontario will often use out-of-province cancer services in the neighboring province of Manitoba. Data sharing agreements exist between provinces for the exchange of such information.

### **Death Clearance**

Death certificates are used for the purpose of death clearance, a process that uses the coded cause of death (where cancer is the underlying cause) to identify individuals who were not previously recognized as having cancer. These “death certificate only” cases represent under 2 % of incident cases (Table 3). Unless fact of death is established otherwise, death certificates are necessary to keep accurate survival and prevalence rates. Currently, there are no routinely

scheduled releases of death certificates from the Registrar General of Ontario. As a result, the death clearance process may occur long after incident cases from other sources have been identified.

---

## **The OCR Adopts a New Approach to Counting Cancers**

Counting practices for OCRIS incident cases were modeled after standards set by the International Agency for Research on Cancer (IARC) and the International Association of Cancer Registries (IACR). These counting rules were very conservative and inflexible for patients diagnosed with multiple primaries. Given that approximately 10–14 % of cases with a single primary will develop a subsequent cancer within 25 years, cancer incidence counts would be under-reported by overlooking such subsequent primaries. The modified IARC/IACR rules used by OCRIS did not recognize paired organs (e.g., left or right breast or lung) or colon and skin melanoma subsites, nor did it have timing rules to recognize new, subsequent primary cancers in the same organ. As a result, OCRIS likely reported lower rates of multiple primaries than other registries with more liberal rules, including those using the SEER Multiple Primary and Histology (MPH) coding rules.

However, starting with cases diagnosed in 2010, the OCR implemented the SEER MPH coding rules, which use four criteria for counting multiple primaries: topography, morphology, laterality, and timing (Johnson et al. 2012).

### **Topography and Morphology**

Topography refers to a cancer’s anatomic site of origin, while morphology describes the type of cell and its biological activity. The morphology of cancers is recorded with two codes, describing the cancer’s histology and behavior. In OCRIS, additional primaries were only added to the registry when cancers expressed both a different topography and morphology from the initial primary

cancer. As shown in Table 4, the OCR accepts cancers that are morphologically identical but have different topography, and vice versa, as being multiple primaries.

## Laterality

Laterality applies mainly to paired organs and differentiates similar cancers by organ subsite. The IARC/IACR rules do not recognize laterality. In cases where both paired organs, such as the left

and right kidney, were reported with invasive tumor, only a single primary would be recognized. Using the SEER rules, paired sites are considered in the registration of multiple primaries. As outlined in Table 5, only specific topographic sites are subject to the rules on laterality. With cancers of the central nervous system, the laterality rule only applies to benign and borderline tumors. Malignant central nervous system tumors remain exempt.

## Timing

The diagnosis of multiple primaries can be typically described as synchronous or metachronous. Synchronous cancers are those that develop at the same time or within a small time frame, while metachronous cancers occur in sequence of one another. Data on metachronous cancers are of particular importance to researchers as they provide insight into causal mechanisms involved in the formation of subsequent neoplasia. IARC/IACR rules dictate that the existence of two or more primary cancers does not depend on time and are therefore recognized as a single primary case. The SEER rules on timing allow metachronous cancers to exist as multiple primaries. As shown in Table 5, a cancer must have developed after a specified period of time to qualify as a multiple primary.

**Table 4** Criteria for classifying cancers as multiple primaries under the modified IARC/IACR rules in OCRIS compared to SEER Multiple Primary and Histology rules in the OCR

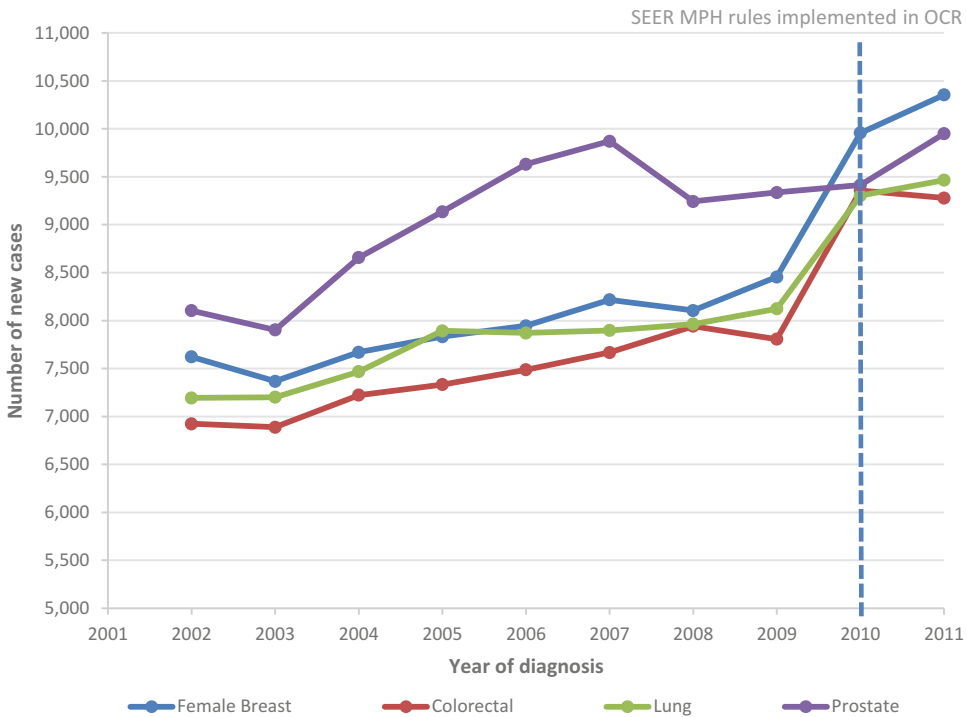
Criteria	Multiple primary rule	
	IARC/IACR (in OCRIS)	SEER MPH (in OCR)
<b>Same topography and different morphology</b>	No (in general)	Yes
<b>Different topography and same morphology</b>	No (in general)	Yes
<b>Laterality</b>	No	Yes
<b>Timing</b>	No	Yes (in general)

*IARC* International Agency for Research on Cancer, *IACR* the International Association of Cancer Registries, *SEER* Surveillance, Epidemiology, and End Results Program, *MPH* multiple primary and histology

**Table 5** Applicable SEER multiple primary counting rules for laterality and timing (Source: *SEER Multiple Primary and Histology Coding Rules Manual, 2007*)

Cancer type	Laterality	Timing
<b>Breast</b>	Yes	5 years
<b>Head and neck</b>	Yes	5 years
<b>Kidney</b>	Yes	3 years
<b>Lung and pleura</b>	Yes	3 years
<b>Urinary</b>	Yes	3 years
<b>Colon</b>	Yes	1 year
<b>Melanoma</b>	Yes	60 days
<b>Benign and borderline central nervous system tumors</b>	Yes	Does not apply
<b>Malignant central nervous system tumors</b>	Does not apply	Does not apply
<b>Other sites</b>	Yes, if considered a paired site	1 year, if applicable
<b>Invasive diagnosis 60 days after an in situ diagnosis</b>	Does not apply	60 days





**Fig. 5** Effect of implementing the SEER Multiple Primary and Histology counting rules on incidence, by cancer type in Ontario, 2002–2011. Note the seemingly disproportionate rise in the incidence of breast, colorectal, and

lung cancers while the incidence of prostate cancer remained relatively fixed (Source: Ontario Cancer Registry, 2015 (Cancer Care Ontario))

### Implications of Counting Rules on Data and Analysis

There was a substantial increase in incident cases following the adoption of the SEER MPH rules by the OCR. This change is due to how cancers were being counted rather than indicating that more people in Ontario were being diagnosed with or dying of cancer. The new rules allow for a more complete accounting of cancer incidence, which improves the ability for regions and communities in Ontario to plan for the future needs of the cancer system.

To further examine the change imposed by the new counting rules, IARC/IACR and SEER MPH rules were compared for 2010 and 2011 incident cases (Candido et al. 2015). According to this analysis, overall there were 5.8 % more cases using the SEER MPH rules, the increase varying by morphology, topography, sex, and age. The greatest change was observed in older age groups

and those diagnosed with melanoma of the skin, female breast cancer, and colorectal cancer. The incidence of colorectal, female breast and lung cancers rose considerably following the implementation of the SEER MPH counting rules (Fig. 5). However, the incidence of prostate cancer remained largely the same.

### Best Practices for Analysis

From an analytic perspective, if an analysis spans the OCRIS and OCR datasets, special care must be taken to reconcile the two. More specifically, data from 2010 onward must first be made IARC/IACR-compatible by using those multiple primary counting rules, which then allow for trend analyses under a common rule.

For cancer projections, the projections must be undertaken based on incidence counts using the IARC/IACR rules and then be modified with a

correction factor that accounts for the effect of the SEER MPH counting rules.

---

## Cancer Stage at Diagnosis

Cancer stage at diagnosis reports the extent of a cancer's invasion and spread beyond the primary site. Factors used in staging include the tumor's topographic site, size, multiplicity, invasiveness, lymphatic penetration, and metastases. In a clinical setting, this information helps determine a patient's appropriate course of treatment and provides an estimate of their prognosis. The dominant clinical staging method is the tumor, node, and metastasis (TNM) staging system and the collaborative staging (CS) framework (which is based on TNM) used by North American cancer registries.

TNM staging reports cancer stage as a function of tumor, node, and metastasis. First, the primary tumor is classified by type, size, and extent. Next, the level of lymph node involvement is determined. Lastly, any metastases are examined to assess the cancer's spread from the primary site. By taking these data into consideration, an overall stage value, ranging from 0 to IV, can be assigned.

CS is a unified data collection framework designed to use a set of data elements based on the extent of disease and clinically relevant factors. The primary objective of CS was to reconcile the American Joint Committee on Cancer's (AJCC) TNM staging system, Union for International Cancer Control TNM staging system, and the SEER Summary Staging system. This change brought about a significant reduction in data redundancy and duplicate reporting. Furthermore, it retained data integrity for both clinical and public health researchers while improving accessibility and compatibility.

The input data items for CS that are collected from the medical record include both clinical diagnostic results, like imaging, biopsy, and other tests, and any cancer resection surgery results. Each data element has an additional field that identifies whether it was collected from clinical or resection surgery findings and an indicator if neoadjuvant therapy was performed prior to surgery. The CS algorithm then automatically

derives a single set of T, N, M, and stage group, which will be clinical or pathologic, depending on how the extent of disease was discovered within the diagnostic and treatment process.

The staging guidelines for CS require significantly more information than is included in clinical or pathological TNM reports. CCO staging analysts require data on tumor size, depth of invasion, the number and location of positive lymph nodes, as well as site-specific factors (see details below). TNM often does not specify these raw data elements, nor does it provide a cancer stage indicator that combines clinical and pathology data. In order to derive the CS, staging analysts must also review patient pathology and medical records in addition to TNM reports. One calendar year is typically dedicated to the CS capture process for a given diagnosis year.

One significant change that accompanied the adoption of CS was the introduction of site-specific factors (SSFs). SSFs provide supplementary information unique to a cancer type to assist in the staging process. This information expands the understanding of tumor characteristics, prognosis, and predicted treatment response. SSFs for several cancer types were introduced with AJCC 7th edition for cases diagnosed in 2010 onward. Furthermore, the implementation of SSFs allows registries to collect data on biomarkers and other factors that were previously not collected.

## CS Automation and Integration

CS is generated in a two part hybrid system (see section "[Data Systems and Consolidation](#)," Fig. 4). The first part, CS automation, is an automated process that populates CS abstracts by identifying stageable registry cases and linking them to synoptic pathology reports. Stageable cases are those that contain data sufficient to derive a TNM stage, either using the CS data collection system or by manual TNM staging. CS abstracts are required to organize and summarize case data pertinent to the staging process. Staging analysts remotely access hospital electronic patient records to determine if clinical diagnostic tests need to be added to the CS input information. This also

occurs when synoptic pathology data are insufficient for abstraction purposes or are unavailable.

The second process called CS integration requires a more fine-tuned approach. It involves a probabilistic tumor linkage between case and abstract followed by manual review of unlinked abstracts. CS integration involves reviewing abstracts to determine a “best stage” and linking it back to cases in the OCR. This process necessitates remote access to the electronic patient record at hospitals. Currently, CCO is the only organization in Ontario authorized to exercise this level of direct access to electronic patient records.

### Source of Staging Data

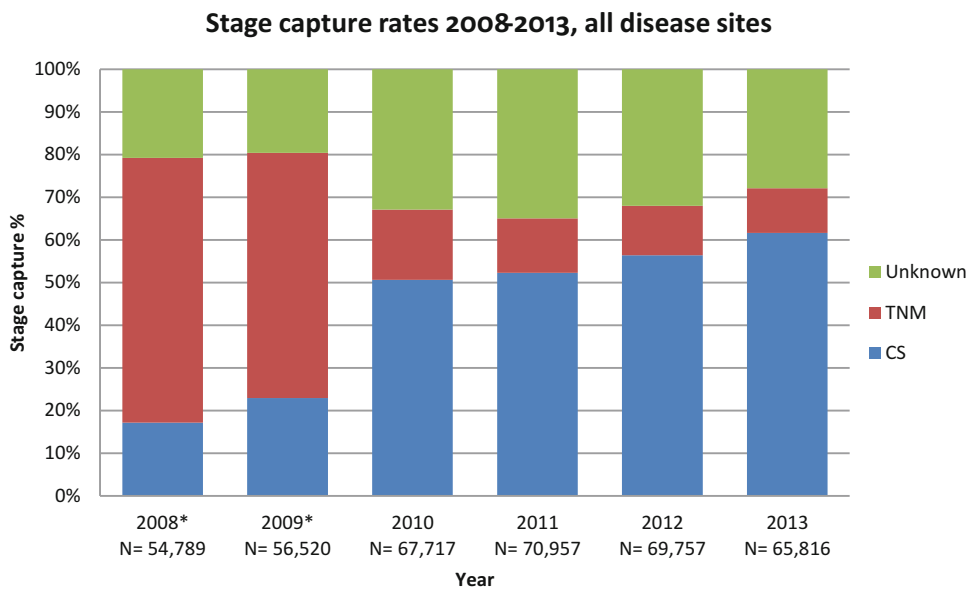
In 2005, CS was captured for only a subset of patients from outside RCCs, representing less than 15 % stage capture in the first year of data collection. It has since expanded to include both RCCs and non-RCC hospitals. Staged TNM data is received from Ontario RCCs, while non-RCC

hospitals have made patient records available to the OCR’s staging analysts.

### Stage Capture Rates

Stage capture refers to the completeness of stage information on all stageable cancer cases identified by the registry. Of the 65,816 cases identified in 2013, approximately 72 % were stageable (Fig. 6). CS was derived for 85 % of those stageable cases, a significant improvement from its introduction in 2008.

The proportion of CS cases increased substantially after 2010, when a national initiative led by the Canadian Partnership Against Cancer supported the Canada-wide adoption of CS. The rise in stage capture rates can also be attributed to the progressive rollout by CCO of CS to a greater number of cancer types (Table 6). CS was officially implemented in the OCR for the four most common cancers in 2010. In 2011, the use of CS grew to include melanoma of the skin and gynecologic cancers, followed by



\* 2008 and 2009 capture rates determined using modified IARC/IACR rules for counting multiple primaries.

**Fig. 6** Stage capture rates using Collaborative Stage (CS) and Tumor Node Metastasis (TNM), Ontario, 2008–2013 (Source: Ontario Cancer Registry, 2015 (Cancer Care

Ontario). \* 2008 and 2009 capture rates determined using modified IARC/IACR rules for counting multiple primaries)

**Table 6** Progressive rollout by OCR of collaborative staging by cancer site

Cancer type	Year of full implementation of collaborative stage by OCR
<b>Breast</b>	2010
<b>Lung</b>	2010
<b>Colorectal</b>	2010
<b>Prostate</b>	2010
<b>Gynecologic cancers</b>	2011
<b>Melanoma of skin</b>	2012
<b>Thyroid</b>	2013

further expansion in 2013 to include thyroid cancer.

In 2015, there was a decision to retire the CS system and to implement TNM for population-based staging in North America. The change in staging system is expected to take effect in Canadian provincial registries with cases diagnosed in 2017. Despite the decision to return to TNM staging, the AJCC has stated that it is committed to keeping SSFs an integral part of the staging process. Discussions are still underway in Canada about additional factors relevant to population-based staging that may need to be collected.

## Linkage of the OCR to Other Datasets

By linking the OCR with other datasets, researchers can obtain a more comprehensive understanding of healthcare issues. Whether linked with CCO or external datasets, the OCR can serve as the basis for research studies, especially when patient-level data is available. Datasets regularly linked with the OCR for the purpose of health services research include the following healthcare utilization databases:

- Ontario Health Insurance Plan claims
- Ontario Drug Benefit claims
- CIHI's Discharge Abstracts Database (DAD) for inpatient hospitalizations
- CIHI's National Ambulatory Care Reporting System (NACRS)

This section describes the dataset linkage process within the OCR and outlines several of CCO's other data holdings that may be of interest to health services researchers.

## Other Linkage Processes

Pending approval by CCO's data disclosure team, CCO may process cohort files submitted by external researchers (see section "[Data Privacy and Access](#)" for more information on data access). At minimum, these cohort files must include names and birthdates of all patients to be processed. Additional identifiable information such as health card numbers (HCNs), postal codes, and gender may be included in the cohort file to expedite the linkage procedure and any necessary manual resolutions. After a suitable cohort file has been received by CCO, a linked file may be produced. In the interest of efficiency, cohort files for less than 300 individuals are linked to the OCR manually through a name search function. Cohort files for over 300 individuals necessitate a probabilistic linkage in the same manner as described in section "[Data Systems and Consolidation](#)," but with the use of Automatch software. The software compares records from client files to the OCR and assigns a total score corresponding to how closely the records match. Matches on uncommon names will receive a higher score than matches on common names, indicating greater confidence in the link.

These linkages are to a subset of OCR data. Subsets are pared down to comply with research parameters. For example, if the cohort represents females enrolled in a research survey which commenced in 2002, the subset will not contain female patients who died prior to 2002 or any male patients. Typical information released from the OCR includes person key (a unique identifier for an OCR person), date of diagnosis, topography, morphology, vital status, and date of last contact or death.

The probabilistic linkage will match the cohort file to the OCR person records and assign a total match score (weight). Matches with the highest weights will be automatically accepted, matches with low weights will be rejected and links falling

between the high/low thresholds are manually reviewed. The high/low thresholds will be determined by an OCR data analyst through analysis of the data. The unique identifier for a OCR person will then be used to select case level data from the OCR for the cohort members that were identified as matches to the OCR. The final product of the linkage is a file of matched records which typically includes information related to the cancer diagnosis and vital status information.

### CCO's Other Data Holdings

CCO data holdings store information collected from healthcare service providers across the province. This information enables the planning and funding of cancer and other healthcare services, development of guidelines, and management of the cancer and renal care systems in Ontario. The major data holdings are shown in Table 7. Details about the data held within each of these repositories can be found on CCO's website, [www.cancercare.on.ca](http://www.cancercare.on.ca).

Other provincial organizations with which CCO maintains a data sharing agreement (DSA) will sometimes create linkages with OCR data. One such example is the Institute for Clinical Evaluative Sciences, which uses their version of the OCR data received from CCO to perform in-house linkages for research purposes. The differences in dataset versions between CCO and organizations that receive CCO data through data sharing agreements can be identified through their respective data dictionaries, which are often available online.

---

### Health Services Research Using the OCR

The OCR has been a source of data for projects across the cancer continuum. A review of the peer-reviewed literature suggests that use of Ontario cancer data in health research dates back to the 1970s. A series of papers in the 1970s by MacKay and Sellers reported on the burden of cancer by using the OCR. Such cancer

surveillance reports eventually evolved to describe province-wide patterns and trends in healthcare delivery aimed at managing and planning for the cancer system, allocating resources, as well as evaluating and monitoring the cancer system. Between 1973 and 2014, more than 460 peer-reviewed articles were published using data from the OCR. The frequency of OCR data use grew substantially following the 1990s (Fig. 7). In the last 2 years of available data (2013–2014), 120 peer-reviewed research articles were published citing use of the OCR. This growth may be attributed to improvements in information capture in the healthcare sector and the growing availability of healthcare data. For instance, within CCO, ALR has evolved in its ability to measure and monitor activity related to systemic treatment, including chemotherapy and radiation therapy. Similarly, CCO's recent implementation of the Wait Time Information System increases the scope of data the patient experiences in the healthcare system. Moreover, as healthcare-related information has become more readily available in electronic format, the potential for data linkage and exploration of research topics has continued to grow.

This section presents specific examples of how the OCR has been used for health services research. The works cited in this section provide some recent examples of data linkage between the OCR and other administrative data sources or linkage with primary data collected by the research study.

### Examples of Health Services Research Using the OCR

Using date of diagnosis, geography, and demographic information, researchers frequently extract data from the OCR for descriptive purposes, to explore trends over time, patterns of care, and potential gaps in access and equity. Using this approach, researchers have described the postoperative mortality risk among the elderly (Nanji et al. 2015), wait times from abnormal mammography to surgery among breast cancer patients

**Table 7** CCO's major data holdings as of September 2015 (*Source: CCO, 2015*)

Data holding	Description	Type of data
<b>Activity Level Reporting (ALR)/Cancer Activity Datamart</b>	Provides an integrated set of data elements from Regional Cancer Centers (RCC) related to systemic treatment and radiotherapy that cannot be obtained from other providers. This information is used to support management decision-making, planning, accountability, and performance management at the RCC, regional, and corporate level.	This dataset contains administrative data, clinical data, and demographic data.
<b>Patient Information Management System (PIMS)/Pathology Datamart</b>	Database comprised of patient and tumor information for cancer and cancer-related pathology reports (tissue, cytology), submitted from public hospital (and some commercial) laboratories. PIMS documents patient, facility, report identifiers, and tumor identifiers, such as site, histology, and behavior. This information is used to support management decision-making, planning, disease surveillance and research, as well as contributing to resolved incidence case data in the Ontario Cancer Registry.	This dataset contains administrative data, clinical data, and demographic data.
<b>New Drug Funding Program (NDFP)</b>	The NDFP database stores patient and treatment information about systemic therapy drug utilization at RCCs and other Ontario hospitals, for which reimbursement is being sought through the NDFP according to strict eligibility criteria.	This dataset contains: administrative data, clinical data (eligibility criteria) and demographic data.
<b>Ontario Breast Screening Program (OBSP)</b>	The associated Integrated Client Management System database provides an integrated set of data for each client screened in the OBSP for the purposes of program administration, management, and evaluation.	This dataset contains administrative data, clinical data, and demographic data.
<b>Colorectal Screening Data – Colonoscopy Interim Reporting Tool (CIRT)</b>	The data collected through CIRT will be used to understand current colonoscopy activities conducted within participating hospitals from both volume and quality perspectives. It will also be used to validate incremental volume allocations across the province.	This dataset contains: administrative care and clinical data.
<b>Laboratory Reporting Tool (LRT)</b>	LRT contains CCC program FOBT (fecal occult blood test) kit distribution and results data from the CCC partner labs.	
<b>Ontario Cervical Screening Program</b>	Cytobase is comprised of cervical cytology data ("Pap Test" results) collected from participating community laboratories. This cervical cancer screening database contains patient, physician, and laboratory information. This information is used to administer and evaluate the performance of CCO's Cervical Screening Program, for cancer planning and management and for cancer surveillance research.	This dataset contains administrative data, clinical data, and demographic data.
<b>Brachytherapy Funding Program</b>	Stores patient and treatment information about prostate cancer patients at RCC hospitals, for which reimbursement is being sought.	This dataset contains administrative data, clinical data, and demographic data.

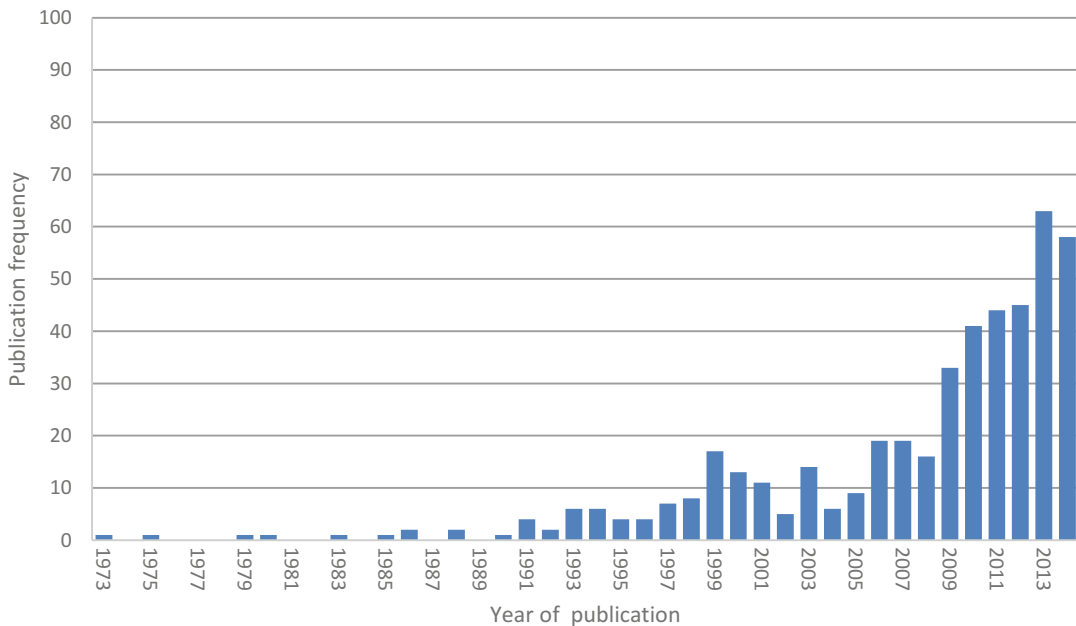
(continued)

**Table 7** (continued)

Data holding	Description	Type of data
<b>Symptom Management Reporting Database</b>	The Symptom Management Reporting Database data is comprised of three components: patient registration, symptom screening using the Edmonton Symptom Assessment System (ESAS) and functional assessment using the Palliative Performance Scale and/or Eastern Cooperative Oncology Group Performance Status. This information is captured by participating sites using the Interactive Symptom Assessment and Collection system and then submitted on a monthly basis to the Symptom Management Reporting Database.	This dataset contains administrative data, clinical data, and demographic data.
<b>Interim Annotated Tumor Project (ATP) Database</b>	The Interim ATP provides an integrated set of data, combining tumor information from the Ontario Institute for Cancer Research's Tumor Bank with CCO's Cancer Registry, for the purpose of increasing the accuracy and utility of the information for both researchers and CCO planners. For example, researchers may use this information to study the association between genetics and response to cancer drugs; in turn, CCO may use this information to create clinical guidelines for the care and treatment of cancer patients in Ontario.	This dataset contains administrative data, clinical data, and demographic data.
<b>Wait Times Information System (WTIS)</b>	<p>The Wait Time Information System is the first-ever information system for Ontario to collect accurate and timely wait time data. This system has been implemented in 82 Ontario hospitals. Work is underway to enhance this system to track wait times for all surgical procedures in Ontario</p> <p>This web-based system performs several functions, which include:</p> <ul style="list-style-type: none"> <li>Enabling the collection of data related to wait times</li> <li>Providing clinicians and other health professionals with the tools required to effectively assess patient urgency according to a defined wait times standard</li> <li>Measuring and reporting wait times and data regarding utilization of procedures</li> <li>Supplying clinicians, administrators, and managers with near real-time information for use in monitoring and managing wait lists</li> <li>Reporting wait time information to the public on a website enabling patients to manage their own care and the public to assess progress on reducing wait times.</li> </ul>	This dataset contains administrative data, clinical data, and demographic data.

(Cordeiro et al. 2015), and rates of thyroid cancer among children, adolescents, and young adults (Zhukova et al. 2015). Through linkage with a dataset identifying 140,000 registered or “Status

Indians” in Ontario, researchers have been able to describe the cancer experience among the First Nations population in Ontario and study their survival rates over a 30-year time frame



**Fig. 7** Distribution of peer-reviewed publications using the OCR as a data source 1973–2014 (Source: CCO Surveillance and OCR, July 2015)

(Nishri et al. 2015). The ability of the OCR to identify patients in specific clinical subgroups has also enabled research studies to test concordance with clinical practice guidelines, such as the treatment of patients with stage II colon cancer (Biagi et al. 2009), follow-up surveillance of patients treated for Hodgkin’s lymphoma (Hodgson et al. 2010), and the use of single fraction radiotherapy for uncomplicated bone metastases (Ashworth et al. 2014).

Population-based retrospective cohort studies have used the OCR to identify cohorts of patients who were diagnosed during a given period of time, underwent particular therapeutic courses, or experienced a particular model of care. This approach has been used to carry out research to look at healthcare costs among colorectal cancer patients (Mittmann et al. 2013), the impact of active surveillance in prostate cancer (Richard et al. 2015) and the use of adjuvant chemotherapy among patients with early breast cancer (Enright et al. 2012). These studies make use of noncancer comparison groups or population-based comparisons through strategies such as random digit dialing. They may also use comparison groups

consisting of individuals with cancer who have experienced standard care, in which comparisons may also be derived from the OCR using treatment-based criteria. The OCR is able to provide covariates necessary for the statistical control of potential confounders in these comparative analyses (e.g., stage at diagnosis or date of diagnosis).

Studies incorporating survival analysis and modeling have been able to uncover factors associated with survival on a population level. Such studies have uncovered clinicopathological factors linked to survival among patients diagnosed with pancreatic adenocarcinoma (Kagedan et al. 2015), survival among bladder cancer patients receiving various treatment modalities (Leveridge et al. 2015; MacKillop et al. 2014a), survival among Ontario men who underwent radical prostatectomy, and general survival trends among individuals with laryngeal cancer (Macneil et al. 2015). By coupling OCR-defined cohorts with clinical data from sources such as surgical pathology reports, researchers have been able to associate the prognostic importance of specific clinical factors and provide direction for best practice in clinical



reporting (e.g., Berman et al. 2015). Other investigators have looked at variability in survival among patients visiting different RCCs in Ontario by linking the OCR with stage and treatment data (e.g., head and neck cancer – MacKillop et al. 2014b).

The OCR is also useful to health services researchers who are interested in the effectiveness of preventive strategies to control cancer, such as population-based screening programs. In this type of research design, the OCR data provides the clinical endpoint that will determine the effectiveness of screening intervention. The OCR has been used to capture rates of colorectal cancer among those individuals who had a positive guaiac fecal occult blood screening test as part of Ontario's Colon Cancer Check program and assess their risk of colorectal cancer over a 30-month time frame (Tinnmouth et al. 2015). The OCR has also been used to ascertain the rates of cervical cancer before and after the introduction of a human papillomavirus immunization program for girls in grade 8 (Smith et al. 2015). The OCR has been used widely to look at the effects of breast cancer screening and its various aspects though linkage with the data from the Ontario Breast Screening Program. This research has shed light on the role of mammographic density in screening outcomes (Boyd et al. 2007), the contribution of clinical breast examination to breast screening (Chiarelli et al. 2009), and the performance of digital compared with screen-film mammography (Chiarelli et al. 2013).

## Patient Contact

Research teams will occasionally approach CCO to gain access to the OCR for the purpose of identifying individuals eligible to participate in cancer-related research studies. In these instances, analysts at CCO will work with research investigators to refine a set of criteria for participation in the study and extract a cohort from the OCR.

Until 2014, CCO had been in the practice of providing cohorts to the research team, who would then make contact with patients to request their participation, often via their physician. The current process for patient contact is initiated with

a letter from CCO. These letters are used to confirm that the individual has been diagnosed with cancer, inform said individual about the research being performed, and obtain consent for the release of their contact information to the researcher. Individuals are also provided the option to opt out of any such studies in the future.

The new and more standardized approach to patient contact minimizes the risk associated with erroneous identification of cancer patients and contacting patients who do not or do not yet know they have cancer, as well as patients who are deceased. The approach also ensures a more consistent and effective process for obtaining informed consent from study participants.

Examples of patient-contact studies using OCR-identified patients as a sampling frame have included a case-control study to identify risk factors associated with ovarian tumors (McGee and Narod 2010), a study of quality of life and health utilities among prostate cancer patient (Krahn et al. 2013), a dietary study among breast cancer patients (Boucher et al. 2012), and a survey of men with prostate cancer about decision-making around the use of complementary and alternative medicine (Boon et al. 2005).

---

## Data Privacy and Access

### Privacy

As a prescribed entity under the Ontario *Personal Health Information Protection Act*, CCO is permitted to collect, use, and disclose personal health information. By way of comparison, other prescribed entities within Ontario include the Pediatric Oncology Group of Ontario, Canadian Institute for Health Information, and the Institute for Clinical Evaluative Sciences.

CCO has robust information management practices, outlined within the Privacy Program, in place to ensure the protection of personal health information within the OCR and its other data holdings. These information management practices are audited on a triennial basis by the Office of the Information and Privacy Commissioner of Ontario.

CCO's Privacy Program includes privacy policies, standards, procedures, and guidelines. Its privacy assurance and risk management activities involve:

- Privacy impact assessments and risk mitigation plans
- Data sharing agreements
- Standard operating procedures

Staff privacy training and awareness activities are in place to maintain a culture of privacy across the organization. A data access program, described below, is implemented to review and approve external and internal requests for access to OCR data.

## Data Request Process

CCO understands the value of health services research and has therefore implemented the data disclosure team to assist researchers and other data requestors in accessing its data holdings. Outlined in Table 8 are the four types of data requests typically received by CCO.

Figure 8 outlines CCO's data disclosure process and the various internal groups involved. The Data Disclosure Subcommittee oversees all research data requests and occasionally some general data requests, in adherence with the *Personal Health Information Protection Act* and CCO's Data Use and Disclosure Standard. This group also reviews CCO's data disclosure policies and

procedures. Before obtaining final approval by the Data Disclosure Subcommittee, research data requests must undergo an extensive review by subject matter experts in the data disclosure working group.

## Technical Appendix

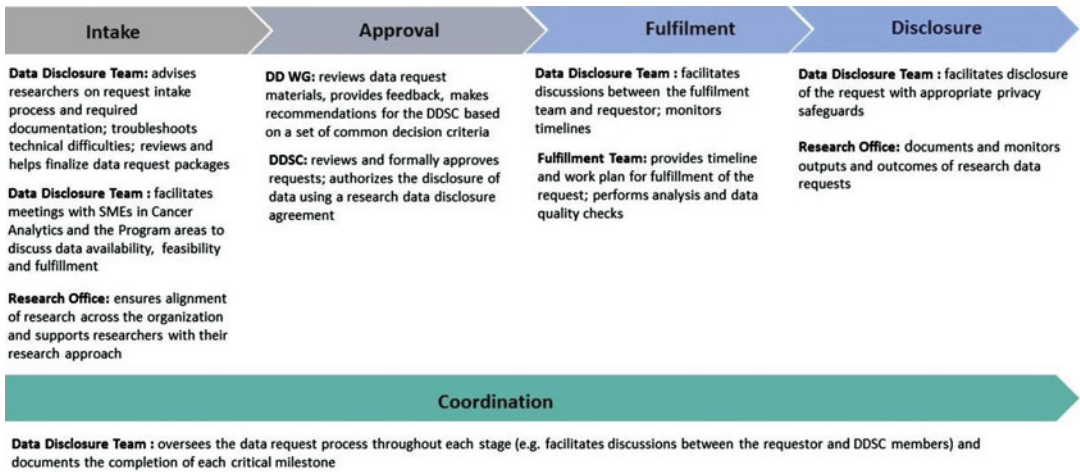
Synoptic pathology reports are an integral component of the EDW and feed the Pathology Data Mart, which is needed for CS integration (Fig. 9). Synoptic pathology reports from the Pathology Data Mart, OCR case files and CS abstracts are utilized by the Registry Plus service to drive CS integration and populate the CS data mart (see section "[Cancer Stage at Diagnosis](#)" for more information on CS and its processes). Registry Plus is a suite of publicly available free software programs for collecting and processing cancer registry data (Centers for Disease Control and Prevention 2015).

### ePath, eMaRC, and ASTAIRE

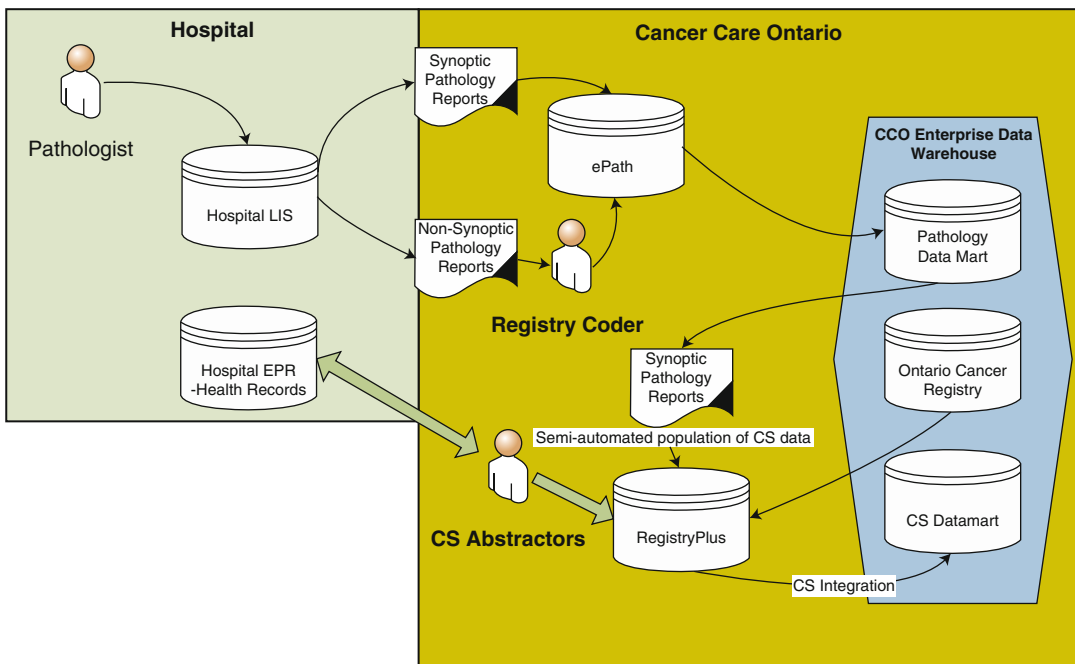
All pathology reports are handled through CCO's ePath electronic pathology reporting system. ePath receives, processes and stores pathology reports, connecting the diagnostic laboratories to the OCR. ePath is comprised of several major subsystems, including the Electronic Mapping, Reporting, and Coding (eMaRC) and the Automated Synoptic Template Analysis Interface and Rule Engine (ASTAIRE).

**Table 8** The four types of data requests received by CCO

Request type	Description
Research data requests	Requests from external researchers for record-level data (personal health information or de-identified data) for the purposes of conducting scientific studies. This type of request also includes patient-contact studies, where CCO contacts prospective participants to obtain consent for permission to be contacted for a research study
General data requests	Nonresearch requests for record-level or aggregate data for a variety of health system planning purposes, including regional performance management, quality assurance, and information dissemination. Currently this type of request also includes private company requests for record-level or aggregate data, for purposes such as marketing and economic analyses
Genetic requests	Requests from genetic counselors for pathology reports, with consent from the individual in question or substitute decision-maker, to facilitate the genetic counseling process
SEER*Stat requests	Requests from external partners for the latest SEER*Stat de-identified data software package to facilitate the production of aggregate cancer incidence and mortality statistics



**Fig. 8** Data disclosure process at CCO. DD WG data disclosure working group, DDSC data disclosure subcommittee



**Fig. 9** Diagram of pathology-driven processes at CCO. LIS laboratory information system, EPR electronic patient record, CS Collaborative Stage

CCO eMaRC is a subcomponent of the ePath system, which processes and stores pathology reports received in HL7 messaging format. CCO eMaRC automatically filters cancer vs non-cancer reports and non-reportable reports and provides

partial automation for numerous ICD-O-3 diagnoses codes, collaborative staging elements, and creates NAACCR compatible abstract records. The system also merges multiple reports for a single patient so as to prevent the creation of extra cases in the OCR.

A data quality assessment tool, ASTAIRE ensures that synoptic data is compliant with the College of American Pathologists standards. ASTAIRE is made up of three components: GINGER, FRED, and ADELE. Combined, GINGER and FRED ensure that synoptic data is sufficiently complete and in line with current eCC versions. ADELE then cleans data so that may be admitted to the EDW.

In the interest of privacy and efficiency, data handled through ePath is coded in Health Level Seven V2 format, which is a secure method of data transmission designed to protect sensitive health information. This data contains three main elements: patient ID (PID), observation report ID (OBR), and observations (OBX). Patient ID contains personal and identifiable information, such as a patient's name, sex, and address. The observation report ID pertains to the pathology report and provides information regarding the pathologist, surgeon, referrals, and specimen collection. The observations data element conveys information regarding the clinical diagnosis, clinical history, gross pathology, submitted tissues, and full diagnosis.

## References

- Allemani C, Weir HK, Carreira H, Harewood R, Spika D, Wang XS. Global surveillance of cancer survival 1995–2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet*. 2015;385(9972):977–1010.
- Ashworth A, Kong W, Chow EL, Mackillop W. The fractionation of palliative radiation therapy for bone metastases in Ontario. Paper presented at: The 56th Annual Meeting of the American Society for Radiation Oncology; San Francisco; Sept 2014.
- Berman DM, Kawashima A, Peng Y, Mackillop WJ, Siemens DR, Booth CM. Reporting trends and prognostic significance of lymphovascular invasion in muscle-invasive urothelial carcinoma: a population-based study. *Int J Urol*. 2015;22(2):163–70.
- Biagi JJ, Wong R, Brierley J, Rahal R, Ross J. Assessing compliance with practice treatment guidelines by treatment centers and the reasons for noncompliance. Paper presented at: The 2009 Annual Meeting of the American Society of Clinical Oncology; Orlando; May 2009.
- Boon H, Westlake K, Deber R, Moineddin R. Problem-solving and decision-making preferences: no difference between complementary and alternative medicine users and non-users. *Complement Ther Med*. 2005;13(3):213–6.
- Boucher BA, Cotterchio M, Curca IA, Kreiger N, Harris SA, Kirsh VA, et al. Intake of phytoestrogen foods and supplements among women recently diagnosed with breast cancer in Ontario, Canada. *Nutr Cancer*. 2012;64(5):695–703.
- Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic density and the risk and detection of breast cancer. *N Engl J Med*. 2007;356(3):227–36.
- Cancer Act, R.S.O 1990, c. C.1 [Internet]. 22 June 2006 [cited 28 Oct 2015]. Available from: <http://www.ontario.ca/laws/statute/90c01>
- Candido E, Young S, Nishri D. One cancer or two? The impact of changes to the rules for counting multiple primary cancers on estimates of cancer burden in Ontario [Internet]. In: Proceedings of the 2015 Canadian Society for Epidemiology and Biostatistics Conference; 1–4 June 2015; Mississauga/Toronto: Cancer Care Ontario; 2015. Available at: [http://csebca.ipage.com/wordpress/wp-content/uploads/2014/06/June-2\\_1430\\_SouthStudio\\_C1.2-Candido.pdf](http://csebca.ipage.com/wordpress/wp-content/uploads/2014/06/June-2_1430_SouthStudio_C1.2-Candido.pdf)
- Centres for Disease Control and Prevention. Registry Plus, a suite of publicly available software programs for collecting and processing cancer registry data [Internet]. Atlanta: National Center for Chronic Disease Prevention and Health Promotion; Jan 2015 [cited 28 Oct 2015]. Available at: <http://www.cdc.gov/cancer/npcr/>
- Chiarelli AM, Majpruz V, Brown P, Theriault M, Shumak R, Mai V. The contribution of clinical breast examination to the accuracy of breast screening. *J Natl Cancer Inst*. 2009;101(18):1236–43.
- Chiarelli AM, Edwards SA, Prummel MV, Muradali D, Majpruz V, Done SJ, et al. Digital compared with screen-film mammography: performance measures in concurrent cohorts within an organized breast screening program. *Radiology*. 2013;268(3):684–93.
- Clarke EA, Marrett LD, Kreiger N. Cancer registration in Ontario: a computer approach. *IARC Sci Publ*. 1991;95:246–57.
- Coleman MP, Quaresma M, Berrino F, Lutz JM, De Angelis R, Capocaccia R, et al. Cancer survival on five continents: a worldwide population-based study (CONCORD). *Lancet Oncol*. 2008;9(8):730–56.
- Cordeiro ED, Dixon M, Coburn N, Holloway C. A patient-centered approach toward wait times in the surgical management of breast cancer in the province of Ontario. *Ann Surg Oncol*. 2015;22(8):2509–16.
- Enright K, Grunfeld E, Yun L, Moineddin R, Dent SF, Eisen A, et al. Acute care utilization (ACU) among women receiving adjuvant chemotherapy for early breast cancer (EBC). Paper presented at: The 2012 Breast Cancer Symposium; San Francisco; Sept 2012.
- Hodgson DC, Grunfeld E, Gunraj N, Del Giudice L. A population-based study of follow-up care for Hodgkin lymphoma survivors: opportunities to improve

- surveillance for relapse and late effects. *Cancer*. 2010;116(14):3417–25.
- International Cancer Benchmarking. Showcasing our findings and impacts. London: Cancer Research; Dec 2014 [cited 28 Oct 2015]. Available from: [http://www.cancerresearchuk.org/sites/default/files/icbp\\_pb\\_1012214\\_booklet\\_final.pdf](http://www.cancerresearchuk.org/sites/default/files/icbp_pb_1012214_booklet_final.pdf)
- Johnson CH, Peace S, Adamo P, Fritz A, Percy-Laurry A, Edwards BK. The 2007 Multiple Primary and Histology Coding Rules [Internet]. Bethesda: National Cancer Institute's Surveillance Epidemiology and End Results Program; Aug 2012. Available at: <http://seer.cancer.gov/tools/mphrules/>
- Kagedan DJ, Raju R, Dixon M, Shin E, Li Q, Liu N, et al. Predictors of actual survival in resected pancreatic adenocarcinoma: A population-level analysis. Paper presented at: The 15th Annual Americas Hepato-Pancreato-Biliary Congress; Miami Beach; Sept 2015.
- Krahn MD, Bremner KE, Alibhai SM, Ni A, Tomlinson G, Laporte A, et al. A reference set of health utilities for long-term survivors of prostate cancer: population-based data from Ontario, Canada. *Qual Life Res*. 2013;22(10):2951–62.
- Leveridge MJ, Siemens DR, Mackillop WJ, Peng Y, Tannock IF, Berman DM, et al. Radical cystectomy and adjuvant chemotherapy for bladder cancer in the elderly: a population-based study. *Urology*. 2015;85(4):791–8.
- MacKillop W, Siemens R, Zaza K, Kong W, Peng P, Berman D, et al. The outcomes of radiation therapy and surgery for bladder cancer: a population-based study. Paper presented at: The 56th Annual Meeting of the American Society for Radiation Oncology; San Francisco; Sept 2014a.
- MacKillop W, Kong W, Zaza K, Owen T, Booth C. Volume of practice and the outcomes of radiation therapy for head and neck cancer. Paper presented at: The 56th Annual Meeting of the American Society for Radiation Oncology; San Francisco; Sept 2014b.
- Macneil SD, Liu K, Shariff SZ, Thind A, Winkvist E, Yoo J, et al. Secular trends in the survival of patients with laryngeal carcinoma, 1995–2007. *Curr Oncol*. 2015;22(2):85–99.
- McGee J, Narod S. Low-malignant-potential tumor risk factor analysis: a matched case-control analysis. Paper presented at: The 41st Annual Meeting of the Society of Gynecologic Oncologists; San Francisco; Mar 2010.
- Mittmann N, Isogai PK, Saskin R, Liu N, Porter J, Cheung MC, et al. Homecare utilization and costs in colorectal cancer. Paper presented at: Healthcare Cost, Quality, and Policy: Driving Stakeholder Innovation in Process and Practice Conference; Toronto; Nov 2013.
- Nanji S, Mackillop WJ, Wei X, Booth CM. Management and outcome of colorectal cancer (CRC) liver metastases in the elderly: A population-based study. Paper presented at: The 15th Annual Americas Hepato-Pancreato-Biliary Congress; Miami Beach; Sept 2015.
- Nishri ED, Sheppard AJ, Withrow DR, Marrett LD. Cancer survival among First Nations people of Ontario, Canada (1968–2007). *Int J Cancer*. 2015;136(3):639–45.
- Parkin DM, Bray F. Evaluation of data quality in the cancer registry: principles and methods part II: Completeness. *Eur J Cancer*. 2009;45:756–64.
- Personal Health Information Protection Act; June 2016 [cited July 2016]. Available from: <https://www.ontario.ca/laws/statute/04p03>
- Richard PO, Alibhai S, Urbach D, Fleshner NE, Timilshina N, Klotz L, et al. The uptake of active surveillance in prostate cancer: Results of a population based-study. Paper presented at: The 2015 Annual Meeting of the American Urological Association; New Orleans; Apr 2015.
- Smith LM, Strumpf EC, Kaufman JS, Lofters A, Schwandt M, Levesque LE. The early benefits of human papillomavirus vaccination on cervical dysplasia and anogenital warts. *Pediatrics*. 2015;135(5):1131–40.
- Tinmouth JM, Lim T, Kone A, Mccurdy B, Dube C, Rabeneck L. Risk of colorectal cancer among those who are gFOBt positive but have had a recent prior colonoscopy: experience from an organized screening program. Paper presented at: Digestive Disease Week 2015; Washington, DC; May 2015.
- Zhukova N, Pole J, Mistry M, Fried I, Bartels U, Huang A, et al. Clinical and molecular determinants of long-term survival in children with low grade glioma; a population based study. Paper presented at: The 16th International Symposium on Pediatric Neuro-Oncology in Conjunction with the 8th St. Jude-VIVA Forum; Singapore; 28 June 2015–2 July 2015.



# Challenges of Measuring the Performance of Health Systems

# 17

Adrian R. Levy and Boris G. Sobolev

## Contents

<b>Introduction</b> .....	391
<b>Background</b> .....	392
<b>Performance Measurement in the Canadian Health-Care System</b> .....	392
Data Requirements .....	394
<b>A Case Study on Performance Measurement: Health Technology Assessment</b> .....	394
<b>Existing Research on Performance Measurement in Health Technology Assessment</b> .....	397
<b>Data Sources for Performance Measurement in Health Technology Assessment</b> .....	398
<b>Discussion</b> .....	399
<b>Recommendations</b> .....	400
<b>References</b> .....	401

## Abstract

Improving the measurement of the performance of health systems is a wise policy option for federal, provincial, and territorial governments because it provides essential information for understanding the inevitable trade-

offs involved in trying to reduce costs while striving to improve quality of care, access, and the health of the population. Performance measurement – monitoring, evaluating, and communicating the degree to which health-care systems address priorities and meet specific objectives – is also garnering increased attention from many stakeholders at other levels of the system.

---

A. R. Levy (✉)  
Community Health and Epidemiology, Dalhousie  
University, Halifax, NS, Canada  
e-mail: [adrian.levy@dal.ca](mailto:adrian.levy@dal.ca)

B. G. Sobolev (✉)  
School of Population and Public Health, University of  
British Columbia, Vancouver, BC, Canada  
e-mail: [bgsobolev@gmail.com](mailto:bgsobolev@gmail.com)

---

## Introduction

In 2010, the 11th in a series of annual reports was published and presented the most recent health indicator data from the Canadian Institute for

Health Information and Statistics Canada on a broad range of performance. Each indicator falls into one of the four dimensions of the health indicator: (1) health status provides insight on the health of Canadians, including well-being, human function, and selected health conditions; (2) nonmedical determinants of health reflect factors outside of the health system that affect health; (3) health system performance provides insight on the quality of health services, including accessibility, appropriateness, effectiveness, and patient safety; and (4) community and health system characteristics provide useful contextual information, rather than direct measures of health status or quality of care. The goals of this chapter are to characterize the goals of a high-functioning health-care system and provide a typology for performance measures in health care. Both of these will be done within the context of the renewal of the First Ministers' Accord.

The intent of the 1984 *Canada Health Act* was to ensure that all residents of Canada have access to medically necessary health care on a prepaid basis. However, the act has not been uniformly applied across provinces and territories, leading to variability in available services and treatment in different jurisdictions. The federal government's determination to adhere to national standards while reducing funding to the provinces has recently produced additional challenges. (For more details about federalism and health care in Canada, see Wilson (2000)).

The system currently used to measure the performance of the health system in Canada lags behind that of other countries such as the United States and the United Kingdom, both in terms of standardized indicators and research in the area. As a result, there is evidence indicating that the values of Canadians are misaligned with the funding and performance of the health-care system (Snowdon et al. 2012).

In this chapter, the authors review the current state of knowledge about performance measurement in health care and examine current efforts in Canada. We describe the structural, political, conceptual, and methodological challenges of performance measurement in the field of health technology assessment. We argue that without more clarity around ethics and perspectives and a more

systematic approach to performance measurement, it will not be possible to develop a coherent strategy for informing policy-making and decision-making throughout the entire health-care system.

---

## Background

At a fundamental level, "the primary aim of evaluation is to aid stakeholders in their decision making on policies and programs" (Alkin 2004). It is intended to provide evidence on the degree to which government policies and spending are effectively addressing specific issues identified by bureaucrats and elected officials.

Performance management in the public sector became a focus of interest in the late 1980s, starting with the reinventing government movement (Osbourne and Gaebler 1992). In the United States, the 1993 *Government Performance and Results Act* obligated all federal departments and agencies to present 5-year teaching plans linked to performance measures; annual performance plans were required after 1998. In the United Kingdom, the financial management initiative was introduced in the early 1980s.

In Canada, the federal government introduced a centralized evaluation policy in 1977. Using evidence from peer-reviewed sources and from reports of the auditor general, Shepherd argued that, between 1977 and 2009, Canada's evaluation policy was focused on operational issues directed primarily toward program managers (Shepherd 2012). In 2009, federal evaluation policy was refocused on fiscal prudence and accountability.

---

## Performance Measurement in the Canadian Health-Care System

Over the past 25 years, there has been an increase in measuring and reporting on the performance of the Canadian health-care system at the federal, provincial, and territorial levels. On the demand side, provincial and territorial governments and health authorities have been subjected to intense pressure to contain costs; patients have greater expectations to be involved in decisions about

their treatment; and health-care professionals and health authorities expect more oversight and accountability be built into the health-care system. On the supply side, the information revolution and progress in information technology have made it less expensive and more straightforward to collect, process, and disseminate data.

There have been several attempts to define the problem of how to measure health-care performance in Canada, the necessary first step toward aligning goals and objectives. In 2000, the First Ministers' Meeting Communiqué on Health directed Canada's health ministers to meet to collaborate on the development of a comprehensive framework to report on health status, health outcomes, and quality of service using jointly agreed-upon comparable indicators. The intent was that such reporting would meet several objectives by providing information to Canadians on government performance, as well as assisting individuals, governments, and health-care providers to make more informed health choices. In September 2002, all fourteen federal, provincial, and territorial governments released comparable indicator reports on a set of 67 indicators. The 2003 First Ministers' Accord on Health Care Renewal (Appendix 1) directed health ministers to develop more indicators to supplement work undertaken in response to the September 2000 communiqué and identified the following priority areas for reform: healthy Canadians, primary health care, home care, catastrophic drug coverage and pharmaceutical management, diagnostic and medical equipment, and health human resources. Federal, provincial, and territorial jurisdictions agreed on 70 indicators with 81 sub-indicators and established the Health Indicators Project to have them collated and make them publicly available.

Priorities and directions for the Health Indicators Project were broadly revisited at a second consensus conference in March 2004. The resulting consensus statement established that health indicators must be:

- Relevant to established health goals
- Based on standard (comparable) definitions and methods

- Broadly available and able to be disseminated electronically across Canada at the regional, provincial, and national level

The primary goal of the Health Indicators Project was to support health regions in monitoring progress in improving and maintaining the health of the population and the functioning of the health system for which they are responsible through the provision of good-quality comparative information on:

- The overall health of the population served, how it compares with other regions in the province and country, and how it is changing over time
- The major nonmedical determinants of health in the region
- The health services received by the region's residents
- The characteristics of the community or the health system

No mention was made of other potential uses for performance indicators, including establishing the competence of organizations and identifying the effectiveness of programs to meet specific objectives.

The communiqué from the 2004 First Ministers' Meeting on the Future of Health Care, called "A 10-Year Plan to Strengthen Health Care," included an explicit commitment to "accountability and reporting to citizens" that read: "all governments agree to report to their residents on health system performance including the elements set out in this communiqué." In so doing, the first ministers agreed that performance indicators would be required and would be used for reporting purposes. The intent of the effort was to hold health ministries accountable for stewardship of the health-care system using performance indicators. The communiqué did not specify whether such reporting would be used in a formative (to improve specific health systems) or in a summative (to implement corrective measures or impose penalties) fashion.



Consultations continued with provincial and regional health authorities to ensure that relevant data were collected and consistent methods were used for performance measurement. In 2012, the 13th in a series of annual reports presented health indicator data from the Canadian Institute for Health Information and Statistics Canada on a broad range of performance measures (CIHI 2012). The data were grouped into four dimensions of health: (1) health status, which provides insight on the health of Canadians, including well-being, human function, and selected health conditions; (2) nonmedical determinants of health, which reflect factors outside of the health system that affect health; (3) health system performance, which provides insight on the quality of health services, including accessibility, appropriateness, effectiveness, and patient safety; and (4) community and health system characteristics, which provide useful contextual information rather than direct measures of health status or quality of care.

That report used the following principles to categorize disparities in the health system:

- Same access to available care for the same need
- Same utilization for the same need
- Same quality of care for all

## Data Requirements

Those considering performance measurement are faced with many competing needs when designing information systems to serve a range of stakeholders (Table 1). A set of consensus performance measures needs to be developed iteratively, and those involved in the process must have a deep understanding of existing and potential data sources that can be used to create the measures. The specific circumstances of health care in Canada – such as Canada’s single-payer financing and several provincial and federal initiatives – have led to the development of key elements needed to produce some routine performance measures, including population registries, vital statistics, administrative health databases containing records of patients’ interactions with various elements of the health-care system, and

patient and treatment registries. As a result of the large amount of data collected in Canada, this country has been characterized as a data-rich environment (Roos et al. 2005). This is reflected by the activities of provincial data centers, which both serve as data custodians and collate and use administrative health and other databases for research and evaluation (Suissa et al. 2012).

Existing performance measures reported by the Canadian Institute for Health Information depend on information from provincial and territorial population registries, vital statistics, hospital discharge abstracts, and physician claims. Even though performance measures have been reported annually since 2003, there are concerns about the provinces’ ability to produce unbiased performance measures because of data quality; in Manitoba, the auditor was “unable to form an opinion on the accuracy of the data or on the adequacy of disclosure” for 21 of 56 health indicators used in the provincial report (Manitoba Minister of Health and Healthy Living 2004).

---

## A Case Study on Performance Measurement: Health Technology Assessment

In general, three types of outcomes are studied in health-care evaluations: those related to patients, those related to treatments, and those related to the system (Levy 2005). Patient-related outcomes represent the effects of delivering care in a particular system on the patient’s ability to care for himself or herself, physical function and mobility, emotional and intellectual performance, and self-perception of health. Treatment-related outcomes represent the biological and physiological changes in the patient’s condition that occur as a result of administering therapy within the health-care system. System-related outcomes represent the effect on the health-care system produced by the provision of medical services to a patient population.

Examples of the outcomes include performance benchmarks, requirements for pain medication, length of hospital stay, waiting times, frequency of readmission, and frequency and

**Table 1** Examples of health-care performance indicators and information needs according to the type of stakeholder

Stakeholder	Goals	Types of needed information
Citizens	<ul style="list-style-type: none"> <li>To see evidence that resources on health are being spent efficiently and align with stated priorities</li> <li>To have the information they need to hold policy and decision-makers accountable for health policies and health-care delivery that align with societal values</li> <li>To be reassured that necessary care will be forthcoming in time of need</li> </ul>	<ul style="list-style-type: none"> <li>Transparent descriptions of stated priorities</li> <li>Comparative information on the health of the population versus that in other countries</li> <li>Comparative information on the performance of the health-care system versus that in other countries</li> <li>Transparent access to indicators of access, quality of care, and resource use in the health-care system</li> </ul>
Patients	<ul style="list-style-type: none"> <li>To be reassured that they will have access to specific health care when they need it, within a safe timeframe and at adequate proximity</li> <li>To obtain information on the intended and unintended consequences of alternative health-care options and on the out-of-pocket expenses associated with these options</li> </ul>	<ul style="list-style-type: none"> <li>Information on available health-care services and modalities</li> <li>Information on trade-offs between services in terms of potential intended and unintended health outcomes and out-of-pocket costs</li> </ul>
Health-care professionals	<ul style="list-style-type: none"> <li>To provide high-quality and appropriate health care to patients</li> <li>To maintain and improve their knowledge and skills in health-care delivery</li> </ul>	<ul style="list-style-type: none"> <li>Data on individual performance against benchmarks</li> <li>Up-to-date information on best practices, guidelines</li> </ul>
Hospitals	<ul style="list-style-type: none"> <li>To monitor and improve the use of health-care resources</li> <li>To manage local budgets</li> <li>To identify and prioritize health technology acquisition and disinvestment</li> <li>To ensure patient safety</li> <li>To conduct continuous quality improvement</li> </ul>	<ul style="list-style-type: none"> <li>Collective data on health-care quality, including patient safety indicators measured against benchmarks</li> <li>Information on distributions of access (utilization, waiting lists, and waiting times) measured against benchmarks</li> <li>A transparent health technology assessment process</li> <li>Information on patient experience and satisfaction</li> <li>Hospital-level costing information</li> </ul>
Health authorities	<ul style="list-style-type: none"> <li>To ensure that hospitals and health-care professionals provide appropriate and cost-effective health care</li> <li>To ensure that patients have access to the specific health care they need, within a safe timeframe and at adequate proximity</li> <li>To manage regional budgets</li> <li>To assess the impact of health care on the regional health needs of the population</li> <li>To ensure equitable distribution of resources</li> </ul>	<ul style="list-style-type: none"> <li>Information on the comparative health of their population versus that of populations served by other health authorities</li> <li>Information on the health needs of their region</li> <li>Information on the equity of health-care resource distribution</li> <li>Information on distributions of access (utilization, waiting lists, and waiting times) across health authority</li> <li>Health authority-level costing information</li> </ul>
Governments	<ul style="list-style-type: none"> <li>To assess the impact of health care on patients and on population health</li> <li>To establish current and future health policy goals and priorities</li> <li>To set and manage governmental budgets</li> <li>To plan for the viability and sustainability of the health-care system</li> <li>To demonstrate the adequacy and proper functioning of regulatory procedures for health care</li> <li>To provide appropriate assessment and research infrastructure</li> <li>To promote investment and innovation in health care</li> </ul>	<ul style="list-style-type: none"> <li>Comparative data on the health of their population versus that of populations in other provinces and territories and in other countries</li> <li>Information on the societal value of health care, elicited using transparent citizen engagement processes</li> <li>Information on the health needs of the region</li> <li>Information on the equity of health-care resource distribution</li> <li>Information on distributions of access (utilization, waiting lists, and waiting times) across the jurisdiction</li> <li>Aggregate and decomposed expenditure data at the provincial, territorial, and national level</li> <li>Information on societal productivity attributable to health and health care</li> </ul>
Regulators	<ul style="list-style-type: none"> <li>To protect patient safety</li> <li>To ensure protection of health-care professionals and other consumers beyond patients</li> <li>To uphold their fiduciary responsibility</li> <li>To promote efficiency in health-care markets</li> </ul>	<ul style="list-style-type: none"> <li>Safety signals from health care</li> <li>Integrity in reporting financial performance</li> <li>Information on innovation in health care</li> </ul>

severity of secondary health complications. In health-care evaluation, a performance measure summarizes the distribution of a health-care outcome in the patient population. In most studies, the performance measure combines the observed responses for all patients or hospitals into a single number. For example, a performance study might record the timing and occurrence of a clinic appointment for each patient, with the distribution of time to clinic appointment (the health-care outcome) being summarized by the weekly rate of appointments (the performance measure).

There have been large investments in health technology assessment over the past decades, and the use of new health-care technology is an important driver of ongoing increases in health-care expenditures. Before an expensive new technology is implemented and covered in a jurisdiction, the expected impacts are assessed at the provincial level, and the technology's incremental cost-effectiveness is often assessed by the Canadian Agency for Drugs and Technology in Health; by several provinces, such as Ontario and Quebec; and by some Canadian hospitals (Levin et al. 2007; McGregor and Brophy 2005).

At the time a new health technology comes to market, there is typically little information on its benefits, safety, and cost implications for the population among whom the technology will be used. As such, health technology assessment provides an incomplete picture. It examines short-term safety, with a focus on the most common, serious (potentially life threatening), and severe (potentially debilitating) unintended consequences; efficacy, often using data from the restricted conditions in randomized trials; the acquisition costs; and, sometimes, estimated cost-effectiveness on the basis of long-term project models drawing on the limited information available at market launch.

Once the technology is marketed, some information becomes available on the geographic distribution of the technology and sometimes its utilization. However, this descriptive information alone is not adequate for assessing the performance of the technology. Decision-makers need to understand how new technologies affect patients once they have been adopted for use in

the real world, or they need to understand how they affect the health system in terms of who is actually treated, the long-term clinical benefits, severe unintended consequences, health-related quality of life, and productivity. Even less is known about the impact of less severe unintended consequences, downstream medical and health consequences (for the population to whom the technology is actually applied), population effectiveness, or incremental cost-effectiveness in actual use.

Many innovations have led to less invasive technologies being introduced to treat conditions previously managed surgically, such as percutaneous transluminal coronary angiography, which is now being undertaken in patients who were previously managed with coronary artery bypass grafting (Weintraub et al. 2012), and extracorporeal shock-wave lithotripsy, which has displaced surgical removal of kidney stones. Noninvasive technologies typically reduce patient morbidity and the length of hospital stay, often resulting in lower unit costs of treatment, and should therefore result in potential cost savings to the health-care system. However, understanding the long-term consequences of such technologies requires formal assessment because those savings are often not realized. Angioplasty leads to a greater need for repeat revascularization over time, which reduces the cost differential, and, perhaps because of reduced morbidity, the number of patients and treatments may increase after a new technology becomes established (Levy and McGregor 1995).

Although measuring the performance of new health-care technologies once they have been introduced into practice is crucial, it is done only rarely. The work of the Ontario Health Technology Advisory Committee is an exception (Levin et al. 2007). One reason is that there is a lack of indicators on a new health technology and a time lag of at least several years before administrative data becomes available for analysis in Canada. This knowledge gap is becoming increasingly problematic as governments, health authorities, and hospitals struggle to work within fixed budgets, with the federal government planning on indexing its spending to inflation. Decision-makers in these organizations have said clearly

that they suffer from a lack of straightforward information about which technologies work, on whom, and under what circumstances (Health Technology Assessment Task Group on behalf of the Federal/Provincial/Territorial Advisory Committee on Information and Emerging Technologies 2004). There is no consensus on, or even an understanding of, what should be measured or how performance should be measured.

---

### **Existing Research on Performance Measurement in Health Technology Assessment**

At least four groups of investigators have proposed methods to measure performance in health technology assessment. A group of investigators from the United Kingdom proposed a framework for describing decision-making systems that use health technology assessment to determine reimbursement of health technologies (Hutton et al. 2006). The framework groups systems under four main headings (constitution and governance, objectives, use of evidence and decision processes, and accountability) and identifies three processes (assessment, decision, and outputs and implementation). Hutton et al. assessed the feasibility of implementing the framework using published information on constitution and governance, methods and processes, the use of evidence, and transparency and accountability, at the stages of assessment, decision-making, and implementation. They found that most of the information needed for their framework was not publicly available.

A group of researchers from l'Université de Montréal proposed a framework for performance assessment in health technology assessment organizations (Lafortune et al. 2008). Their conceptual model includes four functions and organizational needs that must be balanced for a health technology agency to perform well: goal attainment, production, adaptation to the environment and culture, and value maintenance. Although this model has a strong conceptual grounding, it has yet to be applied in practice. It requires analysts to make qualitative judgments, which may make

it more useful for improving performance within an organization than for comparing performance between organizations.

More recently, a group of European investigators proposed an input-throughput-outcome model of the health-care system in relation to the different types of health-care technologies (Velasco et al. 2010). The thrust of their argument is that "health technology assessment should develop to increase its focus on the 'technologies applied to health care' (i.e., the regulatory and policy measures for managing and organizing health-care systems)." They recommend that health technology assessment should have an increased focus on regulatory, financial, and policy measures for managing and organizing health-care systems. They recommend that "countries embarking on health technology assessment should not consider establishing completely separate agencies for health technology assessment, quality development, performance measurement, and health service development, but should rather combine these agencies into a common knowledge strategy for evidence-informed decision-making in the health services and the health system." Although ambitious, there would be much to be gained from such a strategy.

The framework closest to assessing some of the performance measures listed in Table 1 was developed in Quebec (Jacob and McGregor 1993). These authors outlined a new methodology for evaluating the impact of health technology assessments on policy and expenditures and applied it to 21 assessments produced by the Quebec Council for Health Technology Assessment between 1990 and 1995. Using published documents, interviews, questionnaires, and administrative health data, the authors sought to evaluate the impact of health technology assessments by addressing three fundamental questions: (1) What impact was intended? (2) To whom was the message directed? (3) To what extent was the hoped-for impact achieved, first in terms of policy and second in terms of actual distribution and the use of the technology? The authors determined that 18 of the 21 assessments had an influence on policy and that there were substantial savings to the health-care system. They concluded that it will

rarely be possible to precisely estimate impact, but systematic documentation of effects can be achieved. The self-stated limitations of their methodology included the identification of what they called critical incidents, systematic categorization of policies about health technology, and the use of documentation, which led to a degree of objectivity but also led to limitations relating to the reliance on analysts' judgment. The interpretations were improved by consulting with important stakeholders. They also acknowledged that the impact of any health technology assessment is influenced by many other factors, substantially complicating interpretations. (Assessing causality when measuring performance of health technology is among the most pernicious challenges facing the careful analyst. This is made particularly challenging because of the impossibility of randomization in most studies. The thoughtful study by Jacob and McGregor (1993) is notable for its rigor and critical thinking in this area.)

None of the existing frameworks for performance measurement of health technology assessment have gained widespread acceptance or have been used widely to help guide allocation decisions. One reason for this lack of uptake may be that these frameworks are too complicated to be easily applied or understood. Part of the reason the frameworks are complex is that the variables that comprise the frameworks are not clearly defined. Without proper definition it is difficult to access the appropriate indicators, which in turn makes it difficult to examine the outcomes.

Other than the efforts of Jacob and McGregor (1993), existing publications on performance measurement in health technology assessment have focused on processes and not on outcomes. One reason for this is that outcomes are harder to measure in an unbiased fashion. Instead, existing performance measurement systems for health technology assessment are scattered and generated in a nonsystematic fashion. Additionally, health technology assessments must presently rely on data that are made available because it is relatively convenient to do so, such as information generated using routinely collected administrative health data (Roos et al. 2005) and registries (Tu et al. 2007); only rarely is a performance

assessment done using a primary data collection procedure (Goeree et al. 2009).

## Data Sources for Performance Measurement in Health Technology Assessment

In terms of using existing data sources for performance measurement, investigators in the United Kingdom have proposed a typology of databases according to their potential uses in the following elements of health technology assessment (Raftery et al. 2005):

- Group I databases can be used to identify both health technologies and health states; these, in turn, can be disaggregated into clinical registries, clinical administrative databases, and population-oriented databases. These databases can be used to assess effectiveness, equity, and diffusion.
- Group II databases can be used to identify health technologies but not health states. These databases can be used to assess diffusion only.
- Group III databases can be used to identify health states but not health technologies; these, in turn, can be disaggregated into adverse event reporting, disease-only registries, and health surveys. These databases have restricted scope; they are focused mainly on unintended adverse consequences of treatment or disease.

In the environmental scan that Raftery et al. conducted in England and Wales, 270 databases were identified, of which an estimated six had some potential for health technology assessment, approximately one-half of which could be assigned to group I. These investigators made important recommendations for policy that are applicable in Canada: responsibility for the strategic development of databases should be clarified (in Canada, this might be refocused on the rationalization of data collection efforts with and across health authorities); more resources should be made available; and issues associated with

coding, confidentiality, custodianship and access, maintenance of clinical support, optimal use of information technology, filling gaps, and remedying deficiencies should be clarified.

---

## Discussion

Efforts to measure and assess the performance of the Canadian health system in Canada are in the early stages, and the research agenda is enormous. Policy questions about what data to collect, and at what cost, now have equally important parallels in terms of how and when to most usefully summarize and report such information, how to integrate the information into governance and efforts to improve performance, and, ultimately, how to make wise decisions to optimize the health of the population.

Developing performance indicators can be seen as a four-step process consisting of policy, development, implementation, and evaluation phases (Ibrahim 2001). The process must address the conceptual, methodological, practical, and political considerations for developing performance measures for the Canadian health system. The lack of a conceptual framework for performance measurement in health means that research in the area is in its infancy. Methodological challenges are created by the nature of funding mechanisms in the Canadian health system and the potentially long time lags between cause and effect. Practical considerations include the daunting volume of work that would be required for greater performance measurement, including the cost and timing of such work. To date, many unresolved questions remain, such as the following: Who will decide the performance indicators? Who will measure them? How will the results of such measurements be presented? To whom and how often? Performance assessment should not be seen as a one-time effort: regular, ongoing follow-up is required. Political challenges include the different levels of governmental jurisdictions in Canada, with standards for care being laid out by the *Canada Health Act*; the federal government is responsible for protecting the health of the population by ensuring safety through the regulation of

medical products by setting and enforcing maximum reimbursement amounts for medications, whereas provision of health care is mostly a provincial and territorial responsibility. This complicated legislative and regulatory environment means that political and health reform cycles must be considered at an early stage in the development of performance measures (Roberts et al. 2008). Performance indicators would be developed and implemented much more effectively if there was cooperation between the federal, provincial, and territorial governments as well as health authorities and individual hospitals.

It is not possible for any subset of performance measures to capture all of the facets of health care that are needed by different stakeholders. What is required is a process of systematically identifying and prioritizing performance measures that will meet at least some of the needs of each stakeholder. Determining what performance measures should be used is, at the most fundamental level, an ethical question because the output must represent the different values and needs of multiple stakeholders. (Depending on the perspective, performance measures could be developed to represent different perspectives, including the following ones. First, the **utilitarian perspective** emphasizes the importance of achieving the greatest good for the greatest number. Bureaucrats require performance indicators to provide wise stewardship of the health-care system and to balance equity of access with efficient distribution. For example, some Canadian midsized cities may seek to establish catheterization laboratories to increase the speed of access to angioplasty for treating acute myocardial infarction, and provincial bureaucrats require access to information on both distributive and allocative efficiencies to balance the merit of these claims (Levy et al. 2010). Health-care professionals and hospital administrators use performance indicators to identify the functional competence of individual practitioners and organizations and to decide which technologies to adopt. Surgeons must maintain their skills to minimize operative complications, and health authority decision-makers may seek detailed information on postoperative infection rates when considering a technology for stapling versus

sewing colorectal anastomoses (when closing the opening left after removal of a colostomy bag). This information is needed when making policy decisions about purchasing and planning skills training. Second, the **libertarian perspective** emphasizes the rights of individuals to access and choose between levels of health care. For example, patients choosing between different treatments may seek detailed comparative information on the intended and unintended consequences of different treatment modalities: for example, when patients are considering angioplasty and stenting or bypass surgery for coronary artery disease, their risk preferences may be elicited if information on benefits and risks is available and synthesized in an understandable fashion. Third, the **communitarian perspective** emphasizes the need to balance the rights of individuals against the rights of the community as a whole. Organ donation (e.g., with a presumption that all persons are organ donors unless donation is actively opposed by the family), abortion and family planning services, and issues associated with the use of tobacco and intravenous drugs are all health-care matters in which communitarian values may be invoked.) Examples from the literature include performance measurement in the delivery of health-care services (Roski and Gregory 2001), health systems (Evans et al. 2001), and the health of the community (Klazinga et al. 2001).

The inherent complexities of health care, such as the diverse expertise of health-care professionals, the variety of organizational arrangements, the array of treatment protocols, and the myriad interactions between managerial and clinical activities, may necessitate that multiple outcomes be integrated in evaluating the effects of an intervention at the level of the patient, treatment, or health-care system (Sobolev et al. 2012). Table 1 provides examples of health-care performance indicators and information needs according to the type of stakeholder. This list is not intended to be exhaustive, and the categories and information needs overlap between stakeholders.

Once a performance measure comes into practice, it permeates the thinking of decision-makers

and becomes normative (Murray and Lopez 1996). In so doing, it has the possibility of influencing policy decisions, spending, and even patterns of thinking about the health system. There is a risk of overreliance on existing performance measures to the detriment of other aspects of care. For instance, in 2004, Canada's first ministers agreed to reduce waiting times in five priority areas – radiation therapy for cancer, cardiac care, diagnostic imaging, joint (hip and knee) replacement, and cataract surgery for sight restoration – by providing hospitals with cash incentives from a \$5.5-billion funding envelope. The Canadian Institute for Health Information now reports on performance measures for waiting times (CIHI 2012b). The current emphasis on these five priority areas means that other necessary procedures not considered a priority are disincentivized. In orthopedics, for example, operations such as surgery to repair feet and ankles are paid for out of a hospital's global budget and are not eligible for the incentive payments, which creates a financial incentive for hospitals to prioritize hip and knee replacements.

---

## Recommendations

A useful performance measure should always begin with detailed documentation of the indicators that constitute the measure, once definitions have been agreed upon. Given the seemingly widespread acceptance in Canada of the four dimensions discussed earlier, indicators should fall into one of these dimensions: health status, nonmedical determinants of health, health system performance, and community and health system characteristics. There should also be a clarification of responsibility for the strategic development of databases, a greater availability of resources, and clarification of issues associated with coding, confidentiality, custodianship and access, maintenance of clinical support, optimal use of information technology, filling gaps, and remedying deficiencies.

The focus of measurement must be on outcomes as well as processes, and health performance measurement should have an increased

focus on regulatory, financial, and policy measures for managing and organizing health-care systems. There should not be separate agencies for quality development, performance measurement, and service development, but rather these should be combined in a common strategy that will inform decision-making throughout the entire health-care system.

There has been, to date, a lack of focus on strategic evaluations of policy and program coherence, that is, whether policies and programs are addressing the issues and values that are most important to Canadians, such as understanding and improving determinants of health by reducing poverty and aligning healthcare spending with the principles embodied in the *Canada Health Act*.

**Acknowledgments** This chapter is reprinted from Levy, Adrian R., and Boris G. Sobolev. “*The Challenges of Measuring the Performance of Health Systems in Canada.*” *Health Care Federalism in Canada*. Eds. Katherine Fierlbeck and William Lahey. Montreal: McGill-Queen’s University Press, 2013. Print.

## References

- Alkin M. *Evaluation roots: tracing theorists’ views and influences*. Thousand Oaks: CA Sage; 2004.
- Canadian Institute for Health Information (CIHI). *Health indicators 2012*. <http://waittimes.cihi.ca/>
- Evans DB, Edejer TT, Lauer J, et al. Measuring quality: from the system to the provider. *Int J Qual Health Care*. 2001;13:439–46.
- Goeree R, Levin L, Chandra K, et al. Health technology assessment and primary data collection for reducing uncertainty in decision making. *J Am Coll Radiol*. 2009;6:332–42.
- Health Canada – Health Technology Assessment Task Group on behalf of the Federal/Provincial/Territorial Advisory Committee on Information and Emerging Technologies *Technology Strategy 1.0*. 2004. Available at <http://www.hc-sc.gc.ca/hcs-sss/pubs/ehealth-esante/2004-tech-strateg/index-eng.php>
- Hutton J, McGrath C, Frybourg JM, et al. Framework for describing and classifying decision-making systems using technology assessment to determine the reimbursement of health technologies (fourth hurdle systems). *Int J Technol Assess Health Care*. 2006;22:10–8.
- Ibrahim JE. Performance indicators from all perspectives. *Int J Qual Health Care*. 2001;13:431–2.
- Jacob R, McGregor M. Assessing the impact of health technology assessment. *Int J Technol Assess Health Care*. 1993;13:68–80.
- Klazinga N, Stronks K, Delnoij D, Verhoeff A. Indicators without a cause. Reflections on the development and use of indicators in health care from a public health perspective. *Int J Qual Health Care*. 2001;13:433–8.
- Lafortune L, Farand L, Mondou I, et al. Assessing the performance of health technology assessment organizations: a framework. *Int J Technol Assess Health Care*. 2008;24:76–86.
- Levin L, Goeree R, Sikich N, et al. Establishing a comprehensive continuum from an evidentiary base to policy development for health technologies: the Ontario experience. *Int J Technol Assess Health Care*. 2007; 23:299–309.
- Levy AR. Categorizing outcomes of health care delivery. *Clin Invest Med*. 2005;28:347–50.
- Levy AR, McGregor M. How has extracorporeal shock-wave lithotripsy changed the treatment of urinary stones in Quebec? *Can Med Assoc J*. 1995;153: 1729–36.
- Levy AR, Terashima M, Travers A. Should geographic analyses guide the creation of regionalized care models for ST-segment elevation myocardial infarction? *Open Med*. 2010;1:e22–5.
- Manitoba, Minister of Health and Healthy Living. *Manitoba’s comparable health indicator report*. Winnipeg: Manitoba Health; 2004.
- McGregor M, Brophy JM. End-user involvement in health technology assessment (HTA) development: a way to increase impact. *Int J Technol Assess Health Care*. 2005;21:263–7.
- Murray CJL, Lopez AD. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020*. Cambridge, MA: Harv Sch Public Health/WHO/World Bank; 1996; Report No. 1.
- Osbourne D, Gaebler T. *Reinventing government*. Lexington: Addison-Wesley; 1992.
- Rafferty J, Roderick P, Stevens A. Potential use of routine databases in health technology assessment. *Health Technol Assess*. 2005;9:1–iv.
- Roberts MJ, Hsiao W, Berman P, Reich M. *Getting health reform right – a guide to improving performance and equity*. Oxford, UK: Oxford University Press; 2008.
- Roos LL, Gupta S, Soodeen RA, Jebamani L. Data quality in an information-rich environment: Canada as an example. *Can J Aging*. 2005;24 Suppl 1:153–70.
- Roski J, Gregory R. Performance measurement for ambulatory care: moving towards a new agenda. *Int J Qual Health Care*. 2001;13:447–53.
- Shepherd RP. In search of a balanced Canadian federal evaluation function: getting to relevance. *Can J Program Eval*. 2012;26:1–45.
- Snowdon A, Schnarr K, Hussein A, Alessi C. *Measuring what matters: the cost vs. values of health care*. Ivey International Centre for Health Innovation. <http://sites.ivey.ca/healthinnovation/thought-leadership/white-papers/measuring-what-matters-the-cost-vs-values-of-health-care-november-2012/>
- Sobolev B, Sanchez V, Kuramoto L. Health care evaluation using computer simulation: concepts, methods and



- applications. New York: Springer; 2012; 480 pages ISBN: 978-1-4614-2232-7.
- Suissa S, Henry D, Caetano P, et al. CNODES: the Canadian network for observational drug effect studies. *Open Med.* 2012;6, e134.
- Tu JV, Bowen J, Chiu M, et al. Effectiveness and safety of drug-eluting stents in Ontario. *N Engl J Med.* 2007;357:1393–402.
- Velasco GM, Gerhardus A, Rottingen JA, Busse R. Developing health technology assessment to address health care system needs. *Health Policy.* 2010; 94:196–202.
- Weintraub WS, Grau-Sepulveda MV, Weiss JM, et al. Comparative effectiveness of revascularization strategies. *N Engl J Med.* 2012;366:1467–76.

---

**Part II**

**Methods in Health Services Research**



# Analysis of Repeated Measures and Longitudinal Data in Health Services Research

# 18

Juned Siddique, Donald Hedeker, and Robert D. Gibbons

## Contents

<b>Introduction</b> .....	406
Issues Inherent in Longitudinal Data .....	407
Historical Background .....	408
<b>Statistical Models for the Analysis of Longitudinal and Repeated Measures Data</b> .....	
Mixed-Effects Regression Models .....	409
Matrix Formulation .....	412
Covariance Pattern Models .....	413
Calculating Effect Sizes .....	414
<b>Illustrative Example: The WECare Study</b> .....	
Mixed-Effects Regression Models for Continuous Data Using the WECare Study .....	415
Curvilinear Growth Model .....	416
Covariance Pattern Models .....	418
Effect of Treatment Group on Change .....	422
<b>Extensions and Alternatives</b> .....	
Analysis of Longitudinal Data with Missing Values .....	423
Generalized Estimating Equation Models .....	425
Models for Categorical Outcomes .....	426

---

J. Siddique (✉)  
Department of Preventive Medicine, Northwestern  
University Feinberg School of Medicine, Chicago, IL, USA  
e-mail: [siddique@northwestern.edu](mailto:siddique@northwestern.edu)

D. Hedeker  
Department of Public Health Sciences, University of  
Chicago, Chicago, IL, USA  
e-mail: [hedeker@uchicago.edu](mailto:hedeker@uchicago.edu)

R. D. Gibbons  
Departments of Medicine and Public Health Sciences,  
University of Chicago, Chicago, IL, USA  
e-mail: [rdg@uchicago.edu](mailto:rdg@uchicago.edu)

Growth Mixture Models .....	429
<b>Discussion</b> .....	430
<b>References</b> .....	430

### Abstract

This chapter reviews statistical methods for the analysis of longitudinal data that are commonly found in health services research. The chapter begins by discussing issues inherent in longitudinal data and provides historical background on early methods that were used to analyze data of this type. Next, mixed-effects regression models (MRMs) and covariance-pattern models (CPMs) for longitudinal data are introduced with a focus on linear models for normally distributed outcomes. As an illustration of the use of these methods in practice, MRMs and CPMs are applied to data from the Women Entering Care (WECare) study, a longitudinal depression treatment study. Finally, extensions and alternatives to these models are briefly described. Key phrases: mixed-effects models; random-effects models; covariance-pattern models; effect sizes.

baseline severity as well. Laird and Ware (1982) showed that mixed-effects regression models could be used to perform a more complete analysis of all of the available longitudinal data under much more general assumptions regarding the missing data (i.e., missing at random). The net result was a more powerful set of statistical tools for analysis of longitudinal data that led to more powerful statistical hypothesis tests, more precise estimates of rates of change (and differential rates of change between experimental and control groups), and more general assumptions regarding missing data, for example, because of study dropout. This early work has led to considerable related advances in statistical methodology for the analysis of longitudinal data (see Hedeker and Gibbons 2006; Fitzmaurice et al. 2012; Diggle et al. 2002; Goldstein 2011; Longford 1993; Raudenbush and Bryk 2002; Singer and Willett 2003; Verbeke and Molenberghs 2000 for several excellent reviews of this growing literature).

### Introduction

In health services research, a typical study design is the longitudinal clinical trial in which patients are randomly assigned to different treatments and repeatedly evaluated over the course of the study. Since the pioneering work of Laird and Ware (1982), statistical methods for the analysis of longitudinal data have advanced dramatically. Prior to this time, a standard approach to analysis of longitudinal data principally involved using the longitudinal data to impute end-points (e.g., last observation carried forward) and then to simply discard the valuable intermediate time-point data, favoring the simplicity of analyses of change scores from baseline to study completion (or the last available measurement treated as if it was what would have been obtained had it been the end of the study), in some cases adjusting for

The following sections provide a general overview of recent advances in statistical methods for the analysis of longitudinal data. The primary focus is on linear models for continuous data. Their application is illustrated using data from the Women Entering Care (WECare) study, a longitudinal depression treatment study of low income minority women with depression. In order to motivate the use of these advanced methods, the first section discusses issues inherent in longitudinal data and some of the history of earlier methods for the analysis of longitudinal data. Next, linear mixed-effects regression models (MRMs) and covariance-pattern models (CPMs) are described in detail and applied to the WECare study. At the end of the chapter, alternatives to and extensions of linear MRMs are briefly discussed and concluding remarks are provided.

## Issues Inherent in Longitudinal Data

While longitudinal studies provide far more information than their cross-sectional counterparts, they are not without complexities. The following sections review some of the major issues associated with longitudinal data analysis.

### Heterogeneity

Particularly in health services research, individual differences are the norm rather than the exception. The overall mean response in a sample drawn from a population provides little information regarding the experience of the individual. In contrast to cross-sectional studies in which it is reasonable to assume that there are independent random fluctuations at each measurement occasion, when the same subjects are repeatedly measured over time, their responses are correlated over time, and their estimated trend line or curve can be expected to deviate systematically from the overall mean trend line. For example, behavioral and/or biological subject-level characteristics can increase the likelihood of a favorable response to a particular experimental intervention (e.g., a new pharmacologic treatment for depression), leading subjects with those characteristics to have a trend with higher slope (i.e., rate of change) than the overall average rate of change for the sample as a whole. In many cases, these personal characteristics may be unobservable, leading to unexplained heterogeneity in the population. Modeling this unobserved heterogeneity in terms of variance components that describe subject-level effects is one way to accommodate the correlation of the repeated responses over time and to better describe individual differences in the statistical characterization of the observed data. These variance components are often termed “random-effects,” leading to terms like random-effects or mixed-effects regression models.

### Missing Data

Perhaps the most important issue when analyzing data from longitudinal studies is the presence of missing data. Stated quite simply, not all subjects remain in the study for the entire length of the

study. Reasons for discontinuing the study may be differentially related to the treatment. For example, some subjects may develop side effects to an otherwise effective treatment and must discontinue the study. Alternatively, some subjects might achieve the full benefit of the study early on and discontinue the study because they feel that their continued participation will provide no added benefit. The treatment of missing data in longitudinal studies is itself a vast literature, with major contributions by Laird (1988), Little (1995), Rubin (1976), and Little and Rubin (2002) to name a few. The basic issue is that even in a randomized and well-controlled clinical trial, the subjects who were initially enrolled in the study and randomized to the various treatment conditions may be quite different from those subjects that are available for analysis at the end of the trial. If subjects “drop out” because they already have derived full benefit from an effective treatment, an analysis that only considers those subjects who completed the trial may fail to show that the treatment was beneficial relative to the control condition. This type of analysis is often termed a “completer” analysis. To avoid this type of obvious bias, investigators often resort to an analysis in which the last available measurement is carried forward to the end of the study as if the subject had actually completed the study. This type of analysis, often termed an “end-point” analysis, introduces its own set of problems in that (a) all subjects are treated equally regardless of the actual intensity of their treatment over the course of the study, and (b) the actual responses that would have been observed at the end of the study, if the subject had remained in the study until its conclusion, may in fact, be quite different than the response made at the time of discontinuation. Returning to the example of the study in which subjects discontinue when they feel that they have received full treatment benefit, an end-point analysis might miss the fact that some of these subjects may have had a relapse had they remained on treatment. Many other objections have been raised about these two simple approaches of handling missing data, which have led to more statistically reasoned approaches for

the analysis of longitudinal data with missing observations.

### **Irregularly Spaced Measurement Occasions**

It is not at all uncommon in real longitudinal studies either in the context of designed experiments or naturalistic cohorts, for individuals to vary both in the number of repeated measurements they contribute and even in the time at which the measurements are obtained. This may be due to drop-out or simply due to different subjects having different schedules of availability. While this can be quite problematic for traditional analysis of variance based approaches (leading to highly unbalanced designs which can produce biased parameter estimates and tests of hypotheses), more modern statistical approaches to the analysis of longitudinal data are all but immune to the “unbalancedness” that is produced by having different times of measurement for different subjects. Indeed, this is one of the most useful features of the regression approach to this problem, namely the ability to use all of the available data from each subject, regardless of when the data were specifically obtained.

### **Historical Background**

Existing methods for the analysis of longitudinal data are an outgrowth of two earlier approaches for repeated measures data. The first approach, the so-called repeated measures ANOVA was essentially a random intercept model that assumed that subjects could only deviate from the overall mean response pattern by a constant that was equivalent over time. A more reasonable view is that the subject-specific deviation is both in terms of the baseline response (i.e., intercept) and in terms of the rate of change over time (i.e., slope or set of trend parameters). This more general structure could not be accommodated by the repeated measures ANOVA. The random intercept model assumption leads to a compound-symmetric variance-covariance matrix for the repeated measurements in which the variances and covariances of the repeated measurements are constant over time. In general, it is common to find that variances

increase over time and covariances decrease as time-points become more separated in time. Finally, based on the use of least-squares estimation, the repeated measures ANOVA breaks down for unbalanced designs, such as those in which the sample size decreases over time due to subject discontinuation. Based on these limitations, the repeated measures ANOVA and related approaches are mostly no longer used for the analysis of longitudinal data. Mixed-effects regression models, which are described in the next section, build upon the repeated measures ANOVA framework by allowing more than just the intercept term to vary by individual in order to better capture between-subject variability. In addition, mixed-effects regression models use all available data so that not all subjects need to be measured at the same time points.

The second early approach for repeated measures data was multivariate growth curve – or MANOVA – models (Potthoff and Roy 1964; Bock 1975). The primary advantage of the MANOVA approach versus the ANOVA approach is that the MANOVA assumes a general form for the correlation of repeated measurements over time, whereas the ANOVA assumes the much more restrictive compound-symmetric form. The disadvantage of the MANOVA model is that it requires complete data. Subjects with incomplete data are removed from the analysis, leading to potential bias. In addition, both MANOVA and ANOVA models focus on comparison of group means and provide no information regarding subject-specific growth curves. Finally, both ANOVA and MANOVA models require that the time-points are fixed across subjects (either evenly or unevenly spaced) and are treated as a classification variable in the ANOVA or MANOVA model. This precludes analysis of unbalanced designs in which different subjects are measured on different occasions. Finally, software for the MANOVA approach often makes it difficult to include time-varying covariates, which are often essential to modeling dynamic relationships between predictors and outcomes. The MANOVA approach has been extended into a set of methods referred to as CPMs which also estimate the parameters of the

repeated measures variance-covariance matrix, but within a regression framework. Additionally, CPMs allow for incomplete data across time, and thus include subjects with incomplete data in the analysis. These methods are discussed in the next section.

---

## Statistical Models for the Analysis of Longitudinal and Repeated Measures Data

In an attempt to provide a more general treatment of longitudinal data, with more realistic assumptions regarding the longitudinal response process and associated missing data mechanisms, statistical researchers have developed a wide variety of more rigorous approaches to the analysis of longitudinal data. Among these, the most widely used include mixed-effects regression models (Laird and Ware 1982), and generalized estimating equation (GEE) models (Zeger and Liang 1986). Variations of these models have been developed for both discrete and continuous outcomes and for a variety of missing data mechanisms. The primary distinction between the two general approaches is that mixed-effects models are “full-likelihood” methods and GEE models are “partial-likelihood” methods. The advantage of statistical models based on partial-likelihood is that (a) they are computationally easier than full-likelihood methods, and (b) they generalize quite easily to a wide variety of outcome measures with quite different distributional forms. The price of this flexibility, however, is that partial likelihood methods are more restrictive in their assumptions regarding missing data than their full-likelihood counterparts. In addition, full-likelihood methods provide estimates of person-specific effects (e.g., person-specific trend lines) that are quite useful in understanding inter-individual variability in the longitudinal response process and in predicting future responses for a given subject or set of subjects from a particular subgroup (e.g., a county, a hospital, or a community). In the following sections attention is focused on full-likelihood methods, and partial-likelihood methods are only briefly discussed in section “[Generalized Estimating Equation Models](#).”

## Mixed-Effects Regression Models

Mixed-effects regression models (MRMs) are now widely used for the analysis of longitudinal data. Variants of MRMs have been developed under a variety of names: random-effects models (Laird and Ware 1982), variance component models (Dempster et al. 1981), multilevel models (Goldstein 1986), two-stage models (Bock 1989), random coefficient models (de Leeuw and Kreft 1986), mixed models (Longford 1987; Wolfinger 1993), empirical Bayes models (Hui and Berger 1983; Strenio et al. 1983), hierarchical linear models (Raudenbush and Bryk 1986), and random regression models (Bock 1983a, b; Gibbons et al. 1988). A basic characteristic of these models is the inclusion of random subject effects into regression models in order to account for the influence of subjects on their repeated observations. These random subject effects thus describe each person’s trend across time, and explain the correlational structure of the longitudinal data. Additionally, they indicate the degree of between-subject variation that exists in the population of subjects.

There are several features that make MRMs especially useful in longitudinal research. First, subjects are not assumed to be measured the same number of times, thus, subjects with incomplete data across time are included in the analysis. The ability to include subjects with incomplete data is an important advantage relative to procedures that require complete data across time because (a) by including all data, the analysis has increased statistical power, and (b) complete-case analysis may suffer from biases to the extent that subjects with complete data are not representative of the larger population of subjects. Because time can be treated as a continuous variable in MRMs, subjects do not have to be measured at the same time-points. This is useful for analysis of longitudinal studies where follow-up times are not uniform across all subjects. Both time-invariant and time-varying covariates can be easily included in the model. Thus, changes in the outcome variable may be due to both stable characteristics of the subject (e.g., their gender or race) as well as characteristics that change across time (e.g., life-events). Finally, whereas traditional

approaches estimate average change (across time) in a population, MRMs can also estimate change for each subject. These estimates of individual change across time can be particularly useful in longitudinal studies where a proportion of subjects exhibit change that deviates from the average trend.

To help fix ideas, consider the following simple linear regression model for the measurement  $y$  of individual  $i$  ( $i = 1, 2, \dots, N$  subjects) on occasion  $j$  ( $j = 1, 2, \dots, n_i$  occasions):

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} \times Trt_i) + \varepsilon_{ij}. \quad (1)$$

Ignoring subscripts, this model represents the regression of the outcome variable  $y$  on the independent variable time (denoted  $t$ ). The subscripts keep track of the particulars of the data, namely whose observation it is (subscript  $i$ ) and when the observation was made (the subscript  $j$ ). The independent variable  $t$  gives a value to the level of time, and may represent time in weeks, months, etc. Since  $y$  and  $t$  carry both  $i$  and  $j$  subscripts, both the outcome variable and the time variable are allowed to vary by individuals and occasions. The variable  $Trt_i$  is a binary variable that indicates the treatment assigned to individual  $i$ . When  $Trt_i$  is dummy coded as a 1 or 0, with 1 indicating membership in the treatment group, the regression coefficient  $\beta_0$  is the mean of  $y$  when  $t = 0$ ,  $\beta_1$  is the slope or rate of change for the control group, and  $\beta_2$  is the difference in slopes between the treatment and control groups.

In linear regression models, the errors  $\varepsilon_{ij}$  are assumed to be normally and independently distributed in the population with zero mean and common variance  $\sigma^2$ . This independence assumption makes the typical general linear regression model unreasonable for longitudinal data. This is because the outcomes  $y$  are observed repeatedly from the same individuals, and so it is much more reasonable to assume that errors within an individual are correlated to some degree. Furthermore, the above model posits that the change across time is the same for all individuals since the model parameters ( $\beta_0$ , the intercept or initial

level, and  $\beta_1$ , the linear change across time) do not vary by individuals except in terms of treatment assignment. For both of these reasons, it is useful to add individual-specific effects into the model that will account for the data dependency and describe differential time-trends for different individuals. This is precisely what MRMs do. The essential point is that MRMs therefore can be viewed as augmented linear regression models. Note also that here and elsewhere in this chapter, a main effect for treatment is not included in the model. That is, it is assumed that there is no difference in the expected outcomes between treatment groups at baseline. This is a reasonable assumption in a clinical trial where participants are randomized prior to receiving treatment. Alternatively, in an observational study where treatment (or exposure) is not randomized, it usually makes sense to include a main effect for treatment to account for differences between treatment groups at baseline.

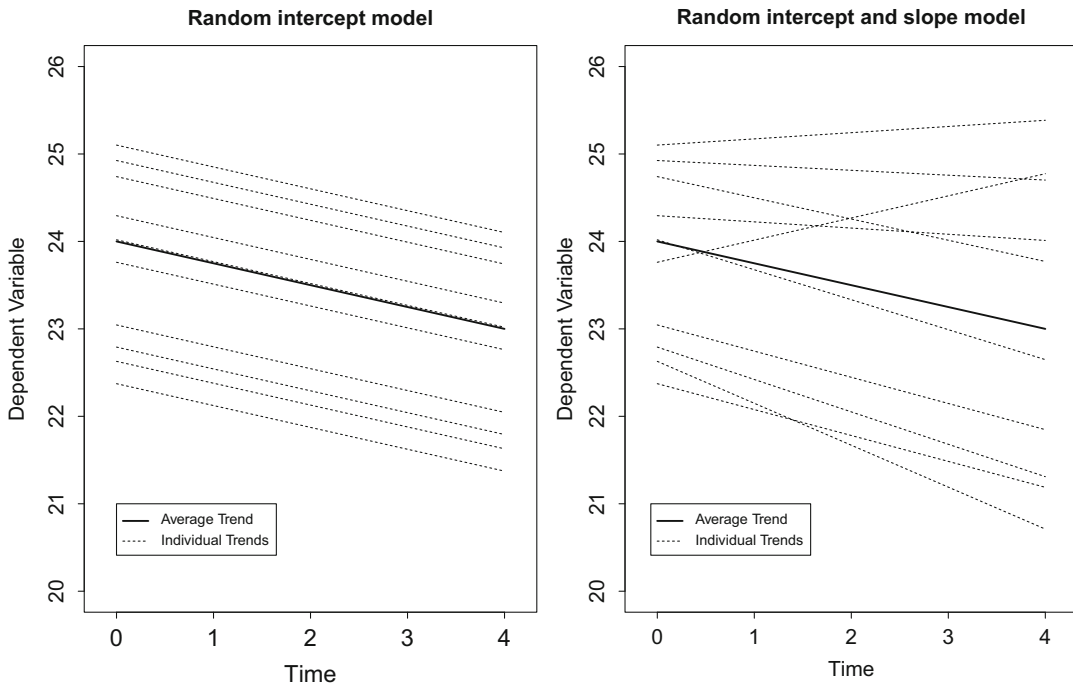
### Random Intercept Model

A simple extension of the linear regression model described in Eq. 1 is the random intercept model, which allows each subject to deviate from the overall mean response by a person-specific constant that applies equally over time:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} \times Trt_i) + v_{0i} + \varepsilon_{ij} \quad (2)$$

where  $v_{0i}$  represents the influence of individual  $i$  on his/her repeated observations. Notice that if individuals have no influence on their repeated outcomes, then all of the  $v_{0i}$  terms would equal 0. However, it is more likely that subjects will have positive or negative influences on their longitudinal data, and so the  $v_{0i}$  terms will deviate from 0. Since individuals in a sample are typically thought to be representative of a larger population of individuals, the individual-specific effects  $v_{0i}$  are treated as random effects. That is, the  $v_{0i}$  are considered to be representative of a distribution of individual effects in the population. The most common form for this population distribution is the normal distribution





**Fig. 1** Simulated longitudinal data based on a random intercept model (*left panel*) and a random intercept and slope model (*right panel*). The *solid bold line* represents

the overall population (average) trend. The *dashed lines* represent individual trends

with mean 0 and variance  $\sigma_v^2$ . In addition, the model assumes that the errors of measurement ( $\varepsilon_{ij}$ ) are conditionally independent, which implies that the errors of measurement are independent conditional on the random individual-specific effects  $v_{0i}$ . Since the errors now have the influence due to individuals removed from them, this conditional independence assumption is much more reasonable than the ordinary independence assumption associated with the linear regression model in Eq. 1. The random intercept model is depicted graphically in the left panel of Fig. 1.

As can be seen, individuals deviate from the regression of  $y$  on  $t$  in a parallel manner in this model (since there is only one subject effect  $v_{0i}$ ) (for simplicity, it is assumed the treatment effect  $\beta_2 = 0$ ). In this figure the solid line represents the population average trend, which is based on  $\beta_0$  and  $\beta_1$ . Also depicted are ten individual trends, both below and above the population (average) trend. For a given sample there are  $N$  such lines,

one for each individual. The variance term  $\sigma_v^2$  represents the spread of these lines. If  $\sigma_v^2$  is near-zero, then the individual lines would not deviate much from the population trend and individuals do not exhibit much heterogeneity in their change across time. Alternatively, as individuals differ from the population trend, the lines move away from the population trend line and  $\sigma_v^2$  increases. In this case, there is more individual heterogeneity in time-trends.

### Random Intercept and Trend Model

For longitudinal data, the random intercept model is often too simplistic for a number of reasons. First, it is unlikely that the rate of change across time is the same for all individuals. It is more likely that individuals differ in their time-trends; not everyone changes at the same rate. Furthermore, the compound symmetry assumption of the random intercept model is usually untenable for most longitudinal data. In general, measurements at points close in time tend to be more highly

correlated than measurements further separated in time. Also, in many studies subjects are more similar at baseline due to entry criteria, and change at different rates across time. Thus, it is natural to expect that variability will increase over time.

For these reasons, a more realistic MRM allows both the intercept and time-trend to vary by individuals:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} \times Trt_i) + v_{0i} + v_{1i} t_{ij} + \varepsilon_{ij}. \quad (3)$$

In this model,  $\beta_0$  is the overall population intercept,  $\beta_1$  is the overall population slope for the group with  $Trt$  coded 0, and  $\beta_2$  indicates how the population slopes vary between treatment groups (by specifically indicating how the slope for  $Trt$  coded 1 is different than the slope for  $Trt$  coded 0). In terms of the random effects,  $v_{0i}$  is the intercept deviation for subject  $i$ , and  $v_{1i}$  is the slope deviation for subject  $i$  (relative to their treatment group). As before,  $\varepsilon_{ij}$  is an independent error term distributed normally with mean 0 and variance  $\sigma^2$ . As with the random intercept model, the assumption regarding the independence of the errors is one of conditional independence, that is, they are independent conditional on  $v_{0i}$  and  $v_{1i}$ . With two random individual-specific effects, the population distribution of intercept and slope deviations is assumed to be bivariate normal  $N(0, \Sigma_v)$ , with the random-effects variance-covariance matrix given by

$$\Sigma_v = \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v0v1} \\ \sigma_{v0v1} & \sigma_{v1}^2 \end{bmatrix}. \quad (4)$$

The model described in Eq. 3 can be thought of as a personal trend or change model since it represents the measurements of  $y$  as a function of time, both at the individual  $v_{0i}$  and  $v_{1i}$  and population  $\beta_0$  and  $\beta_1$  (plus  $\beta_2$ ) levels. The intercept parameters indicate the starting point, and the slope parameters indicate the degree of change over time. The population intercept

and slope parameters represent the overall (population) trend, while the individual parameters express how subjects deviate from the population trends. The right panel of Fig. 1 represents this model graphically.

As can be seen, individuals deviate from the average trend both in terms of their intercept and in terms of their slope. As with the random intercept model, the spread of the lines around the average intercept is measured by  $\sigma_{v0}^2$  in Eq. 4. The variance of the slopes around the average trend is measured by  $\sigma_{v1}^2$  in Eq. 4. By allowing the individual slopes to vary, it is now possible for individual trends to be positive even though the overall trend is negative. The term  $\sigma_{v0v1}$  in Eq. 4 measures the association (covariance) between the random intercept and slope. When this quantity is negative, individuals with larger intercepts ( $\beta_0 + v_{i0}$ ) will have steeper slopes ( $\beta_1 + v_{i1}$ ).

## Matrix Formulation

A more compact representation of the MRM is afforded using matrices and vectors. This formulation helps to summarize statistical aspects of the model. For this, the MRM for the  $n_i \times 1$  response vector  $\mathbf{y}$  for individual  $i$  can be written as:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{v}_i + \boldsymbol{\varepsilon}_i \quad (5)$$

$\begin{matrix} n_i \times 1 & n_i \times p & p \times 1 & n_i \times r & r \times 1 & n_i \times 1 \end{matrix}$

with  $i = 1 \dots N$  individuals and  $j = 1 \dots n_i$  observations for individual  $i$ . Here,  $\mathbf{y}_i$  is the  $n_i \times 1$  dependent variable vector for individual  $i$ ,  $\mathbf{X}_i$  is the  $n_i \times p$  covariate matrix for individual  $i$ ,  $\boldsymbol{\beta}$  is the  $p \times 1$  vector of fixed regression parameters,  $\mathbf{Z}_i$  is the  $n_i \times r$  design matrix for the random effects,  $\mathbf{v}_i$  is the  $r \times 1$  vector of random individual effects, and  $\boldsymbol{\varepsilon}_i$  is the  $n_i \times 1$  residual vector.

For example, in the random intercepts and slopes MRM just considered, for a participant in the treatment group ( $Trt_i = 1$ ) the data matrices are written as

$$y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \dots \\ \dots \\ y_{in_i} \end{bmatrix} \text{ and } X_i = \begin{bmatrix} 1 & t_{i1} & t_{i1} \\ 1 & t_{i2} & t_{i2} \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ 1 & t_{in_i} & t_{in_i} \end{bmatrix} \text{ and } Z_i = \begin{bmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \dots & \dots \\ \dots & \dots \\ 1 & t_{in_i} \end{bmatrix}$$

and the population and individual trend parameter vectors are written as,

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \text{ and } v_{0i} = \begin{bmatrix} v_{0i} \\ v_{1i} \end{bmatrix}$$

respectively. The distributional assumptions about the random effects and residuals are:

$$v_i \sim N(0, \Sigma_v) \\ \varepsilon_i \sim N(0, \sigma^2 I_{n_i}).$$

As a result, it can be shown that the expected value of the repeated measures  $y_i$  is

$$E(y_i) = X_i \beta \tag{6}$$

and the variance-covariance matrix of  $y_i$  is of the form:

$$V(y_i) = Z_i \Sigma_v Z_i' + \sigma^2 I_{n_i}. \tag{7}$$

For example, with  $r = 2, n = 3$ , and

$$X_i = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 2 \end{bmatrix} \text{ and } Z_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$$

The expected value of  $y$  is

$$\begin{bmatrix} \beta_0 \\ \beta_0 + \beta_1 + \beta_2 \\ \beta_0 + 2\beta_1 + 2\beta_2 \end{bmatrix}$$

and the variance-covariance matrix equals  $\sigma^2 I_{n_i} +$

$$\begin{bmatrix} \sigma_{v_0}^2 & \sigma_{v_0}^2 + \sigma_{v_0 v_1} & \sigma_{v_0}^2 + 2\sigma_{v_0 v_1} \\ \sigma_{v_0}^2 + \sigma_{v_0 v_1} & \sigma_{v_0}^2 + 2\sigma_{v_0 v_1} + \sigma_{v_1}^2 & \sigma_{v_0}^2 + 3\sigma_{v_0 v_1} + 2\sigma_{v_1}^2 \\ \sigma_{v_0}^2 + 2\sigma_{v_0 v_1} & \sigma_{v_0}^2 + 3\sigma_{v_0 v_1} + 2\sigma_{v_1}^2 & \sigma_{v_0}^2 + 4\sigma_{v_0 v_1} + 4\sigma_{v_1}^2 \end{bmatrix}$$

which allows the variances and covariances to change across time. For example, if  $\sigma_{v_0 v_1}$  is positive, then clearly the variance increases across time. Diminishing variance across time is also possible if, for example,  $-2\sigma_{v_0 v_1} > \sigma_{v_1}^2$ . Other patterns are possible depending on the values of these variance and covariance parameters.

Models with additional random effects are also possible, as are models that allow autocorrelated errors, that is  $\varepsilon_i \sim N(0, \sigma^2 \Omega_i)$ . Here,  $\Omega$  might, for example, represent an autoregressive (AR) or moving average (MA) process for the residuals. Autocorrelated error regression models are common in econometrics. Their application within an MRM formulation is treated by Chi and Reinsel (1989) and Hedeker (1989), and extensively described in Verbeke and Molenberghs (2000). By including both random effects and autocorrelated errors, a wide range of variance-covariance structures for the repeated measures is possible. This flexibility is in sharp contrast to the traditional ANOVA models which assume either a compound symmetry structure (univariate ANOVA) or a totally general structure (MANOVA). Typically, compound symmetry is too restrictive and a general structure is not parsimonious. MRMs, alternatively, provide these two and everything in between, and so allow efficient modeling of the variance-covariance structure of the repeated measures.

### Covariance Pattern Models

An alternative to using random effects to model correlated measurements over time is to explicitly model the covariance structure through the use of CPMs. These models are a direct outgrowth of the multivariate growth curve models described in the “[Historical Background](#)” section where the covariance structure of the repeated observations

was assumed to follow a general form and all parameters of the matrix were estimated. Rather than estimating every parameter of the covariance matrix, CPMs assume the variance-covariance matrix of the repeated observations follows a specific structure. For example, the *compound symmetry* (CS) covariance model has only two parameters  $\sigma^2$  (variance) and  $\rho$  (correlation) and assumes that observations  $Y_{ij}$  have constant variance over time and the correlation between any two observations on the same subject is the same no matter how far apart those observations occurred. A variety of covariance structures exist and are available in most software packages. See Weiss (2005) for detailed descriptions of a number of different covariance matrices.

Using the matrix notation in Eq. 5, a CPM would be

$$y_i = X_i\beta + \varepsilon_i \tag{8}$$

Where instead of assuming the residuals are independent, it is assumed  $\varepsilon_i \sim N(0, \Omega_i)$ . Some common choices for  $\Omega_i$  include the previously mentioned compound symmetry where for three observations on subject  $i$  the covariance matrix is

$$V(y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

and the parameter  $\rho$  is the correlation between any two observations on the same subject. An *autoregressive* or AR(1) covariance structure also has two parameters like the compound symmetry structure but takes on a different form, namely,

$$V(y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}.$$

Thus, the farther apart two observations are in time, the lower the correlation between them (assuming  $\rho > 0$ ). In general, CPMs apply structure by specifying a specific relationship between repeated observations on the same

subject and assuming constant (homogenous) variance over time (though the homogeneity of variance can be relaxed).

When choosing a covariance model for repeated measures data, one wishes to choose the most parsimonious model that fits the data well. This can be done by first modeling the mean of observations over time and then using likelihood ratio tests as well as model fit indices such as the Bayesian Information Criteria (BIC) and the Akaike Information Criteria (AIC) to select the model that best fits the correlation and variance structure of the data. More details on methods for assessing and comparing model fit of the variance-covariance structure are described by Wolfinger (1993) and Grady and Helms (1995).

### Calculating Effect Sizes

#### Effect Sizes for Mixed-Effects Models

It is often of interest to summarize results from an intervention in terms of effect sizes. The effect size of an intervention is defined as the difference in means between the intervention and the control (or its comparator) divided by the standard deviation of the outcome. Assume a random intercept and slope MRM as in Eq. 16, that is

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} \times Trt_i) + v_{0i} + v_{1i} t_{ij} + \varepsilon_{ij}$$

To estimate the effect size of the treatment effect at time 2, begin by calculating the predicted mean for a subject in the treatment group at time 2 ( $Trt_i = 1, t_{ij} = 2$ ):

$$E(y_{ij} | Trt_i = 1, t_{ij} = 2) = \beta_0 + 2\beta_1 + 2\beta_2 \tag{9}$$

and the predicted mean for a control subject at time 2 is

$$E(y_{ij} | Trt_i = 1, t_{ij} = 2) = \beta_0 + 2\beta_1 \tag{10}$$

since the mean of the random effects and variance terms are 0. Thus the difference between the two

groups is  $2\beta_2$ . The variance for both groups at time 2 is

$$\begin{aligned} \text{Var}(y_{ij}|t_{ij} = 2) &= \text{Var}(v_{0i}) + 2^2\text{Var}(v_{1i}) \\ &\quad + 4\text{Cov}(v_{0i}, v_{1i}) \\ &\quad + \text{Var}(\varepsilon_{ij}) \end{aligned} \quad (11)$$

$$= \sigma_{v_0}^2 + 4\sigma_{v_1}^2 + 4\sigma_{v_0v_1} + \sigma_{v_0v_1} + \sigma^2. \quad (12)$$

In matrix notation, this is written as

$$\text{Var}(y_{ij}|t_{ij} = 2) = [1 \quad 2]\Sigma_v[1 \quad 2]^T + \sigma^2.$$

Thus, the effect size of the intervention at time 2 is

$$\text{Effect Size} = \frac{2\beta_2}{\sigma_{v_0}^2 + 4\sigma_{v_0}^2 + 4\sigma_{v_0v_1} + \sigma^2}.$$

### Effect Sizes for Covariance Pattern Models

Calculating effect sizes for a covariance pattern model is slightly different than for the mixed-effect model in Eq. 16 because, although it is not necessary take into account the variance of the random effects, the error terms are no longer independent. The model is

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} \times Trt_i) + \varepsilon_{ij} \quad (13)$$

where  $\varepsilon_i \sim N(0, \Omega_i)$ . As in Eqs. 9 and 10 the difference in predicted means between treatments and controls is  $2\beta_2$ . The variance for both groups at time 2 is simply

$$\text{Var}(y_{ij}|t_{ij} = 2) = \text{Var}(\varepsilon_{i3}) \quad (14)$$

$$= \sigma_{33}^2 \quad (15)$$

That is, the variance at time 2 is the third term on the diagonal of the error variance covariance matrix. Thus, the effect size of the intervention at time 2 is

$$\text{Effect Size} = \frac{2\beta_2}{\sigma_{33}^2}.$$

### Illustrative Example: The WECare Study

This section implements and extends the above methods using data from the WECare Study. The WECare Study investigated depression outcomes during a 12-month period in which 267 low-income, mostly minority, women in the suburban Washington, DC, area were treated for depression. The participants were randomly assigned to one of three groups: medication, cognitive behavioral therapy (CBT), or treatment-as-usual (TAU), which consisted of referral to a community provider. Depression was measured every month or every other month through a phone interview using the Hamilton Depression Rating Scale (HDRS). Information on ethnicity, income, number of children, insurance, and education was collected during the screening and baseline interviews. All screening and baseline data were complete except for income, with 10 participants missing data on income. After baseline, the percentage of missing interviews ranged between 24 and 38 per cent across months. Outcomes of the study were reported in Miranda et al. (2003, 2006). In these papers, the primary research question was whether the medication and CBT treatment groups had better depression outcomes as compared with the treatment-as-usual (TAU) group.

Table 1 provides mean HDRS scores, percent missing, and cumulative measurement dropout at each time point by treatment group. By month 6, approximately 84% of participants had been retained in the study. By month 12, the retention rate was 76%. The difference in dropout rates across the three treatment groups was not significant ( $p = 0.27$ ). Figure 2 provides a spaghetti plot of depression trajectories for all 267 participants (top panel) and also plots the mean depression score by treatment group (bottom panel). Two features of the data are readily apparent. First, as shown by the spaghetti plot, there is quite a bit of

**Table 1** WECare mean Hamilton Depression Rating Scale (HDRS) scores, percent missing, and cumulative measurement dropout at each time point

Mean HDRS score (% missing, % cumulative measurement dropout)			
Month of study	Medication (n = 88)	CBT (n = 90)	TAU (n = 89)
Baseline	17.95 (0%, 0%)	16.28 (0%, 0%)	16.48 (0%, 0%)
Month 1	14.00 (20%, 2%)	13.11 (27%, 6%)	12.80 (27%, 4%)
Month 2	10.74 (16%, 5%)	11.42 (27%, 7%)	11.30 (29%, 10%)
Month 3	9.60 (28%, 8%)	10.24 (36%, 9%)	13.05 (27%, 11%)
Month 4	9.54 (31%, 9%)	9.07 (38%, 13%)	11.81 (35%, 12%)
Month 5	8.62 (40%, 14%)	10.47 (34%, 14%)	11.85 (40%, 13%)
Month 6	9.17 (28%, 18%)	10.73 (33%, 14%)	11.92 (29%, 15%)
Month 8	8.07 (36%, 24%)	9.62 (30%, 17%)	11.55 (33%, 18%)
Month 10	9.04 (40%, 27%)	8.31 (31%, 20%)	10.92 (31%, 19%)
Month 12	9.71 (30%, 30%)	8.38 (24%, 24%)	10.22 (19%, 19%)

Note. CBT cognitive behavioral therapy, TAU treatment as usual

between-subject variability in the data. Second, as shown by the plots of means over time, the trends in depression scores do not appear to be linear. Instead, they appear curvilinear, with an initial strong downward trend and then a leveling off over time.

### Mixed-Effects Regression Models for Continuous Data Using the WECare Study

This section illustrates the use of MRMs for continuous data using the WECare data. The section begins by fitting the WECare data using the

random intercept and slope model in Eq. 16. Here, time corresponds to the month of the interview and takes on values from 0 to 12. As noted above, the change in depression scores across time do not appear to be linear. For now, time is treated as linear in order to demonstrate the role of diagnostics in addressing model fit. Subsequently, quadratic and cubic terms are incorporated as well as the effect of treatment group in the model. The initial model is

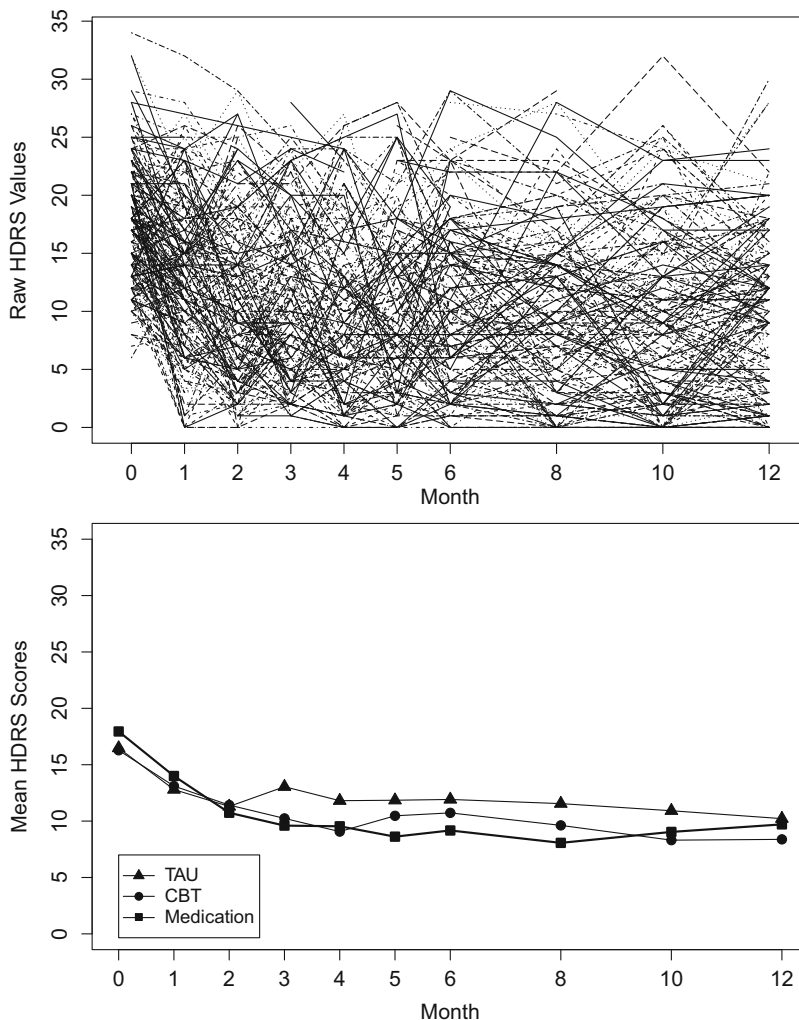
$$y_{ij} = \beta_0 + \beta_1 t_{ij} + v_{0i} + v_{1i} t_{ij} + \varepsilon_{ij} \quad (16)$$

where  $\beta_0$  is the average month 0 (baseline) HDRS level and  $\beta_1$  is the average HDRS monthly linear change. The random effect  $v_{0i}$  is the individual deviation from the average intercept, and  $v_{1i}$  is the individual deviation from the average linear change. Fitting this model yields the results given in Table 2.

Focusing first on the estimated regression parameters, this model indicates that patients start, on average, with a HDRS score of 14.08 and change by  $-0.51$  points each month. Lower scores on the HDRS reflect less depression, so patients are improving over time. The estimated HDRS score at a given month equals  $14.08 - (0.51 \times \text{month})$ . So for example, at month 2 the average depression score is  $15.64 - (1.56 \times 2) = 12.88$ . Both the intercept and slope are statistically significant ( $p < 0.0001$ ). The intercept being significant is not particularly meaningful; it just indicates that HDRS scores are different than zero at baseline. However, because the slope is significant, one can conclude that the rate of improvement is significantly different from zero in this study. On average, patients are improving across time.

For the variance and covariance terms of the random effects, there are concerns in using the standard errors in constructing Wald test statistics (estimate divided by its standard error) particularly when the population variance is thought to be near zero and the number of subjects is small (Bryk and Raudenbush 1992). This is because variance parameters are bounded; they cannot be less than zero and so using the standard normal for the sampling distribution is not reasonable. As a result, statistical significance is not indicated for

**Fig. 2** WECare depression scores over the course of the study. The *top panel* plots the raw HDRS scores for all 267 participants where each *line* represents a single individual. The *bottom panel* is plots of mean HDRS scores by treatment group. There is substantial heterogeneity in the raw scores and nonlinear trends in the means



the variance and covariance parameters in the tables. However, the magnitude of the estimates does reveal the degree of individual heterogeneity in both the intercepts and slopes. For example, while the average intercept in the population is estimated to be 14.08, the estimated population standard deviation for the intercept is 4.52 ( $=\sqrt{20.44}$ ). Similarly, the average population slope is  $-0.51$ , but the estimated population standard deviation for the slope equals 0.42, and so approximately 95% of subjects are expected to have slopes in the interval  $-0.51 \pm (1.96 \times 0.42) = -1.33$  to 0.31. That the interval includes positive slopes reflects the fact that not all subjects improve across time.

Thus, there is considerable heterogeneity in terms of patients' initial level of depression and in their change across time. Finally, the covariance between the intercept and linear trend is negative; expressed as a correlation it equals  $-0.13$ , which is small in size. This suggests that baseline depression level (i.e., intercept) is not related to the amount of linear change over time. Later on, it is seen that baseline level is positively correlated with quadratic trend – patients who are initially more depressed tend to level off over time more than patients who are less depressed at baseline. Using the estimated population intercept ( $\hat{\beta}_0$ ) and slope ( $\hat{\beta}_1$ ) one can estimate the average HDRS score at each time-point. These are displayed in

Fig. 3 along with the observed means at each time-point. As can be seen, a linear trend does not result in close agreement between the observed and estimated means. In particular, there is an initial sharp downward trend that the linear model is unable to capture. For a more quantitative assessment, the interested reader is referred to Kaplan

and George (1998) which describes use of econometric forecasting statistics to assess various forms of fit between observed and estimated means. The lack of fit of the estimated means to the observed means suggests the inclusion of curvilinear trends in the model – a point made in the next section.

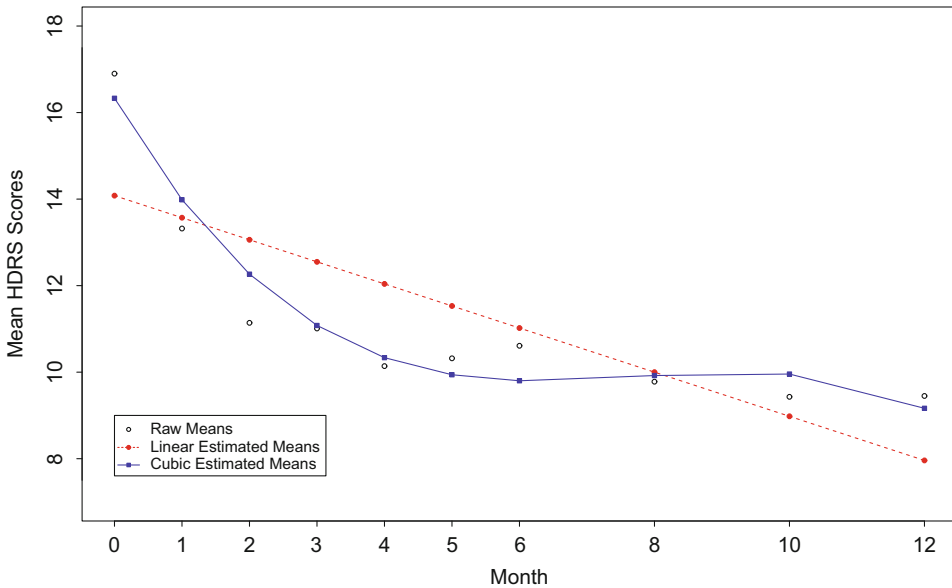
**Table 2** MRM regression results for WECare data with random intercepts and slopes and assuming linear change over time

Parameter name	Symbol	Estimate	SE	<i>t</i>	p-value
Intercept	$\beta_0$	14.08	0.33	42.30	<0.0001
Linear slope	$\beta_1$	-0.51	0.04	-12.27	<0.0001
Intercept variance	$\sigma_{v_0}^2$	20.44	2.53		
Intercept/linear slope covariance	$\sigma_{v_0v_1}$	-0.25	0.23		
Linear slope variance	$\sigma_{v_1}^2$	0.18	0.04		
Error variance	$\sigma^2$	23.67	0.88		

Note.  $-2 \log L = 12305.7$ .

### Curvilinear Growth Model

In many situations, it is too simplistic to assume that the change across time is linear. In the present example, for instance, it appears that the depression scores diminish across time in a curvilinear manner. A curvilinear trend would allow a leveling off of the improvement across time. This is clearly plausible for rating scale data, like the HDRS scores, where values below zero are impossible. Here, a curvilinear growth model is considered by adding both a quadratic and cubic term to the model. A plot of observed versus estimated means using linear and quadratic terms (not shown) did not appear to fit the observed data well so a cubic term is also added. When random cubic effects were included in



**Fig. 3** Observed and predicted WECare mean depression scores. Mean scores based on a linear or quadratic model do not fit the observed data as well as a model that includes cubic effects



the model, they were perfectly correlated with the random quadratic effects so the updated model only has random intercepts, slopes, and quadratic slopes. This produces the following model

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + v_{0i} + v_{1i} t_{ij} + v_{2i} t_{ij}^2 + \varepsilon_{ij}. \tag{17}$$

Where  $\beta_0$  is the average month 0 HDRS level,  $\beta_1$  is the average HDRS monthly linear change,  $\beta_2$  is the average HDRS monthly quadratic change, and  $\beta_3$  is the average HDRS monthly cubic change. Similarly,  $v_{0i}$  is the individual deviation from average intercept,  $v_{1i}$  is the individual deviation from average linear change, and  $v_{2i}$  is the individual deviation from average quadratic change. Fitting this model yields the results given in Table 3.

Focusing first on the estimated regression parameters, this model indicates that patients start off, on average, with an HDRS score of 16.33. Note that this value is higher than the intercept of the linear model of 14.08 and closer to the observed baseline mean of 16.9. The linear, quadratic, and cubic terms in the model are all highly significant ( $p < 0.0001$ ). The coefficient of the linear effect of month is  $-2.69$ , the coefficient of the quadratic term is  $0.36$ , and the coefficient of the cubic term is

$-0.02$ . Thus, change in depression from baseline to a given month is calculated as  $16.33 - (2.69 \times \text{month}) + (0.36 \times \text{month}^2) - (0.02 \times \text{month}^3)$ . So for example, at month 2 the average depression score is  $16.33 - (2.69 \times 2) + (0.36 \times 4) - (0.02 \times 8) = 12.26$ . Average HDRS scores at each month are displayed in Fig. 3 along with the observed means and estimated means based on a linear model. Including a cubic effect in the model does a better job capturing trends in depression scores over time. Note that at months 8 and 10, the quadratic term dominates so that mean depression scores begin to increase, and then at month 12 the cubic term dominates so that HDRS scores decrease again. Most of the improvement in depression is occurring during the first few months of the study. Because the scale for each of these terms is different (e.g., the linear effect ranges from 0 to 12, the cubic effect ranges from 0 to  $12^3 = 1728$ ), it is difficult to compare them to each other in terms of magnitude. The  $t$ -statistics provide some evidence of the magnitude and suggest that although the linear effect is strongest, all three effects contribute to the effect of time on depression symptoms.

As before, the variance and covariance terms in Table 3 provide information regarding the amount of heterogeneity in the data. The 95% confidence interval for subject-specific intercepts is  $16.33 \pm 3.87$  and the 95% confidence interval

**Table 3** MRM results for the WECare data with cubic trends and random intercept, slope, and quadratic slopes effects

Parameter name	Symbol	Estimate	SE	$t$	p-value
Intercept	$\beta_0$	16.33	0.34	47.99	<0.0001
Month	$\beta_1$	-2.69	0.22	-12.03	<0.0001
Month <sup>2</sup>	$\beta_2$	0.36	0.05	7.97	<0.0001
Month <sup>3</sup>	$\beta_3$	-0.015	0.003	-6.12	<0.0001
Intercept variance	$\sigma_{v_0}^2$	15.02	2.38		
Intercept/linear slope covariance	$\sigma_{v_0v_1}$	0.67	0.69		
Linear slope variance	$\sigma_{v_1}^2$	1.55	0.36		
Intercept/quadratic slope covariance	$\sigma_{v_0v_2}$	-0.10	0.05		
Linear/quadratic slope covariance	$\sigma_{v_1v_2}$	-0.11	0.03		
Quadratic slope variance	$\sigma_{v_2}^2$	0.01	0.002		
Error variance	$\sigma^2$	19.75	0.79		

Note.  $-2 \log L = 12095.1$

for the subject-specific quadratic terms in the model includes zero reflecting the fact that there is considerable heterogeneity in terms of patients' initial level of depression and in their changes across time.

Finally, the covariance between the linear effect and the quadratic effect is negative; expressed as a correlation it equals  $-0.94$ , which is very high. This is partially due to multicollinearity but also suggests that those patients who make the most initial gains (i.e., steep slopes) tend to level off at a greater rate (i.e., greater quadratic effects) than patients who have flatter slopes in the early stages of the study. An alternative explanation is that of a floor effect due to the HDRS rating scale. Simply put, once patients achieve low depression scores they no longer have room to keep improving and thus tend to level off.

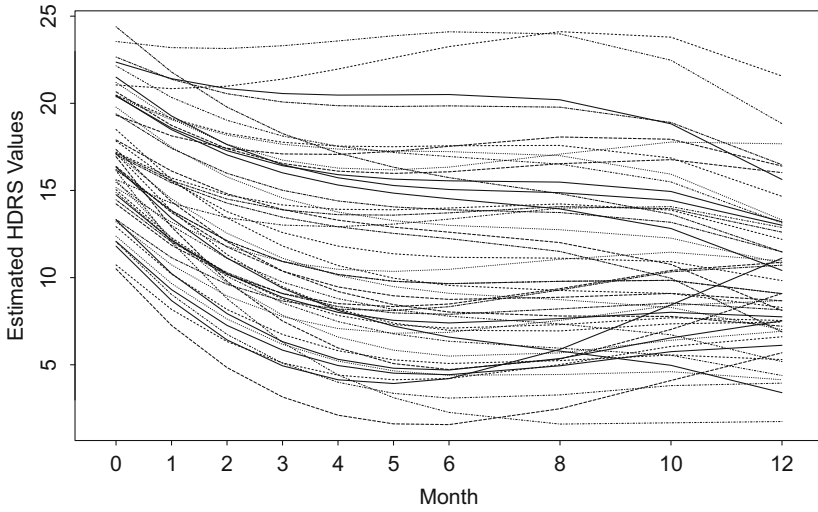
An interesting question, at this point, is whether it is necessary to include random effects for the linear and quadratic terms or whether a less complicated model is sufficient. Fitting the more restrictive model with random intercepts and linear terms (not shown) yields  $-2 \log L = 12155.6$ . Note that both models still include fixed effects for linear slope, quadratic slope, and cubic slope. Because these are nested models, they can be compared using a likelihood-ratio test. For this, one compares the difference in model deviance values (i.e.,  $-2 \log L$ ) to a chi-square distribution, where the degrees of freedom equals the number of parameters set equal to zero in the more restrictive model. Comparing the full model to the restricted model with only random intercepts and slopes,  $\chi^2_3 = 12155.6 - 12095.1 = 60.5$ ,  $p < 0.0001$  for  $H_0 : \sigma_{v_0 v_2} = \sigma_{v_1 v_2} = \sigma_{v_2}^2 = 0$ . It should be noted that the use of the likelihood ratio test for this purpose also suffers from the variance boundary problem mentioned above (Verbeke and Molenberghs 2000). Based on simulation studies it can be shown that the likelihood-ratio test is too conservative (for testing null hypotheses about variance parameters), namely, it does not reject the null hypothesis often enough. This would then lead to accepting

a more restrictive variance-covariance structure than is correct. As noted by Berkhof and Snijders (2001), this bias can largely be corrected by dividing the  $p$ -value obtained from the likelihood-ratio test (of variance terms) by two. In the present case it doesn't really matter, but this modification yields  $p < 0.0001/2 = 0.00005$ . Thus, there is clear evidence that the assumption of only random intercepts and linear slopes is rejected, and the inclusion of the random quadratic slopes is necessary.

In addition to plots of the overall means over time, estimates of the individual trends, based on the random effects  $\hat{v}_{0i}$ ,  $\hat{v}_{1i}$  and  $\hat{v}_{2i}$  are often of interest. Figure 4 contains a plot of the individual trend estimates from this model. These are obtained by calculating  $\hat{y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 t_{ij} + \hat{\beta}_2 t_{ij}^2 + \hat{\beta}_3 t_{ij}^3 + \hat{v}_{0i} + \hat{v}_{1i} t_{ij} + \hat{v}_{2i} t_{ij}^2$  for  $t = 0, 1, \dots, 12$ , and then connecting the time point estimates for each individual. For clarity, 50 of the 267 WECare participants were randomly selected to display in Fig. 4.

The plot makes apparent the wide heterogeneity in trends across time, as well as the increasing variance in HDRS scores across time. Some individuals have initial accelerating downward trends suggesting immediate improvement and then a leveling off over time, while others appear to have more modest improvements and then perhaps a slight worsening of symptoms. Some individuals even have positive trends indicating a worsening of their depressive symptoms across time. This is not too surprising given that not all depression interventions work for everyone. At the end of this chapter, growth mixture models are briefly introduced which attempt to classify individuals into discrete latent classes based on the shape of their trajectories.

It is worth noting that the estimates of the individual trends presented in Fig. 4 are empirical Bayes (EB) estimates, which reflect a compromise between an estimate based solely on an individual's data and an estimate for the population of interest. Thus, they are not equivalent to ordinary least squares (OLS) estimates (i.e.,



**Fig. 4** Subject-specific estimated WECare HDRS means over time based on a model with cubic fixed effects and random intercept, slope, and quadratic slope effects. For clarity, only a random sample of 50 participants is displayed

fitting a regression line for each participant separately) which would only rely upon an individual’s data. An important advantage of EB estimates relative to OLS estimates is that they are not as prone to the undue influence of outliers. This is especially true when an individual has few measurements by which to base these estimates on. Because of this, the EB estimates are said to be *shrunk to the mean*, where the mean of the random effects equals zero in the population. The degree of shrinkage depends on the number of measurements an individual has. Thus, if a subject has few measurements, then

the EB estimate will be smaller (in absolute value) than the corresponding OLS estimate. Alternatively, if the subject has many measurements across time, then the EB and OLS estimates would be very similar. These EB estimates are readily available from most MRM software programs.

Finally, the fit of the observed variance-covariance matrix of the repeated measures is addressed. These are calculated based on the pairwise data for the covariances and the available data for each of the variances. The observed variance-covariance matrix is

$$= \begin{matrix} V(y) \\ \left[ \begin{matrix} 26.87 & & & & & & & & & & \\ 16.52 & 42.64 & & & & & & & & & \\ 17.19 & 30.54 & 49.54 & & & & & & & & \\ 12.03 & 22.64 & 28.47 & 47.00 & & & & & & & \\ 12.65 & 28.68 & 29.47 & 32.39 & 52.74 & & & & & & \\ 9.37 & 21.22 & 20.28 & 24.95 & 30.09 & 49.88 & & & & & \\ 9.10 & 21.82 & 29.03 & 26.73 & 29.34 & 28.15 & 49.75 & & & & \\ 7.32 & 23.62 & 23.98 & 26.49 & 24.74 & 27.88 & 31.67 & 50.83 & & & \\ 7.93 & 22.11 & 22.79 & 22.69 & 26.19 & 23.96 & 27.05 & 33.33 & 53.32 & & \\ 5.48 & 17.17 & 17.83 & 18.78 & 21.53 & 22.44 & 22.86 & 30.53 & 30.97 & 50.14 & \end{matrix} \right] \end{matrix}$$



**Table 4** Fit indices for various covariance patterns fit to the WECare data

Covariance pattern	No. of parameters	$-2 \log L$	AIC	BIC	p-value versus unstructured
Autoregressive/oving Average	3	12115.6	12129.6	12129.6	0.0001
MRM	7	12095.1	12117.1	12156.5	<0.0001
Toeplitz	10	12108.5	12136.5	12186.7	<0.0001
Heterogeneous Toeplitz	19	12079.3	12125.3	12207.8	0.004
Factor analytic (2)	29	12064.8	12130.8	12249.1	0.006
Factor analytic (1)	20	12130.7	12178.7	12264.8	<0.0001
Heterogeneous CS	11	12202.1	12232.1	12286.0	<0.0001
Autoregressive (1)	2	12257.1	12269.1	12290.7	<0.0001
Heterogeneous Autoregressive(1)	11	12227.8	12257.8	12311.7	<0.0001
Antependence	19	12209.6	12255.6	12338.1	<0.0001
Unstructured	55	12016.9	12134.9	12346.6	NA

As can be seen, while none of the covariance patterns provide a statistically similar fit to the data than the unstructured covariance in terms of a likelihood ratio test, the MRM with random intercepts, slopes, and quadratic slopes has the smallest AIC and the second smallest BIC among all the models. BIC imposes a high penalty on models with many parameters so it is not surprising that the unstructured covariance has the worst BIC. For this reason, Fitzmaurice et al. (2012) recommend against use of BIC for model selection of (co)variance structure. AIC is more useful for comparing models that are not nested when a likelihood ratio test is not appropriate. Still, Table 4 suggests that the MRM provides a relatively parsimonious fit to the WECare data. Perhaps a model with both random subject effects and autocorrelated errors could be considered here.

### Effect of Treatment Group on Change

At this point, the effect of treatment group on depression outcomes is examined by augmenting the model to include interactions of time with treatment group. Setting the TAU group as the reference group, two new variables are created:  $MEDS_i$  which equals 1 if participant  $i$  was randomized to antidepressants and 0 otherwise; and  $CBT_i$  which equals 1 if participant

$i$  was randomized to CBT and 0 otherwise. The mixed-effects model is now

$$\begin{aligned}
 y_{ij} = & \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 \\
 & + \beta_4 t_{ij} MEDS_i + \beta_5 t_{ij}^2 MEDS_i \\
 & + \beta_6 t_{ij}^3 MEDS_i + \beta_7 t_{ij} CBT_i \\
 & + \beta_8 t_{ij}^2 CBT_i + \beta_9 t_{ij}^3 CBT_i + v_{0i} \\
 & + v_{1i} t_{ij} + v_{2i} t_{ij}^2 + \varepsilon_{ij}.
 \end{aligned} \quad (18)$$

The parameters  $v_{0i}$ ,  $v_{1i}$ ,  $v_{2i}$ , and  $\varepsilon_{ij}$  have the same interpretation as in section “[Curvilinear Growth Model](#).”

The unstructured covariance model is

$$\begin{aligned}
 y_{ij} = & \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 \\
 & + \beta_4 t_{ij} MEDS_i + \beta_5 t_{ij}^2 MEDS_i \\
 & + \beta_6 t_{ij}^3 MEDS_i + \beta_7 t_{ij} CBT_i \\
 & + \beta_8 t_{ij}^2 CBT_i + \beta_9 t_{ij}^3 CBT_i + \Sigma_{ij}.
 \end{aligned} \quad (19)$$

where  $\Sigma_{ij}$  represents the  $j$ th entry on the diagonal of the  $n_i \times n_i$  unstructured covariance matrix for subject  $i$ .

Equations 18 and 19 highlight the difference between a mixed-effects model and a covariance pattern model. The mixed-effects model partitions the variance of  $y_{ij}$  into between-subject variance (estimated via the random effects) and within-subject variance (estimated

via the error term). The covariance pattern model does not make this distinction. When the focus of inference is on the fixed-effects in the model, this distinction is less important. In other settings, where there is interest in determining the degree of subject heterogeneity and/or examining individual subject trends, it may be more important.

In both models,  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  represent the linear, quadratic, and cubic effects of time for the TAU group which has been chosen to be the reference group. The coefficients  $\beta_4$ ,  $\beta_5$ , and  $\beta_6$  are the time by Medication group interactions with the three time effects and indicate the difference in time trends between the Medication and TAU group. The coefficients  $\beta_7$ ,  $\beta_8$ , and  $\beta_9$  are the time by CBT group interactions and indicate the difference in time trends between the CBT and TAU group. A likelihood-ratio test can be used to test the null hypothesis that there is no effect of Medication versus TAU (i.e.,  $\beta_4$ ,  $\beta_5$ , and  $\beta_6$  are zero) by fitting model 18 with and without the time by Medication interaction effects. This yields  $\chi^2_3 = 12091.2 - 12067.1 = 24.1$ , which has a p-value  $< 0.0001$ . A similar test for the effect of the CBT group yields  $\chi^2_3 = 12074.0 - 12067.1 = 6.9$  which has a p-value 0.075. In model 19, the corresponding likelihood ratio tests are  $\chi^2_3 = 12012.6 - 11988.9 = 23.7$  ( $p < 0.0001$ ) for the Medication group and  $\chi^2_3 = 11995.2 - 11988.9 = 6.3$  ( $p = 0.10$ ) for the CBT group. Thus, both models give similar results regarding the significance of the Medication and CBT treatment groups versus the TAU group.

Table 5 reports the results from fitting the model described in Eq. 18 to the WECare data and Table 6 reports the results from the model described in Eq. 19. As can be seen, the estimates from both models are similar.

It is interesting to note that among the time by treatment interactions, only the interaction of Medication with linear time is significant. This suggests that the effect of the Medication intervention takes place early on in the study, during the initial sharp decline in depression scores. This is clearer in Fig. 5, which displays the estimated means at each time

point by treatment group using the parameter estimates in Table 5. Even though the other treatment by time interactions are not significant, their magnitude is large enough such that the three different growth curves have very different shapes.

Once it has been established that the Medication intervention (but not the CBT intervention) produces significantly different outcomes than the TAU group (via likelihood ratio tests), it may be of interest to estimate the mean HDRS scores of these interventions at specific time points, their differences, and their corresponding effect sizes. This can be done using the methods described in section “Calculating Effect Sizes.”

For example, to calculate the effect size of the Medication intervention versus the TAU intervention at month 6, one begins by estimating the mean HDRS scores for both groups at month 6. For both Eqs. 18 and 19 the difference in mean HDRS scores at month 6 between the Medication and TAU interventions is  $6\beta_4 + 6^2\beta_5 + 6^3\beta_6$ . The variance at month 6 in the mixed-effects model is

$$\begin{aligned} \text{Var}(y_{ij}|t_{ij}=6) &= \text{Cov}(v_{0i} + 6v_{1i} + 6^2v_{2i} + \varepsilon_{i6}, v_{0i} + 6v_{1i} + 6^2v_{2i} + \varepsilon_{i6}) \\ &= \text{Var}(v_{0i}) + 2\text{Cov}(v_{0i}, 6v_{1i}) + 2\text{Cov}(v_{0i}, 6^2v_{2i}) \\ &\quad + \text{Var}(6v_{1i}) + 2\text{Cov}(6v_{1i}, 6^2v_{2i}) + \text{Var}(6^2v_{2i}) \\ &\quad + \text{Var}(\varepsilon_{i6}) \\ &= \sigma_{v_0}^2 + 12\sigma_{v_1v_2} + 72\sigma_{v_0v_2} + 36\sigma_{v_1}^2 + 432\sigma_{v_1v_2} \\ &\quad + 1296\sigma_{v_2}^2 + \sigma^2 \\ &= 54.57. \end{aligned} \tag{20}$$

In matrix notation, this is written as

$$\text{Var}(y_{ij}|t_{ij}=6) = [1 \ 6 \ 6^2]\Sigma_v[1 \ 6 \ 6^2]^T + \sigma^2.$$

Using the estimates from Table 5, the effect size based on the mixed-effects model is

$$\text{Month 6 effect size} = \frac{-4.39}{\sqrt{54.57}} = -0.60.$$

For the covariance pattern model, the variance at month 6 is simply the seventh term on the

**Table 5** Results from a mixed-effect regression model fit to the WECare data

Parameter name	Symbol	Estimate	SE	<i>t</i>	p-value
Intercept	$\beta_0$	16.330	0.34	47.96	<0.0001
Month	$\beta_1$	-2.081	0.36	-5.84	<0.0001
Month <sup>2</sup>	$\beta_2$	0.325	0.07	4.38	<0.0001
Month <sup>3</sup>	$\beta_3$	-0.016	0.00	-3.88	0.0001
Month*MEDS	$\beta_4$	-1.356	0.48	-2.8	0.005
Month <sup>2</sup> *MEDS	$\beta_5$	0.099	0.10	0.96	0.34
Month <sup>3</sup> *MEDS	$\beta_6$	0.001	0.01	0.11	0.92
Month*CBT	$\beta_7$	-0.424	0.49	-0.87	0.38
Month <sup>2</sup> *CBT	$\beta_8$	-0.005	0.10	-0.05	0.96
Month <sup>3</sup> *CBT	$\beta_9$	0.002	0.01	0.39	0.70
Intercept variance	$\sigma_{v_0}^2$	15.062	2.387		
Intercept, slope covariance	$\sigma_{v_0v_1}$	1.052	0.658		
Slope variance	$\sigma_{v_1}^2$	1.182	0.322		
Intercept, quadratic slope covariance	$\sigma_{v_0v_2}$	-0.134	0.050		
Slope, quadratic slope covariance	$\sigma_{v_1v_2}$	-0.078	0.025		
Quadratic slope variance	$\sigma_{v_2}^2$	0.006	0.002		
Error variance	$\sigma^2$	19.741	0.792		

Note.  $-2 \log L = 12067.1$

**Table 6** Results from a covariance-pattern model fit to the WECare data

Parameter name	Symbol	Estimate	SE	<i>t</i>	p-value
Intercept	$\beta_0$	16.817	0.31	54.22	<0.0001
Month	$\beta_1$	-2.118	0.39	-5.49	<0.0001
Month <sup>2</sup>	$\beta_2$	0.319	0.08	4.06	<0.0001
Month <sup>3</sup>	$\beta_3$	-0.015	0.00	-3.61	0.0004
Month*MEDS	$\beta_4$	-1.497	0.53	-2.81	0.005
Month <sup>2</sup> *MEDS	$\beta_5$	0.138	0.11	1.25	0.21
Month <sup>3</sup> *MEDS	$\beta_6$	-0.002	0.01	-0.29	0.77
Month*CBT	$\beta_7$	-0.438	0.54	-0.82	0.41
Month <sup>2</sup> *CBT	$\beta_8$	0.009	0.11	0.08	0.94
Month <sup>3</sup> *CBT	$\beta_9$	0.001	0.01	0.18	0.86

Note.  $-2 \log L = 11988.9$

diagonal of the covariance matrix which is equal to 49.52. Thus, using parameter estimates from Table 6, the effect size based on the covariance pattern model is

$$\text{Month 6 effect size} = \frac{-4.45}{\sqrt{49.52}} = -0.62.$$

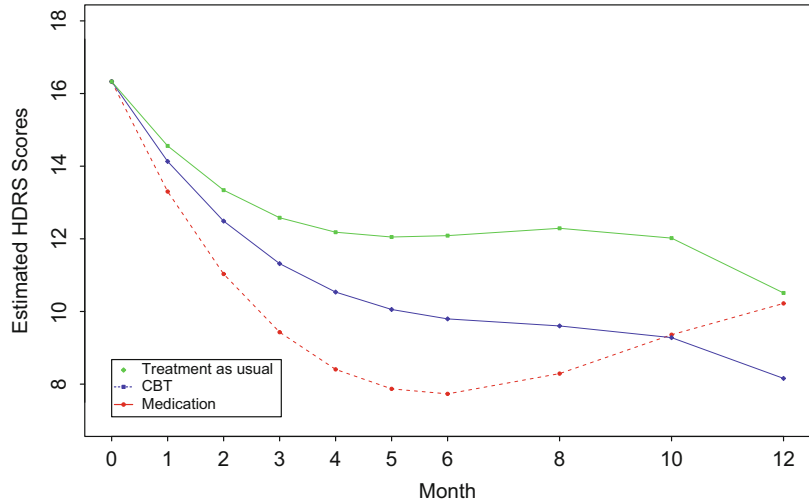
Both effect sizes are similar and suggest a medium effect of the Medication intervention.

## Extensions and Alternatives

### Analysis of Longitudinal Data with Missing Values

While longitudinal designs have many benefits, measuring participants repeatedly over time also leads to repeated opportunities for missing data, either through failure to answer certain items, missed assessments, or permanent withdrawal

**Fig. 5** Estimated WECare HDRS means over time by treatment group



from the study. As noted above, the treatment of missing data in longitudinal studies is itself a vast literature. An important consideration when drawing inferences from longitudinal data is the reason for the missing data, also referred to as the missing data mechanism (Rubin 1976). Most of the methods described in this chapter – with the exception of GEE methods – provide valid estimates under the assumption that the missing data mechanism is missing at random (MAR) as described by Rubin (1976), where the probability that a value is missing does not depend on unobserved information such as the value itself. When data are not missing at random (NMAR), that is, the probability that a value is missing *does* depend on unobserved information, it is necessary to model both the outcome as well as the missing data mechanism itself.

NMAR is an untestable assumption since the mechanism by definition depends on unobserved information. Thus, it is difficult to identify those situations where one is dealing with data that are NMAR. However, one situation where data that are NMAR is often a concern is participant drop-out where subjects withdraw from a study and are never heard from again. In this situation, two common approaches for handling drop-outs in longitudinal designs are pattern-mixture models and shared-parameter models. In pattern-mixture models, the data are stratified by the different

dropout patterns with distinct model parameters for each stratum. Marginal estimates across the patterns can be derived as a weighted average across pattern specific estimates (Little 1995) or by using multiple imputation (Demirtas and Schafer 2003). Shared-parameter models are identified by using common random effects to relate the response with the missing-data indicator (Daniels and Hogan 2000; Guo et al. 2004).

Limitations due to space prevent an in-depth discussion of this topic. Instead, readers are referred to recent review articles including Kenward and Molenberghs (1999), Siddique et al. (2008), and Ibrahim and Molenberghs (2009). Also the books by Little and Rubin (2002), Fitzmaurice et al. (2012), Hedeker and Gibbons (2006), and Daniels and Hogan (2008) which contain useful material on this topic.

## Generalized Estimating Equation Models

In the 1980s, alongside development of MRMs and CPMs for incomplete longitudinal data, generalized estimating equations (GEE) models were developed (Liang and Zeger 1986; Zeger and Liang 1986). Essentially, GEE models extend generalized linear models (GLMs) to the case of correlated data. This class of models has become



very popular – especially for the analysis of categorical and count outcomes – though they can be used for continuous outcomes as well. One difference between GEE models and MRMs is that GEE models are based on quasi-likelihood estimation, and so the full likelihood of the data is not specified. GEE models are termed marginal models, and they model the regression of  $y$  on  $x$  and the within subject dependence (i.e., the association parameters) separately. The term “marginal” in this context indicates that the model for the mean response depends only on the covariates of interest, and not on any random effects or previous responses. In terms of missing data, GEE assumes that the missing data are missing completely at random (MCAR) where the probability that a value is missing does not depend either on observed or missing values. This is a stricter (and possibly less realistic) assumption than that assumed by the models employing full-likelihood estimation which assume missing data are MAR.

Conceptually, GEE reproduces the marginal means of the observed data, even if some of those means have limited information because of subject drop-out. Standard errors are adjusted (i.e., inflated) to accommodate the reduced amount of independent information produced by the correlation of the repeated observations over time. By contrast, mixed-effects models use the available data from all subjects to model temporal response patterns that would have been observed had the subjects all been measured to the end of the study. Because of this, estimated mean responses at the end of the study can be quite different for GEE versus MRM, if the future observations are related to the measurements that were made during the course of the study. If the available measurements are not related to the missing measurements (e.g., following dropout), GEE and MRM will produce quite similar estimates. This is the fundamental difference between GEE and MRM, that is, the assumption that the missing data are dependent on the observed responses for a given subject during that subject’s participation in the study. It is hard to imagine that a subject’s responses that would have been obtained following dropout would be

independent of their observed responses during the study. This leads to a preference for full-likelihood approaches over quasi or partial likelihood approaches, and MRM over GEE, at least for longitudinal data. There is certainly less of an argument for a preference for data that are only clustered (e.g., providers nested within clinics), in which case advantages of MAR over MCAR are not as germane.

A basic feature of GEE models is that the joint distribution of a subject’s response vector  $y_i$  does not need to be specified. Instead, it is only the marginal distribution of  $y_{ij}$  at each time point that needs to be specified. To clarify this further, suppose that there are two time-points and suppose that the outcome is a continuous normal random variable. GEE would only require us to assume that the distribution of  $y_{i1}$  and  $y_{i2}$  are two univariate normals, rather than assuming that  $y_{i1}$  and  $y_{i2}$  form a (joint) bivariate normal distribution. Thus, GEE avoids the need for multivariate distributions by only assuming a functional form for the marginal distribution at each time-point. This leads to a simpler quasi-likelihood approach for estimating the model parameters, rather than the full-likelihood approach of the MRM and CPM. The disadvantage, as mentioned above, is that because a multivariate distribution is not specified for the response vector, the assumption for the missing data is more stringent for the GEE than the full-likelihood estimated MRMs and CPMs. A complete treatment of GEE can be found in Hardin and Hilbe (2012).

## Models for Categorical Outcomes

Reflecting the usefulness of mixed-effects modeling and the importance of categorical outcomes in many areas of research, generalization of mixed-effects models for categorical outcomes has been an active area of statistical research. For dichotomous response data, several approaches adopting either a logistic or probit regression model and various methods for incorporating and estimating the influence of the random effects have been developed (Gibbons 1981; Stiratelli et al. 1984; Wong and Mason 1985; Gibbons and Bock 1987;

Conaway 1989; Goldstein 1991). Here, briefly described is a mixed-effects logistic regression model for the analysis of binary data. Extensions of this model for analysis of ordinal, nominal, and count data are described in detail by Hedeker and Gibbons (2006).

To set the notation, let  $i$  denote individuals and let  $j$  denote the repeated measurement occasions within each individual. Assume that there are  $i = 1, \dots, N$  individuals and  $j = 1, \dots, n_i$  measurement occasions nested within each individual. Let  $Y_{ij}$  be the value of the dichotomous outcome variable, coded 0 or 1. The logistic regression model is written in terms of the log odds (i.e., the logit) of the probability of a response, denoted  $p_{ij}$ . Considering first a random-intercept model, augmenting the logistic regression model with a single random effect yields:

$$\ln \left[ \frac{p_{ij}}{1 + p_{ij}} \right] = \mathbf{x}'_{ij}\boldsymbol{\beta} + v_i \tag{21}$$

where  $x_{ij}$  is the  $(p + 1) \times 1$  covariate vector (includes a 1 for the intercept),  $\boldsymbol{\beta}$  is the  $(p + 1) \times 1$  vector of unknown regression parameters, and  $v_{i0}$  is the random subject effect. These random effects are assumed to be distributed in the population as  $N(0, \sigma_v^2)$ . For convenience and computational simplicity, in models for categorical outcomes the random effects are typically expressed in standardized form. For this,  $v_{0i} = \sigma_v \theta_i$  and the model is given as:

$$\ln \left[ \frac{p_{ij}}{1 + p_{ij}} \right] = \mathbf{x}'_{ij}\boldsymbol{\beta} + \sigma_v \theta_i. \tag{22}$$

Notice that the random-effects variance term (i.e., the population standard deviation  $\sigma_v$ ) is now explicitly included in the regression model. Thus, it and the regression coefficients are on the same scale, namely, in terms of the log-odds of a response.

The model can also be expressed in terms of a latent continuous variable  $y$ , with the observed dichotomous version  $Y$  being a manifestation of the unobserved continuous  $y$ . Here, the model is written as:

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \sigma_v \theta_i + \varepsilon_{ij} \tag{23}$$

in which case the error term  $\varepsilon_{ij}$  follows a standard logistic distribution under the logistic regression model (or a standard normal distribution under the probit regression model). This representation helps to explain why the regression coefficients from a mixed-effects logistic regression model do not typically agree with those obtained from a fixed-effects logistic regression model, or for that matter from a GEE logistic regression model which has regression coefficients that agree in scale with the fixed-effects model. In the mixed model, the conditional variance of the latent  $y$  given  $\mathbf{x}$  equals  $\sigma_v^2 + \sigma_\varepsilon^2$ , whereas in the fixed-effects model this conditional variance equals only the latter term  $\sigma_\varepsilon^2$  (which equals either  $\pi^2/3$  or 1 depending on whether it is a logistic or probit regression model, respectively). As a result, equating the variances of the latent  $y$  under these two scenarios yields:

$$\boldsymbol{\beta}_M \approx \sqrt{\frac{\sigma_v^2 + \sigma_\varepsilon^2}{\sigma_\varepsilon^2}} \boldsymbol{\beta}_F$$

where  $\boldsymbol{\beta}_F$  and  $\boldsymbol{\beta}_M$  represent the regression coefficients from the fixed-effects and (random-intercepts) mixed-effects models, respectively. In practice, Zeger et al. (1988) have found that  $(15/16)^2 \pi^2/3$  works better than  $\pi^2/3$  for  $\sigma_\varepsilon^2$  in equating results of logistic regression models.

Several authors have commented on the difference in scale and interpretation of the regression coefficients in mixed-models and marginal models, like the fixed-effects and GEE models (Neuhaus et al. 1991; Zeger et al. 1988). Regression estimates from the mixed model have been termed “subject-specific” to reinforce the notion that they are conditional estimates, conditional on the random (subject) effect. Thus, they represent the effect of a regressor on the outcome controlling for, or holding constant, the value of the random subject effect. Alternatively, the estimates from the fixed-effects and GEE models are “marginal” or “population-averaged” estimates which indicate the effect of a regressor averaging over the population of subjects. This difference of scale

and interpretation only occurs for nonlinear regression models like the logistic regression model. For the linear model this difference does not exist.

## Growth Mixture Models

A frequent characteristic of depression clinical trials (such as the WECare study) is that outcomes over time are subject to considerable between-subject heterogeneity due to the fact that patients often follow different trajectories over time. Some participants may see immediate gains, only to relapse at a later date, while others will improve gradually overtime. Some participants will not improve at all. When comparing the effectiveness of different treatments, it is important to identify and take into account these different trajectories because the effectiveness of an intervention may depend on the trajectory class of the participants. Despite the fact that heterogeneity of outcomes is common in depression studies, most analyses such as mixed-effects regression models assume that all individuals are drawn from a single population with common population parameters (Muthén 2004). That is, they assume that all individual trajectories vary around a single mean trajectory. This assumption goes counter to clinical observations and empirical data where variation in trajectory shapes is routinely observed. When individuals follow several different trajectory shapes, conventional repeated measures modeling may lead to a distorted assessment of treatment effects.

Growth mixture modeling (Muthén and Shedden 1999; Muthén et al. 2002; Xu and Hedeker 2002) relaxes the single population assumption to allow for parameter differences across several unobserved populations. Instead of considering individual variation around a single trajectory, a growth mixture model (GMM) allows different classes of individuals to vary around several different trajectories. In this way, growth mixture modeling may do a better job of capturing between-subject variability because it does not require that all individuals follow the same average trajectory over time.

Once multiple trajectories have been identified, analyses can be performed to predict trajectory class as a function of other covariates. This approach is particularly useful in randomized trials because it may suggest that for some groups of individuals one treatment may be better than another treatment based on the subject's predicted trajectory. For example, if a subject's age, number of children, and ethnicity are predictive of a trajectory where outcomes are more favorable under medication rather than CBT, then one would consider treating a patient with similar characteristics with medication. On the other hand, it may be that a subject's predicted trajectory suggests that both medication and CBT are effective. In that case, either treatment can be offered. In this way, growth mixture modeling may provide insights on personalized depression treatments that are tailored based on patient characteristics as well as preferences.

More specifically, let  $c_i$  be a latent categorical variable representing the unobserved membership in a trajectory class for participant  $i$ , where  $c_i = 1, 2, \dots, K$ . The variable  $c$  is referred to as a trajectory class variable. Define  $y_{ij}$  as the outcome for participant  $i$  at time  $j$ ,  $j = 0, 1, \dots, n_i$ . Then, conditional on trajectory class  $k$ , the GMM augments Eq. 16 as follows

$$\begin{aligned} (y_{ij} | c_i = k) &= \beta_{0k} + \beta_{ik}t_{ij} + \beta_{2k}t_{ij}Trt_i \\ &+ v_{0ik} + v_{1ik}t_{ij} + \varepsilon_{ijk} \end{aligned} \quad (24)$$

Both the random and fixed effects have the same interpretation as before, but now they are indexed by trajectory class  $k$ , so that they may vary by trajectory class.

Class membership is expressed by a multinomial logistic regression of the form:

$$P(c_i = k | x_i) = \frac{e^{x_i' \delta_k}}{\sum_{s=1}^K e^{x_i' \delta_s}} \quad (25)$$

where the variable  $x$  can represent baseline covariates. When there are only two classes, Eq. 25 is a logistic regression estimating the probability of being in one class versus another.

For binary variables  $x$  in Eq. 25,  $e^\delta$  can be interpreted as the odds ratio of being in one class versus another. For example, if  $x$  is gender, then one can estimate the odds of a male participant being in one trajectory versus a female.

The number of trajectories in a GMM must be specified a priori. Typically, several GMMs are fit assuming a different number of trajectory classes and the “correct” number of trajectories is chosen based on model fit criteria such as BIC. See Muth'en et al. (2002) and Muth'en et al. (2009) for more detail on fitting GMMs in clinical trial settings and Siddique et al. (2012) for an example of a GMM fit to the WECare data.

---

## Discussion

This chapter reviewed methods for the analysis of longitudinal data commonly encountered in health services research. The chapter began by discussing issues inherent in longitudinal data and then described methods for analyzing these data, focusing on linear mixed-effects models and covariance-pattern models for continuous data. These methods were applied to data from a longitudinal depression treatment trial, going into specific detail on model selection, estimation of treatment effects, calculation of effect sizes, and interpretation.

Data from health services research are often missing and/or not continuous. These types of data suggest the use of models in addition to those discussed in this chapter. Due to space limitations, extended models for missing data and nonlinear models for noncontinuous data were only briefly mentioned. As described, MRMs and CPMs do allow for missing data and provide valid results under the assumption of missing at random (MAR). Thus, the extended missing data models are useful to the extent that researchers suspect that the missing data are missing not at random, a situation that is impossible to ascertain with the observed data. Finally, the chapter briefly described generalized estimating equation (GEE) models and growth mixture models (GMMs) for longitudinal data, noting some distinguishing

features of these classes of models relative to MRMs and CPMs.

Mixed-effects models, which allow one to estimate subject-specific change over time and provide valid estimates in the presence of data missing at random should be considered as the preferred methodology for analysis of longitudinal data by health services researchers. Most current statistical software packages include functions for estimating MRMs and their various extensions, thus making them easily accessible to the interested researcher.

**Acknowledgments** The authors wish to thank Jeanne Miranda for use of the WECare data. Dr. Siddique's work was supported by grant K07 CA154862-01 from the National Cancer Institute and R03 HS018815-01 from the Agency for Healthcare Research and Quality. Dr. Hedeker's work was supported by Award Number P01 CA098262 from the National Cancer Institute. Dr. Gibbons' work was supported by R01 MH8012201 from the National Institute of Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute, Agency for Healthcare Research and Quality, or the National Institutes of Health.

---

## References

- Berkhof J, Snijders TAB. Variance component testing in multilevel models. *J Educ Behav Stat.* 2001;26:133–52.
- Bock RD. Multivariate statistical methods in behavioral research. New York: McGraw-Hill; 1975.
- Bock RD. Within-subject experimentation in psychiatric research. In: Gibbons RD, Dysken MW, editors. *Statistical and methodological advances in psychiatric research.* New York: Spectrum; 1983a. p. 59–90.
- Bock RD. The discrete Bayesian. In: Wainer H, Messick S, editors. *Modern advances in psychometric research.* Hillsdale: Erlbaum; 1983b. p. 103–15.
- Bock RD. Measurement of human variation: a two stage model. In: Bock RD, editor. *Multilevel analysis of educational data.* New York: Academic; 1989.
- Bryk AS, Raudenbush SW. *Hierarchical linear models: applications and data analysis methods.* Newbury Park: Sage; 1992.
- Chi EM, Reinsel GC. Models for longitudinal data with random effects and AR(1) errors. *J Am Stat Soc.* 1989;84:452–9.
- Conaway MR. Analysis of repeated categorical measurements with conditional likelihood methods. *J Am Stat Assoc.* 1989;84:53–61.

- Daniels MJ, Hogan JW. Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics*. 2000;56:1241–8.
- Daniels MJ, Hogan JW. Missing data in longitudinal studies: strategies for Bayesian modeling and sensitivity analysis. New York: Chapman & Hall/CRC; 2008.
- de Leeuw J, Kreft I. Random coefficient models for multi-level analysis. *J Educ Stat*. 1986;11:57–85.
- Demirtas H, Schafer JL. On the performance of random-coefficient pattern-mixture models for nonignorable dropout. *Stat Med*. 2003;22:2553–75.
- Dempster AP, Rubin DB, Tsutakawa RK. Estimation in covariance component models. *J Am Stat Soc*. 1981;76:341–53.
- Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. Analysis of longitudinal data. 2nd ed. New York: Oxford University Press; 2002.
- Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. 2nd ed. Hoboken: Wiley; 2012.
- Gibbons RD. Trend in correlated proportions. PhD thesis, University of Chicago, Department of Psychology, 1981.
- Gibbons RD, Bock RD. Trend in correlated proportions. *Psychometrika*. 1987;52:113–24.
- Gibbons RD, Hedeker D, Waternaux CM, Davis JM. Random regression models: a comprehensive approach to the analysis of longitudinal psychiatric data. *Psychopharmacol Bull*. 1988;24:438–43.
- Goldstein H. Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*. 1986; 73:43–56.
- Goldstein H. Nonlinear multilevel models, with an application to discrete response data. *Biometrika*. 1991;78:45–51.
- Goldstein H. Multilevel statistical models. 4th ed. Hoboken: Wiley; 2011.
- Grady JJ, Helms RW. Model selection techniques for the covariance matrix for incomplete longitudinal data. *Stat Med*. 1995;14:1397–416.
- Guo W, Ratcliffe SJ, Ten Have TR. A random pattern-mixture model for longitudinal data with dropouts. *J Am Stat Assoc*. 2004;99:929–37.
- Hardin JW, Hilbe JM. Generalized estimating equations. 2nd ed. New York: Chapman and Hall; 2012.
- Hedeker D. Random regression models with auto-correlated errors. PhD thesis, University of Chicago, Department of Psychology, 1989.
- Hedeker D, Gibbons RD. Longitudinal data analysis. New York: Wiley; 2006.
- Hui SL, Berger JO. Empirical Bayes estimation of rates in longitudinal studies. *J Am Stat Assoc*. 1983;78:753–9.
- Ibrahim J, Molenberghs G. Missing data methods in longitudinal studies: a review (with discussion). *TEST*. 2009;18:1–43.
- Kaplan D, George R. Evaluating latent growth models through ex post simulation. *J Educ Behav Stat*. 1998; 23:216–35.
- Kenward MG, Molenberghs G. Parametric models for incomplete continuous and categorical longitudinal data. *Stat Methods Med Res*. 1999;8(1):51–83.
- Laird NM. Missing data in longitudinal studies. *Stat Med*. 1988;7:305–15.
- Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982;38:963–74.
- Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:13–22.
- Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *J Am Stat Assoc*. 1995;90:1112–21.
- Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. New York: Wiley; 2002.
- Longford NT. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*. 1987;74:817–27.
- Longford NT. Random coefficient models. New York: Oxford University Press; 1993.
- Miranda J, Chung JY, Green BL, Krupnick J, Siddique J, Revicki DA, Belin T. Treating depression in predominantly low-income young minority women. *J Am Med Assoc*. 2003;290:57–65.
- Miranda J, Chung JY, Green BL, Krupnick J, Siddique J, Revicki DA. One year outcomes of treating depression in predominantly low-income young minority women. *J Clin Consult Psychol*. 2006;74:99–111.
- Muth'en BO. Latent variable analysis: growth mixture modeling and related techniques for longitudinal data. In: Kaplan D, editor. Handbook of quantitative methodology for the social sciences. Newbury Park: Sage; 2004.
- Muth'en B, Shedden K. Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*. 1999;55:463–9.
- Muth'en B, Brown CH, Masyn K, Jo B, Khoo ST, Yang CC, Wang CP, Kellam SG, Carlin JB, Liao J. General growth mixture modeling for randomized preventive interventions. *Biostatistics*. 2002;3(4):459–75.
- Muth'en BO, Brown CH, Leuchter A, Hunter A. General approaches to analysis of course: applying growth mixture modeling to randomized trials of depression medication. In: Shrout PE, editor. Causality and psychopathology: finding the determinants of disorders and their cures. Washington, DC: American Psychiatric Publishing; 2009. Forthcoming.
- Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int Stat Rev*. 1991;59:25–35.
- Potthoff RF, Roy SN. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*. 1964;51:313–6.
- Raudenbush SW, Bryk AS. A hierarchical model for studying school effects. *Sociol Educ*. 1986;59:1–17.
- Raudenbush SW, Bryk AS. Hierarchical linear models. 2nd ed. Thousand Oaks: Sage; 2002.
- Rubin DB. Inference and missing data. *Biometrika*. 1976;63:581–92.
- Siddique J, Brown CH, Hedeker D, Duan N, Gibbons RD, Miranda J, Lavori PW. Missing data in longitudinal trials—part B, analytic issues. *Psychiatr Ann*. 2008; 38(12):793–801.

- Siddique J, Chung JY, Brown CH, Miranda J. Comparative effectiveness of medication versus cognitive behavioral therapy in a randomized controlled trial of low-income young minority women with depression. *J Consult Clin Psychol.* 2012;80:995–1006.
- Singer JD, Willett JB. *Applied longitudinal data analysis.* New York: Oxford University Press; 2003.
- Stiratelli R, Laird NM, Ware JH. Random-effects models for serial observations with binary response. *Biometrics.* 1984;40:961–71.
- Strenio JF, Weisberg HI, Bryk AS. Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates. *Biometrics.* 1983;39:71–86.
- Verbeke G, Molenberghs G. *Linear mixed models for longitudinal data.* New York: Springer; 2000.
- Weiss RE. *Modeling longitudinal data.* New York: Springer; 2005.
- Wolfinger RD. Covariance structure selection in general mixed models. *Commun Stat Simul Comput.* 1993;22:1079–106.
- Wong GY, Mason WM. The hierarchical logistic regression model for multilevel analysis. *J Am Stat Assoc.* 1985;80:513–24.
- Xu W, Hedeker D. A random-effects models for classifying treatment response in longitudinal clinical trials. *J Biopharm Stat.* 2002;11:253–73.
- Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics.* 1986;42:121–30.
- Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics.* 1988;44:1049–60.



Melania Pintilie

## Contents

<b>Introduction</b> .....	434
Motivation and Examples .....	434
The Need to Analyze Time to Event of Interest .....	434
The Follicular Lymphoma Example .....	435
The Pressure Ulcer Healing (PUH) Example .....	435
<b>Estimation of the Probability of Event</b> .....	435
Necessity for Special Techniques .....	435
Nonparametric Estimation of Probability of Event in the Presence of Competing Risks .....	436
The Justification of the Kalbfleisch and Prentice Formula (1) .....	437
The Intuitive Justification for Formula (1) .....	437
Confidence Intervals .....	438
<b>Theoretical Background</b> .....	438
General Remarks .....	438
A Theoretical Example .....	439
<b>Regression Model</b> .....	440
Fine and Gray Model .....	440
Interpretation of the Fine and Gray Model .....	441
Cox Regression in the Presence of Competing Risks .....	442
<b>Other Developments</b> .....	443
Analyzing Correlated Data .....	443
Analyzing Case-Cohort Design .....	443
<b>Sample Size and Power</b> .....	443
<b>Software</b> .....	445
<b>References</b> .....	445

---

M. Pintilie (✉)  
University Health Network, Toronto, ON, Canada  
e-mail: [pintilie@uhnresearch.ca](mailto:pintilie@uhnresearch.ca)

### Abstract

In the time-to-event analysis when more than one type of event can occur and not all are of interest, the situation of competing risks appears. In this chapter the competing risks will be defined, and the need for special statistical analysis techniques will be justified. The methodology for estimation and modeling in the presence of competing risks will be presented. The cumulative incidence function and the Fine and Gray model will be introduced as the main methods to analyze competing risks data. The cumulative incidence function will be contrasted to Kaplan-Meier method. For a deeper understanding of the modelling, the subdistribution hazard will be defined.

The importance of considering the competing risks in the process of designing a study will be emphasized, and the steps needed to be taken in the calculation will be presented. For a better understanding of the material and of the interpretation, examples will be given at each step.

... the event whose occurrence either precludes the occurrence of another event under investigation or fundamentally alters the probability of occurrence of this other event.

As an example, consider a cohort of patients with chronic kidney disease. The interest is to study the time to dialysis. However, patients could die due to comorbidities and never reach the point of dialysis. The death before dialysis initiation is a competing risks event.

The time to local recurrence as the event of interest in cancer treatment is another example. In this case, the occurrence of a distant recurrence could be considered a competing risks event because the treatment for such a recurrence could change the probability of developing a local recurrence.

The existence of competing risks was recognized by David and Moeschberger (1978) in their monograph, and later Kalbfleisch and Prentice (1980) introduced a nonparametric estimation of the probability of the event of interest. And yet in medical research, it was completely ignored until recently. Most of the statistical analyses published before 1990 used inappropriate techniques to analyze the time-to-event data when a competing risk was present. Basically the competing risks event was considered censored as if for that observation the event of interest could still be observed in the future.

## Introduction

### Motivation and Examples

In the time-to-event analysis, the outcome is given by two pieces of information: a continuous part representing the duration of time under the follow-up and a binary part indicating whether at that time an event was observed or not. The observation for which the event was not observed is called censored. It is assumed that with enough follow-up, the events will be observed for all observations. The obvious example is the time to death. Death is an event that eventually will be observed for each observation. However, in some situations more than one type of event can happen. The occurrence of one type of event can hinder the observation or change the probability of other types of events being observed. Such a situation is called competing risks. Gooley et al. (1999) gave a formal definition of a competing risks situation as:

### The Need to Analyze Time to Event of Interest

There are many examples in medical research when it is important to study a specific event of interest.

In cancer research one of the standard treatments is radiation therapy (RT). RT treats a small part of the body, where the tumor is. Thus, if one is interested in the effect of the treatment, it is fair to think of the effect in the treated area, local control of disease. Yet, a patient could experience other events: a relapse in a different part of the body, another malignancy, or death of a different cause. Sometimes all the events (event of interest and the competing risks event) are combined in a



composite end point. However, this approach could diminish the effect on the event of interest or even suggest a totally different conclusion.

Sometimes a composite end point is not feasible. During the treatment a patient needs to be assisted by temporarily inserting a feeding tube. After the treatment and as the patient recovers, the tube is taken out. The time at which the tube is taken out could be considered as a surrogate for response. This end point cannot be considered together with death, for example, as the former is a positive outcome and the latter a negative one.

The following two examples will be utilized along this chapter to illustrate the different aspects of the analysis in the presence of competing risks. The datasets were slightly modified to help illustrate competing risks analysis. Clinical conclusion cannot be drawn from these analyses.

### The Follicular Lymphoma Example

Consider as an example a cohort of patients with early-stage follicular lymphoma with the follow-up ranging between 1 and 31 years. For this disease, the prognosis is good with 10 year survival of approximately 75%. These patients could experience relapses (local and/or distant), a second malignancy, or die of other causes. Each of these events can be of interest with the rest being competing risks with the exception of death which cannot have any competing risks.

### The Pressure Ulcer Healing (PUH) Example

This is a cohort of patients with advanced illness who were admitted to a palliative care center and followed until death (Maida et al. 2008, 2012). All patients had at least one pressure ulcer at the time of admittance, and the time from admittance to complete healing was recorded for all pressure ulcers that healed. The life expectancy for the cohort is low with median survival less than a month. The goal of this analysis is to study the

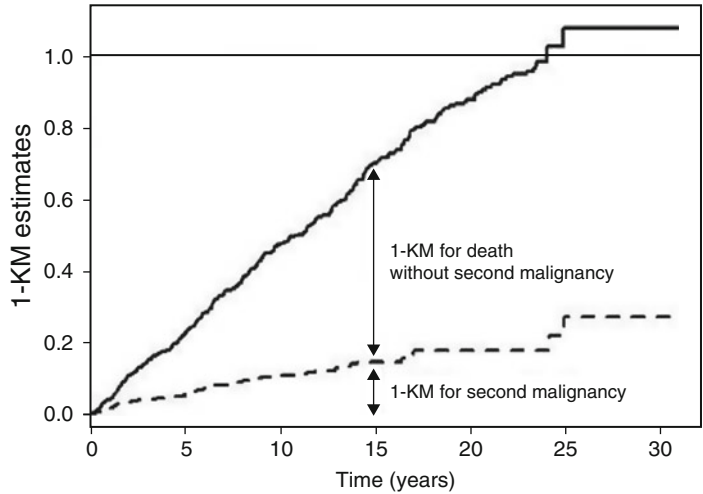
time to pressure ulcer healing as a function of a patient's Palliative Performance Scale status, an important clinical factor which for this analysis is dichotomized at 40, bedridden vs. ambulatory. If a patient had more than one pressure ulcer, one was chosen at random for analysis to avoid having to deal with the added complexity of correlated observations (see section "[Analysing Correlated Data](#)"). Dr. Vincent Maida and Dr. Marguerite Ennis graciously allowed the use of the pressure ulcer healing data (Maida et al. 2012) for the illustration of the concepts of this chapter.

---

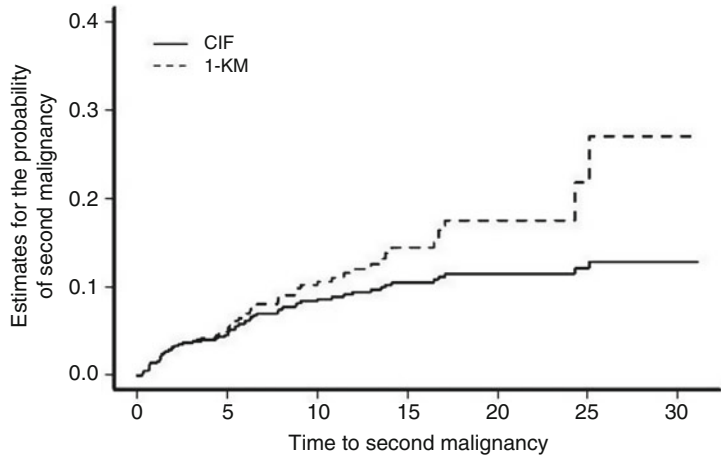
## Estimation of the Probability of Event Necessity for Special Techniques

In the presence of competing risks, the estimates based on the Kaplan-Meier (KM) method when the competing risk is censored are not probabilities. This concept is illustrated using the cohort of follicular lymphoma described in section "[The Follicular Lymphoma Example](#)." The event of interest is the time to second malignancy following the lymphoma diagnosis. For the moment, the competing risks (the deaths without second malignancy) are ignored and censored. With this assumption, KM estimates can be obtained. KM estimates can also be calculated for the deaths without second malignancy as event and with the second malignancy censored. If the KM estimates can be interpreted as probabilities, then the calculated 1-KM would be the probability for each of the two specific types of event to happen. Since the two types of events are mutually exclusive, the sum of the 1-KM estimates calculated at each time point should be the probability of any of the two events to occur, namely, the probability for either second malignancy or death without second malignancy. In Fig. 1, the broken line is the 1-KM estimate for the second malignancy, while the solid line represents the sum of the 1-KM estimates for second malignancy and the death without second malignancy. The fact that the top line goes beyond the possible upper limit of a probability is a proof that

**Fig. 1** (1-KM) Estimates for second malignancy and death without second malignancy in the follicular lymphoma dataset



**Fig. 2** Plot of the CIF and 1-KM for second malignancy in follicular lymphoma



when competing risks are present, the KM estimates cannot be interpreted as probabilities.

and  $d_{ev j}$  are the number of events of interest at time  $t_j$ . The probability of event can be estimated as:

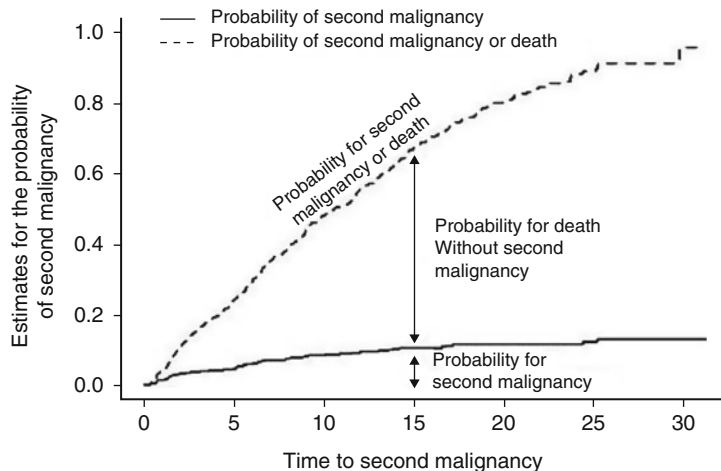
**Nonparametric Estimation of Probability of Event in the Presence of Competing Risks**

$$\hat{F}_{ev}(t) = \sum_{all j, t_j \leq t} \frac{d_{ev j}}{n_j} \hat{S}(t_{j-1}) \quad (1)$$

Kalbfleisch and Prentice (1980) modified the KM estimator to obtain the probability of event in the presence of competing risks. Briefly, suppose  $t_1 < t_2 < \dots$  are the ordered time points for all types of events,  $n_j$  are the number at risk at time  $t_j$ ,

Here  $\hat{S}(t_{j-1})$  is the KM estimate for the complement of the probability of all types of events. In the literature,  $\hat{F}_{ev}(t)$  is sometimes called cumulative incidence function (CIF). Figure 2 shows the estimation based on (1) and on the KM method for the second malignancy in follicular

**Fig. 3** Probability of second malignancy and death without second malignancy in the follicular lymphoma



lymphoma. Clearly the two are different, and the 1-KM estimates are larger than the CIF. It can be proven algebraically that in the presence of competing risks, 1-KM is always larger than the CIF.

$$\begin{aligned}
 \hat{F}(t) &= \sum_{t_j \leq t} \frac{d_j}{n_j} \hat{S}(t_{j-1}) = \sum_{t_j \leq t} \frac{d_{1j} + d_{2j}}{n_j} \hat{S}(t_{j-1}) \\
 &= \sum_{t_j \leq t} \frac{d_{1j}}{n_j} \hat{S}(t_{j-1}) + \sum_{t_j \leq t} \frac{d_{2j}}{n_j} \hat{S}(t_{j-1}) \\
 &= \hat{F}_1(t) + \hat{F}_2(t)
 \end{aligned} \tag{3}$$

**The Justification of the Kalbfleisch and Prentice Formula (1)**

The well-known formula for the KM estimates can be written as a sum for its complement, the estimator for the probability of all events:

$$\begin{aligned}
 \hat{F}(t) &= 1 - \hat{S}(t) \\
 &= 1 - \prod_{t_j \leq t} \frac{n_j - d_j}{n_j} \\
 &= \sum_{t_j \leq t} \frac{d_j}{n_j} \hat{S}(t_{j-1})
 \end{aligned} \tag{2}$$

where  $t_1 < t_2 < \dots$  are the ordered time points for the events,  $n_j$  are the number at risk at time  $t_j$ , and  $d_j$  are the number of events at time  $t_j$ . Suppose that there are two types of events which can occur at time  $t_j$ ; then the total number of events that can happen can be written as the sum of the number of events of type 1,  $d_{1j}$ , and number of events of type 2,  $d_{2j}$ . Then the probability of all events (2) can be written as a sum of the probabilities of the two types of events:

It is easy to recognize the formula for the estimation of the probability of the event of interest (1) in the two terms in (3). Thus the probability of all events can be partitioned in the probabilities of the constituent types of events. Figure 3 shows the partition of the probability of second and death with second malignancy or death into probability of second malignancy and probability of death without second malignancy in the follicular lymphoma dataset.

**The Intuitive Justification for Formula (1)**

In the absence of censoring or competing risks, the estimation of the probability of event for a time point  $t_0$  using the KM method gives an identical result to the intuitive calculation of the ratio between the number of events occurred before  $t_0$  and the total number of subjects. In this sense the KM method can be considered an extension for calculating the probability of event in the presence

**Table 1** Table of percentages

Time point	CIF (%)	1-KM (%)	Naive estimates (%)
1	1.5	1.5	1.5
2	3.1	3.2	3.1
3	3.7	3.8	3.7
4	4.1	4.2	4.1
5	4.6	4.9	4.6
6	6.1	6.7	6.1
7	6.8	7.6	6.8
8	7.2	8.1	7.2
9	7.8	8.9	7.8
10	7.9	9.2	7.9

of censoring. Along the same lines, the CIF estimation (given by (1)) is an extension of the KM method for calculating the probability of event in the presence of competing risks. Thus, if competing risks are not present, the CIF is identical to 1-KM. If competing risks exist but there is no censoring, the CIF is identical to the ratio of the number of events of interest to the number of subjects. To illustrate, the follow-up of the follicular lymphoma dataset was completed to 10 years in an artificial way.

Note that (Table 1) the CIF is identical to the naïve estimates (the ratio between the number of events up to the time point of interest and the total number of subjects) while the 1-KM estimates are larger. It must be emphasized that the equality between the CIF and the naïve estimates holds only when there are no censored observations with shorter follow-up time than the time point at which the calculation is made.

**Confidence Intervals**

As for any estimation, it is desirable to be able to assess the degree of confidence. The customary way is to present the 95% confidence interval of the estimate. This involves the knowledge of the distribution of the estimate and its variance. The  $\ln(1-CIF)$  can be considered to be normally distributed. The variance can be calculated in several ways, but the differences are minimal (Pintilie 2006). In general

the software calculates the variance but may not give the confidence interval. The confidence interval can be calculated using the same technique as in a noncompeting risks situation (Kalbfleisch and Prentice 1980). If  $cCIF$  is the complement of CIF (i.e.,  $1-CIF$ ), then the limits of the confidence interval for  $cCIF$  are given by:

$$cCIF^{\exp(\pm A)}, \quad \text{where } A = \frac{z_{1-\alpha/2} \sqrt{\widehat{Var}(cCIF)}}{cCIF \times \ln(cCIF)} \tag{4}$$

and  $z_{1-\alpha/2}$  is the quantile of the standard normal distribution for 95% confidence interval,  $z_{1-\alpha/2} = 1.96$ .

**Theoretical Background**

This presentation of the competing risks issue is not intended to be mathematical in nature. However, for a thorough understanding of the subject, it was decided to include some theoretical details. The reader who is not mathematically inclined could skip this section.

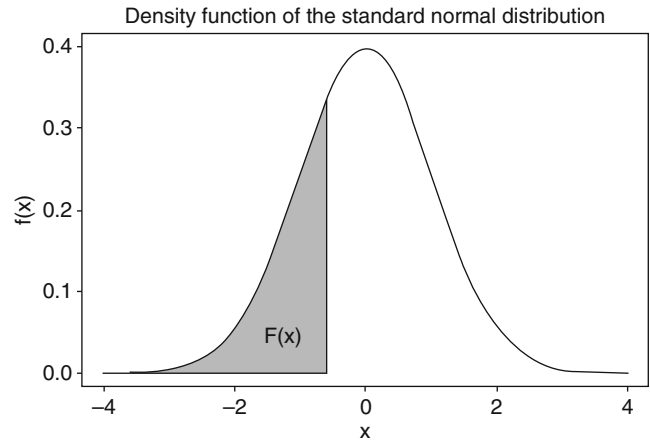
**General Remarks**

In statistics there are two interrelated functions: the density, usually denoted by  $f$ , and the distribution function, usually denoted by  $F$ . The known bell shape of the normal distribution is the plot of the density function. The integral to a certain point  $x$  measures the area under that curve,  $F(x)$ , and it represents the probability that a number generated from the normal distribution is smaller than  $x$ . (See Fig. 4.)

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du \tag{5}$$

In the time-to-event analysis, the distribution function appears usually as its complementary function ( $1 - F(x)$ ) and is called the survivor

**Fig. 4** The density function of the standard normal distribution



function, usually denoted by  $S(t)$ . An important property of any distribution function,  $F$ , is that it is an increasing function ranging between 0 and 1. Another important function in the time-to-event analysis is the hazard function,  $h(t)$ , which is the instantaneous risk for event. It can be calculated as the ratio between the density and the survivor function. For the exponential distribution, the density, distribution, and the hazards functions are:

$$\begin{aligned}
 f(x) &= \lambda e^{-\lambda t} \\
 F(t) &= 1 - e^{-\lambda t} \\
 h(t) &= \frac{f(t)}{1 - F(t)} = \lambda
 \end{aligned}
 \tag{6}$$

The importance of the hazard stems from the way modeling is performed. If the choice is for parametric modeling, the decision on the distribution is based on the shape of the hazard. If the modeling is performed utilizing the ubiquitous Cox proportional hazards model, then the hazard itself is modeled.

**A Theoretical Example**

It was shown in (3) that the estimator of the probability of all events can be partitioned in the probabilities of the constituent types of events. This can be formulated more generally as:

$$\begin{aligned}
 F(t) &= P(T \leq t) \\
 &= P(T \leq t | ev = 1) + P(T \leq t | ev = 2) \\
 &= F_1(t) + F_2(t)
 \end{aligned}
 \tag{7}$$

where it is assumed that there are only two types of events and  $ev = 1$  and  $ev = 2$  refer to events of types 1 and 2, respectively. As mentioned above,  $F(t)$  is an increasing function ranging between 0 and 1. Since all terms are probabilities and thus positive, and the probability of all events is at most 1, it follows that each of the probabilities for a specific event can reach only a value  $p < 1$ . Thus the probability of one event in the presence of another event ranges between 0 and a value  $p < 1$ . It follows that  $F_1$  and  $F_2$  cannot be regarded as true, proper distributions. They are called subdistributions.

For each of these subdistributions, there is a subdensity ( $f_1$  and  $f_2$ ). The hazard for event of type 1 can be defined in two ways:

$$\tilde{h}_1(t) = \frac{f_1(t)}{1 - F(t)}
 \tag{8}$$

$$\tilde{\gamma}_1(t) = \frac{f_1(t)}{1 - F_1(t)}
 \tag{9}$$

Each of these hazards can be modeled, and the results could be different as is their interpretation. The hazard from (8) is called the subhazard while the hazard from (9) is called the subdistribution hazard.

As a theoretical example, consider the subdistribution for an event of interest which is exponentially distributed. Under the latent failure time model in which the event of interest and the competing risks event are independent and exponentially distributed with the parameter  $\lambda_1$  and  $\lambda_2$ , respectively, the subdistribution for the event of interest is:

$$F_1(t) = \frac{\lambda_1}{\lambda_1 + \lambda_2} \left(1 - e^{-(\lambda_1 + \lambda_2)t}\right) \quad (10)$$

Note that the quantity which is in brackets is the distribution function of an exponential distribution with parameter  $\lambda_1 + \lambda_2$ . As a distribution function, this quantity spans the 0–1 interval. On the other hand, the factor with which this is multiplied is a positive quantity less than 1. Therefore, the maximum of this function is  $\frac{\lambda_1}{\lambda_1 + \lambda_2}$ , a quantity less than 1. The two hazards are:

$$\tilde{h}_1(t) = \lambda_1 \quad (11)$$

$$\tilde{\gamma}_1(t) = \frac{f_1(t)}{1 - F(t)} = \frac{\lambda_1(\lambda_1 + \lambda_2)e^{-(\lambda_1 + \lambda_2)t}}{\lambda_2 + \lambda_2} \quad (12)$$

Note that the subhazard is the same as the hazard of the marginal distribution. This is always true under the latent failure time assumption and if the two types of event are independent. This lends easily to a nice interpretation of the effect in the absence of the other event. However, the assumption of independence cannot be proven and rarely can be made (Tsiatis 1975). In the absence of independence, the analysis of the subhazard cannot be interpreted.

When the two events are not independent, the subhazard is no longer the hazard of the marginal model:

$$\tilde{h}_1(t) = \lambda_1 + \mu t \quad (13)$$

where  $\mu$  is the parameter which controls the level of dependence between the two types of events.

In contrast, the analysis of the subdistribution hazard does not assume independence, and it can be interpreted as reflecting the observable effect.

More on the interpretation is presented in section “[Interpretation of the Fine and Gray Model.](#)”

---

## Regression Model

### Fine and Gray Model

The Fine and Gray model (1999) is an extension of the Cox model to the situation of competing risks. The effects are estimated by maximizing the pseudo-likelihood, which is a function that depends on the observed covariates and the order in which the events were observed. As in the Cox regression, the hazard is modeled as:

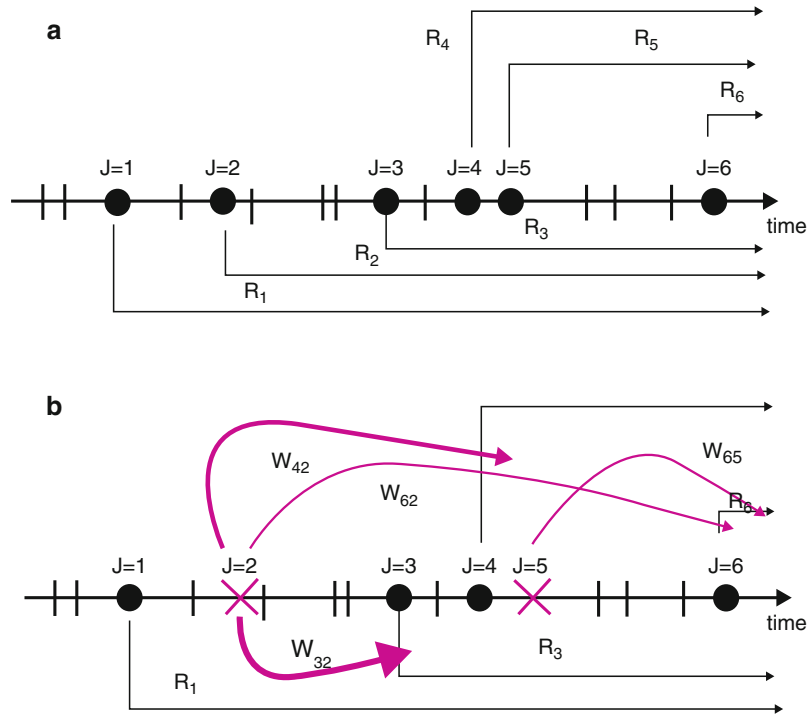
$$\gamma(t|x) = \gamma_0(t)e^{\beta x} \quad (14)$$

where  $x$  is the covariate,  $\gamma_0$  is the baseline hazard, and  $\beta$  is the coefficient estimated by maximizing the pseudo-likelihood given by:

$$PL(\beta) = \prod_{j=1}^r \left( \frac{e^{\beta x_j}}{\sum_{i \in R_j} w_{ij} e^{\beta x_i}} \right) \quad (15)$$

where  $r$  is the number of events of interest and  $R_j$  is the risk set at time  $t_j$ . This formula is written only for one covariate but it can easily be extended to many covariates. The difference between (15) and the partial likelihood of Cox regression is the weight  $w_{ij}$  and the risk set. In Cox regression the risk set is defined as the set of observations with longer observed time than the current event. In addition, for the Fine and Gray model, the risk set also includes all the competing risks events at all time points regardless of the time at which the competing risk was observed. The involvement of the competing risks event is mitigated by the weight: the longer the duration between the current event and the observed competing risks event, the smaller the weight. For example, a competing risks event which happens at 2 years participates fully in the pseudo-likelihood for the terms before 2 years and participates less and less in the pseudo-likelihood for the terms which are farther and farther from

**Fig. 5** The risk set for Cox regression (a) and Fine and Gray regression (b)



2 years. The weights are based on the distribution of the censored time.

In the two diagrams in Fig. 5, the horizontal line represents the time axis, the black circles represent the individual for which the event of interest is observed, the vertical lines are for the censored observations, and the purple crosses in diagram B represent the competing risks events. In diagram A there are no competing risks and all the individuals with the observed time larger than the individual for which the partial likelihood is written are in the risk set. For diagram B the competing risks are always in the risk set, every time with a different weight. Thus, the weight for the individual marked with  $j = 2$  is one for the term  $j = 1$ ,  $w_{32}$  for  $j = 3$ ,  $w_{42}$  for  $j = 4$ , and  $w_{62}$  for  $j = 6$  where  $1 \geq w_{32} \geq w_{42} \geq w_{62} \dots$

**Interpretation of the Fine and Gray Model**

The Fine and Gray regression (1999) models the subdistribution hazard (9). The exponent of a

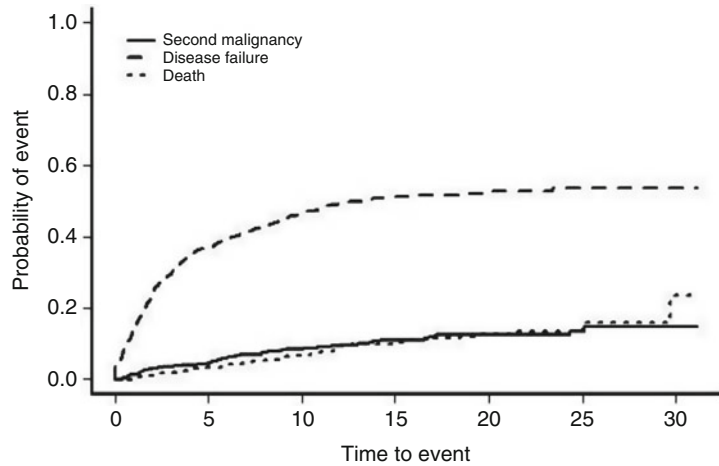
**Table 2** Types of events

Type of event	Frequency	
Second malignancy	56	Event of interest
Relapse before second malignancy	260	Competing risks event
Death without relapse or second malignancy	54	Competing risks event
Censored	171	

coefficient can be interpreted as the subdistribution hazards ratio. As in the Cox regression, the assumption of proportionality of hazards is made and can be checked by visually inspecting the Schoenfeld-type residuals.

Consider the example in section “The Follicular Lymphoma Example” with three types of events: second malignancy, disease failure (relapse), and death without second malignancy or disease failure. Any of these events could be considered as event of interest with the rest of them as competing risks. The types of events and their frequency are listed in Table 2.

**Fig. 6** The probabilities for the three types of event



From Table 2 it is apparent that the most frequent type of event is disease failure. Figure 6 shows that the disease failures occur shortly after the initial diagnosis of follicular lymphoma, while the second malignancies and the death without disease failure happen at a more steady rate.

The Fine and Gray model was applied to second malignancy and to disease failure. Tables 3 and 4 show the results of these models.

Thus, age is the only significant covariate for both types of events. As expected, the disease-specific factors like stage and residual bulk are significant for the disease failure. Furthermore, chemotherapy is marginally not significant. Those with residual bulk or of stage 2 are about 1.5-fold more likely to have disease failures than the ones without residual bulk or stage 1, respectively. Those receiving chemo are less likely to have a disease failure.

Table 5 shows the results when all end points are combined. The results in Table 5 are close to those seen in Table 4 although somewhat weaker for stage, bulk, and chemotherapy. The reason for the resemblance between the last two tables is the fact that there are many more relapses than second malignancies: 260 vs. 56. Thus the results in Table 5 are driven by the number of relapses. Some of the effects are weaker because those covariates have an opposite effect for the second malignancy than for disease failure.

**Table 3** The results of the model with second malignancy as event of interest

	HR	95% conf. int.	p-value
Age	1.03	1.01–1.05	0.0074
Sex: men vs. women	0.98	0.58–1.67	0.94
Stage: 2 vs. 1	0.78	0.41–1.48	0.44
Residual bulk	0.79	0.45–1.41	0.43
Chemotherapy	1.54	0.78–3.02	0.22

**Table 4** The results of the model for disease failure

	HR	95% conf. int.	p-value
Age	1.02	1.01–1.02	0.0019
Sex: men vs. women	1.04	0.81–1.33	0.76
Stage: 2 vs. 1	1.57	1.19–2.08	0.0016
Residual bulk	1.49	1.14–1.95	0.004
Chemotherapy	0.72	0.51–1.01	0.055

**Table 5** The results of the model with all end points combined

	HR	95% conf. int.	p-value
Age	1.04	1.03–1.04	<0.0001
Sex: men vs. women	1.16	0.94–1.42	0.16
Stage: 2 vs. 1	1.41	1.11–1.79	0.0044
Residual bulk	1.45	1.15–1.82	0.0015
Chemotherapy	0.83	0.63–1.11	0.22

### Cox Regression in the Presence of Competing Risks

If the competing risks event is censored, then, from the technical point of view, the analysis could be carried out using the usual Cox model



or Kaplan-Meier estimates, but the interpretation, when possible, is different. In the previous sections, the bias involved in estimating the probability of an event when competing risks are ignored was described. The main question is whether there is a bias when the competing risks are ignored in the modeling process and indeed, if it is possible to predict how large and in which direction this bias is. Another issue is if the results of a model when the competing risks are ignored can be interpreted at all.

In many instances the results of the Cox PH model and Fine and Gray model will be very similar giving the wrong impression that this is a general pattern. However, the two models do not always give similar results. Moreover, the direction of bias cannot be predicted. Finally, the results from the Cox model can be interpreted only under the strict assumption that the distribution of the event of interest and the distribution of competing risks event are independent. This assumption can rarely be made and never substantiated (Tsiatis 1975). The Wound PUH data (given in section “The Pressure Ulcer Healing (PUH) Example”) offers an example when the two models give different results.

Based on the Fine and Gray model (Table 6), the analysis suggests that the performance status is an important prognostic factor with regard to pressure ulcer healing. The patients who are bed-ridden have a longer time to healing than the ambulatory patients. The competing risk of death is ignored in the Cox model, and the effect is much attenuated, the  $p$ -value becomes nonsignificant, and one may reach the wrong conclusion. The probabilities of death and pressure ulcer healing are not independent: knowing that death occurred changes the probability that the pressure ulcer would have healed if the patient could be observed indefinitely. One possible mechanism for this is because the physiological systems needed for

wound healing are part of the system failures associated with death. Only in the rare situation when the event of interest can be assumed independent from the competing events can the Cox model results be interpreted as the effect of a covariate when the competing risks do not exist.

---

## Other Developments

### Analyzing Correlated Data

A notable development is the extension of the Fine and Gray model to accommodate correlated data (strata and/or cluster). For example, in the PUH example, one may wish to analyze all pressure ulcers of a patient rather than just one. This creates clustered data. Zhou et al. (2011, 2012) extended the Fine and Gray model by applying Lee et al.’s (1992) approach.

### Analyzing Case-Cohort Design

When the event of interest is rare, the collection of data for the whole cohort is not feasible. The case-cohort design allows one to take advantage of the number of events of interest while including only a fraction of the data without the event of interest. Pintilie et al. (2010) developed a pseudo-likelihood to analyze a case-cohort design in the presence of competing risks based on Barlow’s work (1999).

---

## Sample Size and Power

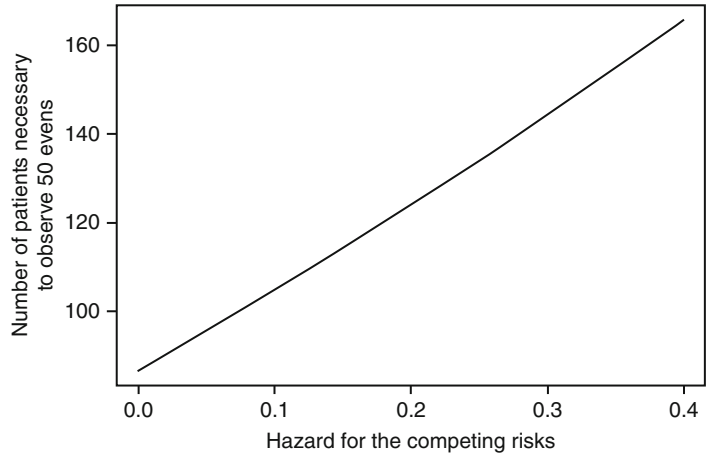
For the time-to-event analysis, the calculation of the sample size necessary to achieve a certain power involves two steps: (a) the calculation of the necessary number of events and (b) the calculation of the necessary number of patients to observe that number of events. The number of events  $n_{ev}$  necessary to detect a specific hazard ratio (HR) is given by:

$$\sqrt{n_{ev}} = \frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{sd(x) \times \ln(HR)} \quad (16)$$

**Table 6** The prognostic value of palliative performance status for wound healing

	HR	95% conf. int.	$p$ -value
Fine and Gray	3.3	1.7–6.7	0.00078
Cox model	1.7	0.8–3.6	0.13

**Fig. 7** The increase in the number of patients as the hazard for CR increases



where  $z_{1-\frac{\alpha}{2}}$  and  $z_{1-\beta}$  are the quantiles of the standard normal distribution for  $\frac{\alpha}{2}$  and  $\beta$ . Thus, for  $\alpha = 0.05$ ,  $z_{1-\frac{\alpha}{2}} = 1.96$  and for  $\beta = 0.2$ ,  $z_{1-\beta} = 0.84$ .  $sd(x)$  stands for the standard deviation of the covariate to be tested. If a randomized trial with equal allocation in two arms is planned, then  $sd(x) = \frac{1}{2}$ . The total number of patients to produce  $n_{ev}$  is:  $n = \frac{n_{ev}}{P_{ev}}$  where  $P_{ev}$  is the probability of the event of interest to occur during the study period. When there are no competing risks,  $P_{ev}$  can be expressed formulaically as:

$$P_{ev} = 1 - \frac{e^{-\lambda f} - e^{-\lambda(\alpha+f)}}{\lambda a} \tag{17}$$

where  $\lambda$  is the hazard rate of the whole cohort,  $a$  is the accrual time, and  $f$  the follow-up time added to the accrual time.

When competing risk are present the formula changes to:

$$P_{ev} = \frac{\lambda_{ev}}{\lambda_{ev} + \lambda_{cr}} \times \left( 1 - \frac{e^{-(\lambda_{ev} + \lambda_{cr})f} - e^{-(\lambda_{ev} + \lambda_{cr})(\alpha+f)}}{(\lambda_{ev} + \lambda_{cr})a} \right) \tag{18}$$

where  $\lambda_{ev}$  and  $\lambda_{cr}$  are the marginal hazards for the events of interest and competing risks event, respectively. It is obvious that if  $\lambda_{cr} = 0$ , i.e., when competing risks do not exist, the formula (18) becomes (17). A close look of formula (18) shows that as  $\lambda_{cr}$  increases, the  $P_{ev}$  decreases

dramatically. This is equivalent to say that as the  $\lambda_{cr}$  increases the total number of patients necessary to observe, a certain number of events of interest increase greatly.

Intuitively, this is obvious since the competing risks hinder the observation of the event of interest. One example is shown in Fig. 7 where an increase of the competing risks from 0 to 0.4 causes a doubling of the final sample size.

The higher the rate of competing risks, the less likely is to observe the event of interest, and therefore a larger initial sample sizes is needed. Therefore, ignoring the competing risks in the design stage will create an underpowered study and will result in a waste of effort and money.

Although the independence between the event of interest and the competing risks event cannot be usually assumed in the analysis phase, this assumption is needed to be made in this section for mathematical tractability. The second assumption made was that the time to the two types of events follows exponential distribution.

**Example 1** Suppose that the researcher wants to validate the prognostic value of a specific marker in a cohort of patients. The marker is measured as present or absent, and the frequency of a positive marker is about half in this population. The cohort is already assembled, and it is known that there are 50 events of interest. The researcher wants to know if there is enough power to detect an effect size corresponding to a subdistribution hazard

ratio of 2 at the level of significance of 0.05. Solving the formula (16) for  $z_{1-\beta}$ , the power is found to be 69%.

**Example 2** A randomized study is being planned to test a new way of delivering radiation for cancer patients. Since radiation is a local treatment, the investigators are interested to test its effect on local disease. Patients may experience a relapse outside the treated area or death of other causes, both representing competing risks events. It is known from previous studies that the rate of local disease in the standard arm is  $\lambda_{ev} = 0.4$  and the rate of other relapses and death of other causes  $\lambda_{cr} = 0.1$ . It is expected that the new treatment will not change the rate of competing risks but it will decrease the rate of local disease to 0.2. The cancer center can accrue 50 patients per year, and it is desirable that the study will accrue the patients in 5 years or less. The analysis will take place 1 year after finishing accrual. The  $\alpha$  level is set to 0.05, and the desired power is 80%. Thus,  $z_{1-\alpha/2} = 1.96$  and  $z_{1-\beta} = 0.84$ .

Note that the given rates for the local relapse refer to the marginal distributions; basically these rates are the hazards of the marginal exponential distributions. The ratio of the two rates for the local relapse (0.4 and 0.2) is not the subdistribution hazard ratio which will be detected. Unfortunately, even in the simple situation when all distributions are exponential and independent, the subdistribution hazards ratio is not independent of time. Its formula can be written as:

$$sHR = \frac{\lambda_1(\lambda_1 + \lambda_{cr})(\lambda_{cr} + \lambda_2 e^{-(\lambda_{ev} + \lambda_{cr})t})e^{-(\lambda_{ev} + \lambda_{cr})t}}{\lambda_2(\lambda_2 + \lambda_{cr})(\lambda_{cr} + \lambda_1 e^{-(\lambda_1 + \lambda_{cr})t})} \quad (19)$$

where  $\lambda_1$  and  $\lambda_2$  are the hazard rates for the local relapse for the standard and the new treatment, respectively, and  $\lambda_{cr}$  is the hazard rate for the competing event for both arms.

For the time span (0–6 years) of this study, sHR ranges between 2 and 1.1. The approximate average is about 1.66 and with this hazard ratio

formula (16) puts the approximate number of events at 122. The formula (18) can be applied for each of the two arms, and probability of event for the standard arm is 0.62 and for the new treatment is 0.41. On average it can be said the probability of event in the study is approximately 0.5. Since the necessary number of events is 122, the total number of patients needs to be 244. This center can accrue 50 patients per year, and thus 244 is a reachable goal. Note that relaxing the accrual effort is not allowed as the maximum number the center can accrue is very close to the total number of patients needed.

---

## Software

The competing risk analysis can be performed almost entirely within R environment using the package *cmprsk* developed by Gray. This package contains functions which give the possibility to estimate the probability of event of interest at any time point, to plot these estimates, to apply the Fine and Gray model, and to plot the predictive probabilities of the event of interest based on this model. The package *crrSC* developed by Zhou extends the Fine and Gray model for stratified or cluster data.

The package *mstate* can be used to modify the data such that the usual Cox model can be applied. This analysis still models the subdistribution hazard, and the obtained coefficients are very close to the results obtained using the function *crr* from *cmprsk*. However, the variance-covariance matrix is slightly different, but for large datasets the differences are minimal.

STATA has a function which allows the user to apply the Fine and Gray model. The plots obtained are the predictive plots from the model.

---

## References

- Barlow EW, Ichikawa L, Rosner D, Izumi S. Analysis of case-cohort design. *J Clin Epidemiol.* 1999;52(12):1165–72.
- David HA, Moeschberger ML. The theory of competing risks. London: Griffin; 1978.

- Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc.* 1999;94:496–509.
- Gooley TA, Leisenring W, Crowley J, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Stat Med.* 1999;18:695–706.
- Kalbfleisch JD, Prentice RL. *The statistical analysis of failure time data.* New York: John Wiley & Sons, Inc.; 1980.
- Lee EW, Wei LJ, Amato D, Leurgans S. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In: Klein JP, Goel PK, editors. *Survival analysis: state of the art.* Dordrecht: Kluwer; 1992.
- Maida V, Corbo M, Dolzhykov M, Ennis M, Irani S, Trozzolo L. Wounds in advanced illness: a prevalence and incidence study based on a prospective case series. *Int Wound J.* 2008;5(2):305–14.
- Maida V, Ennis M, Corban J. Wound outcomes in patients with advanced illness. *Int Wound J.* 2012;9(6):683–92.
- Pintilie M. *Competing risks a practical perspective.* Chichester: Wiley & Sons Ltd; 2006.
- Pintilie M, Bai Y, Yun LS, Hodgson DC. The analysis of case cohort design in the presence of competing risks with application to estimate the risk of delayed cardiac toxicity among Hodgkin Lymphoma survivors. *Stat Med.* 2010;29(27):2802–10.
- Tsiatis A. Nonidentifiability aspect of problem of competing risks. *Proc Natl Acad Sci U S A.* 1975;72:20–2.
- Zhou BQ, Latouche A, Rocha V, Fine J. Competing risks regression for stratified data. *Biostatistics.* 2011; 67(2):661–70.
- Zhou BQ, Fine J, Latouche A, Labopin M. Competing risks regression for clustered data. *Biostatistics.* 2012;13(3):371–83.



Shizhe Chen and XH Andrew Zhou

## Contents

<b>Introduction</b> .....	448
<b>Methods for Mean Inference</b> .....	449
Parametric Methods on Continuous Data .....	449
Nonparametric Methods on Continuous Data .....	454
Zero-Inflated Data .....	455
Two Sample .....	458
Applications on a Simple Example .....	462
<b>Regression</b> .....	462
Parameters of Interest .....	464
Linear Regression on Raw Data .....	464
Transformation on Y .....	466
Transformation on $E[Y]$ .....	469
Two-Part Models .....	471
Quantile Regression .....	472
<b>Prediction</b> .....	473
Some Basic Concepts of Prediction Models .....	473
Difference from Regression Analysis .....	474
<b>Appendix</b> .....	475
Concept of General Pivots .....	475
Variances and Estimators for Back-Transformations .....	476
<b>References</b> .....	477

## Abstract

Cost has become an important outcome in health services research. It can be used not only as a measure for health care spending but also as a measure for a part of health care value. Given ever-increasing rising health care expenditure, the value of health care should include not only traditional measures, such as mortality and morbidity, but also the cost of health care. Due to a limited resource, a new treatment with a slightly

---

S. Chen  
Department of Biostatistics, University of Washington,  
Seattle, WA, USA

X. A. Zhou (✉)  
Beijing International Center for Mathematical Research,  
Peking University, Beijing, China

VA Puget Sound Healthcare System, University of  
Washington, Seattle, WA, USA  
e-mail: [azhou@u.washington.edu](mailto:azhou@u.washington.edu)

better efficacy but much higher cost than an existing treatment may not be a choice of a treatment for a patient. Hence, it is important to be able to approximately analyze cost data. However, appropriately analyzing health care costs may be hindered by special distribution features of cost data, including skewness, zero values, clusters, heteroscedasticity, and multimodality.

Over the decades, various methods have been proposed to address these features. This chapter would be devoted in introducing methods that are able to provide relatively trustworthy results with acceptable efficiency, covering topics on mean inference, regression, and prediction.

---

## Introduction

Rapidly rising of health care costs and health care reforms to containing the health care costs makes the cost be an important outcome in any health services research. It is not straightforward to analyze cost data due to some of its special distributional features, which prevent us from using traditional statistical methods.

The first feature of cost data is its positive skewness or skewed to the right. The skewness arises due to a few patients with high costs, who are accounted for the major part of the total expenses. In addition, cost data often comes with a heavy upper tail, which occurs when the tail of the distribution cannot be bounded by an exponential distribution.

The second feature is discontinuity of the distribution at the zero value, which occurs because not all subjects in the population of interest occur health care costs in a given study period. For example, patients without any hospitalization during a study period have zero in-patient costs. One consequence of the distributional discontinuity at zero is that many standard statistical methods, which require a continuous distribution assumption, cannot be used in the inference of cost data.

The third feature is heteroscedasticity, which occurs when the variance of the cost of a patient is not constant. For example, if the variance of a random variable is a function of the mean, data generated from the distribution of this random

variable will exhibit heteroscedasticity. This kind of the mean-variance relation can also be observed in many known parametric distributions, such as a Poisson distribution and a lognormal distribution. Many traditional statistical methods, such as ordinary least square (OLS), require homoscedasticity in their validity in making statistical inference. Ignoring heteroscedasticity in cost data can lead to wrong statistical inferences.

The fourth feature is censoring of the cost outcome, which occurs when the cost of a patient over a study period is observed. For example, a patient drops out of the study before the study ends; as a result, we only observe the partial cost of this patient over the whole study period. Although the problem of censored cost data is related to survival analysis, analytic techniques are different from traditional survival analysis ones.

The fifth feature is clustering, which occurs due to the effects of clinicians and hospitals. Since some clinician tends to give patients similar prescriptions and uses similar kinds of drugs and treatments, the medical cost of this clinician is expected to be correlated. The same reason goes with clinics and hospitals. Ignoring clustering would lead to invalid statistical inference.

The final feature, not the last one, is multimodality, which occurs when the distribution has more than one mode. This feature may be related to clinician clustering. For example, if the distribution of cost data is generated from patients who are cared by two physicians with different treatment strategies: one physician uses a more liberal approach of ordering tests and describing drugs, and another is more conservative in treating his/her patients, the distribution of the cost data is a mixture of the distributions of two physicians, which may lead to a bimodal distribution.

In this chapter, we are concentrating on a review of statistical methods that can handle the first three distributional features of cost data: (1) skewness, (2) zero values, and (3) heteroscedasticity. We review various methods that have been proposed to address these features. As there is no single method that can handle all features that one might encounter with in a health cost study, in this chapter, we also provide a rough evaluation of those methods to help researchers in choosing methods

that are most suitable. This chapter is organized as follows. Section “[Introduction](#)” focuses on mean inference, which is the very foundation of health cost analysis; section “[Methods for Mean Inference](#)” is about regression models, which is a complicated version of mean inference, and here covariates are taken into consideration; section “[Regression](#)” is a brief introduction on prediction models and some important concepts about prediction models.

---

## Methods for Mean Inference

Methods and theorems are developed to summarize the distribution of health cost data which, as described in the previous section, does not have “nice” properties that we usually assume to be true. The choice of quantity that summarizes the distribution – or, in other words, the summary measure – should be considered on the base of statistical convenience as well as scientific importance. For example, the sample median is known to be a better summary measure for the central location of a skewed distribution than sample mean, but investigators care about the total cost instead of the median cost in most of the time. As will be shown later, a bunch of methods were proposed to find consistent and efficient estimators for the population mean.

Generally speaking, methods with more assumptions perform better than others when the assumptions hold or not being violated too much. Study has shown that using models with inappropriate assumptions on certain data would result in disastrous estimators (Briggs et al. 2005). Some methods depend on few or no assumptions, which can be called robust models, but these methods are often low in efficiency. As the famous quote says “All models are wrong, but some are useful” (Box 1976). The choice of models is important especially in health cost data where the samples behave poorly, though no clear boundary can be drawn in making this decision. It is recommended to check the assumptions when applying certain methods.

Depending on the target population, medical costs have two possible distributions. It might be a continuous distribution with positive values when

the population is defined as subjects who received treatments and paid for them. Such population is interesting for the study of the revenue of a department. It might also be a distribution with a point mass at zero, when the population is defined as a certain group of people like citizens in a city, people in an insurance plan, etc. This kind of distribution is named zero-inflated distribution or delta distribution by Aitchison (1955) The first situation can be seen as a special case of the second one where the point mass at zero is 0. Hence, methods for continuous distributions can be used in the zero-inflated distribution with some modifications. This section will begin with discussions on continuous distributions and then proceed to the case with positive point mass at zero.

## Parametric Methods on Continuous Data

As a classic way of doing statistical analysis, the distribution of data is sometimes assumed to be known and has finite parameters that characterize the distribution. This kind of assumptions is called parametric assumption. For instance, normality is a well-known example of parametric assumption, in which the distribution is characterized by two parameters, the expectation and variance. Unfortunately, this normality assumption does not apply for medical cost data, which is often highly right skewed. A common practice is to transform the data into a more well-behaved form. And then it is possible to assign the normality assumption or some other parametric models on the transformed data.

Box (1976) proposed a family of transformations that can be modified to fit in various situations:

$$\begin{aligned} \frac{y^\lambda - 1}{\lambda} &= x\beta + \varepsilon, & \text{if } \lambda \neq 0; \log(y) \\ &= x\beta + \varepsilon, & \lambda = 0, \end{aligned} \quad (1)$$

where  $y$  is the original dependent variable,  $x$  is a row vector of covariates,  $\varepsilon$  is an additive error term that is independent of the covariates  $x$  and  $\beta$ ,

and  $\lambda$  are parameters to be estimated. Box (1976) stated that under an appropriate transformation, the error term can be approximated by a normal distribution or at least more symmetric than the original scale. Notice that (1) has a dependent variable  $y$  and covariate  $x$  in the formula, and the mean inference is a degenerated version of it where  $x$  is set to be a row of 1 s.

Notice that when  $\lambda$  is set to be zero, the transformation is taking the logarithm of  $y$ . The log transformation is the most widely used transformation in analysis of expenditure data, not only because it reduces the skewness of samples but also because of its real-world interpretations. Manning (1998) gave several rationales for using log transformation in his articles: “(1) A desire for multiplicative or proportional responses to a covariate of interest;... (2) a desire to generate an estimate that easily yields an elasticity; ... or (5) a need to deal with dependent variables that are badly skewed to the right.”

The same reasoning is applicable for medical cost. The expenditures for users are implemented with a log transformation to reduce the skewness inherent in health expenditure data. Under certain circumstances (see Duan 1983), inferences based on logged models are much more precise and robust than direct analysis of original dependent variable. Another attractive property of log transformation is that it has an explicit expression for the untransformed expectation. The expectation of dependent variable  $y$  (untransformed) in a log model is

$$E(y|x) = e^{x\beta} \int e^\varepsilon dF(\varepsilon). \tag{2}$$

If, after transformation, the residuals follow a normal distribution, then the expected value of  $y$  can be written down by straight forward calculations:

$$E(y|x) = \exp(x\beta + 0.5\sigma^2(x)). \tag{3}$$

Notice that (3) shows us that the untransformed mean is a function of both transformed mean and variance.

**Point Estimate**

Several articles in the past decade have been published in searching for efficient estimators of (3). Some of them are well established and have been tested by time (see Zhou 1998). Before proceeding to discuss these methods, there are a few notations that need to be set up. As in (2),  $\{Y_1, \dots, Y_n\}$  is a random sample from a lognormal distribution with mean  $\theta$  and variance  $\tau^2$ . Define  $W_i = \log(Y_i), \forall_i \in (1, \dots, n)$ . Then  $\{W_1, \dots, W_n\}$  comes from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . Let  $\bar{W}$  be defined as  $\sum_{i=1}^n W_i/n$ , and  $S^2$ , the sample variance, be  $\frac{1}{n-1} \sum_{i=1}^n (W_i - \bar{W})^2$ .

1. Maximum Likelihood Estimator (MLE)

$$\hat{\theta}_m = \exp\left(\bar{W} + 0.5 \frac{n-1}{n} S^2\right) \tag{4}$$

The MLE for lognormal distribution is a biased estimator. And the bias is

$$E(\hat{\theta}_m - \theta) = \theta \left( \exp\left(\frac{-n+1}{2n} \sigma^2\right) \left(1 - \frac{\sigma^2}{n}\right)^{-(n-1)/2} - 1 \right),$$

if  $0 < \sigma^2 < n$ . The corresponding mean square error is

$$E(\hat{\theta}_m - \theta)^2 = \theta^2 \left( \exp\left(-\frac{n-2}{2n} \sigma^2\right) \left(1 - 2\frac{\sigma^2}{n}\right)^{-(n-1)/2} - 2\exp\left(\frac{-n+1}{2n} \sigma^2\right) \left(1 - \frac{\sigma^2}{n}\right)^{-(n-1)/2} + 1 \right), \tag{5}$$

when  $0 < \sigma^2 < n/2$ . The MSE in (5) can be estimated by plugging in the estimators of  $\sigma^2$  and  $\mu$ , which are  $S^2$  and  $\bar{W}$ , respectively.

2. Uniformly Minimum-Variance Unbiased Estimator (UMVUE)

$$\hat{\theta}_u = \exp(\bar{W}) g_n(S^2/2), \tag{6}$$

where



$$g_n(t) = \sum_{r=0}^{\infty} \left(\frac{1}{r!}\right) \left(\frac{n-1+2r}{n-1}\right) \times \left(\frac{(n-1)^r}{n}\right)^r \prod_{i=1}^r \left(\frac{n-1}{n-1+2i}\right). \tag{7}$$

It can be tell from its name that  $\hat{\theta}_u$  is an unbiased estimator for  $\theta$ . The mean square error for  $\hat{\theta}_u$  is

$$E(\hat{\theta}_u - \theta)^2 = \theta^2 \left( \exp\left(\frac{1}{n}\sigma^2\right) g_n\left(\frac{1}{2n}\sigma^4\right) - 1 \right). \tag{8}$$

3. Conditionally Minimal Mean Squared Error (MSE) Estimator

$$\hat{\theta}_c = \exp(\bar{W}) g_n\left(\frac{n-4}{2n-2} S^2\right), \tag{9}$$

where  $g_n$  is the same as defined in (7). This estimator is biased; the bias is

$$E(\hat{\theta}_c - \theta) = \theta \left( \exp\left(-\frac{3}{2n}\sigma^2\right) - 1 \right).$$

The MSE of  $\hat{\theta}_c$  is

$$E(\hat{\theta}_c - \theta)^2 = \theta^2 \left( \exp\left(\frac{-2}{n}\sigma^2\right) g_n\left(\frac{(n-4)^2}{2n(n-1)^2}\sigma^4\right) - 2\exp\left(-\frac{3}{2n}\sigma^2\right) + 1 \right). \tag{10}$$

Simulation results by Zhou (1998) show that the conditionally minimal MSE estimator  $\hat{\theta}_c$  is uniformly superior to the alternatives. However, MSEs of those estimators are almost the same when the sample size is sufficiently large ( $n \geq 200$ ). In this case, the MLE  $\hat{\theta}_m$  is recommended because it is easy to compute. With a small sample size, the conditionally minimal MSE estimator  $\hat{\theta}_c$  is more preferable than others.

**Confidence Intervals**

The construction of confidence intervals is more straightforward than the estimators, due to the fact that quantiles are invariant under monotone transformation. The confidence intervals of  $\ln(\theta)$  can

be turned into the CIs of  $\theta$  by simply exponentiating the lower and upper bounds. Recall that  $\{W_i = \log(Y_i)\}$  are normally distributed, so  $\bar{W} + \frac{S^2}{2}$  is the UMVU estimator for  $\ln(\theta)$ . The target now is to estimate the confidence interval of  $\bar{W} + \frac{S^2}{2}$ . Zhou and Gao (1997) summarized several practical procedures with median or large sample sizes. Krishnamoorthy and Mathew (2003) applied the general pivotal quantity on this issue and got asymptotically efficient estimators for the confidence intervals.

In general, one cannot use confidence intervals to make statistical inference as they have slight differences in between them. But in this simple case of one-sample mean inference, hypothesis testing is equivalent with testing whether the mean under null hypothesis lies inside the 100(1 -  $\alpha$ )% confidence intervals or not. And thus, a more desirable confidence interval will be a more reliable approach of hypothesis testing.

Notice that the 100(1 -  $\alpha$ )% confidence intervals can also be used in hypothesis testing under this one-sample setting. The null hypothesis will be rejected with significant level  $\alpha$  when the null mean lies outside of the confidence intervals.

1 Cox’s method: The estimator for the variance of  $\bar{W} + S^2/2$  is  $S^2/n + S^4/(2(n - 1))$ . Cox, in a personal communication to Land (1972), proposed to construct the confidence intervals for  $\ln(\theta)$  by

$$\bar{W} + \frac{S^2}{2} \pm Z_{1-\alpha/2} \sqrt{\frac{S^2}{2} + \frac{S^4}{2(n-1)}}, \tag{11}$$

where  $Z_{1-\alpha/2}$  is the 100(1 -  $\alpha$ )% quantile of a standard normal distribution, i.e., normal distribution with mean zero and standard deviation of 1. The corresponding confidence intervals for  $\theta$  is

$$\left( \exp\left\{ \bar{W} + \frac{S^2}{2} - Z_{1-\alpha/2} \sqrt{\frac{S^2}{2} + \frac{S^4}{2(n-1)}} \right\}, \exp\left\{ \bar{W} + \frac{S^2}{2} + Z_{1-\alpha/2} \sqrt{\frac{S^2}{2} + \frac{S^4}{2(n-1)}} \right\} \right).$$

2. Angus’s conservative method: Although the exact pivotal quantity is not available in this

problem, an approximate pivotal statistics is available as

$$V(\theta) = \frac{\sqrt{n}(\bar{W} + S^2/2 - \ln(\theta))}{\sqrt{S^2(1 + S^2/2)}}, \tag{12}$$

which, in a finite sample, has the same distribution as

$$T(\nu) = \frac{N + \sigma \frac{\sqrt{n}}{2} (\chi_{n-1}^2/(n-1) - 1)}{\sqrt{\frac{\chi_{n-1}^2}{n-1} \left(1 + \frac{\sigma^2 \chi_{n-1}^2}{2(n-1)}\right)}}, \tag{13}$$

where  $N$  and  $\chi_{n-1}^2$  are independent random variables from a standard normal distribution and a  $\chi^2$  distribution with  $n-1$  d.f., respectively. The conservative CIs are

$$L_{1-\alpha} = \bar{W} + \frac{S^2}{2} - \frac{t_{1-\alpha/2}(n-1)}{\sqrt{n}} \times \sqrt{S^2 \left(1 + \frac{S^2}{2}\right)}, \tag{14}$$

$$U_{1-\alpha} = \bar{W} + \frac{S^2}{2} - \frac{q_{\alpha/2}(n-1)}{\sqrt{n}} \times \sqrt{S^2 \left(1 + \frac{S^2}{2}\right)}, \tag{15}$$

where  $q_{\alpha/2}(n-1) = \sqrt{\left(\frac{n}{2}\right) \left\{ \frac{n-1}{\chi_{\alpha/2}^2(n-1)} - 1 \right\}}$ .

Then the  $100(1 - \alpha)\%$  confidence intervals for  $\theta$  is  $(\exp(L_{1-\alpha}), \exp.(U_{1-\alpha}))$ . This approach is called conservative because the probability that  $\ln(\theta)$  falls into the CIs is no less than  $1 - \alpha$ .

3. Parametric bootstrap method: Notice that in (13),  $T$  is determined by  $N$  and  $\chi_{n-1}^2$ . Though  $T$  itself is hard to generate,  $N$  and  $\chi_{n-1}^2$  come from two simple distributions. It is possible to get samples of  $T$  by generating a series of  $N$  and  $\chi_{n-1}^2$ . Suppose  $N_i^* \sim N(0, 1)$  and  $\chi_i^{2*} \sim \chi_{n-1}^2, i = 1, \dots, B$ , where  $B$  is a sufficiently large number. Then

calculate  $T_i^*$  as in (13), and denote the  $t_l^*$  as the  $1 - \alpha/2$  empirical quantile and  $t_u^*$  as the  $\alpha/2$  empirical quantile. The estimated bounds are

$$L_{1-\alpha} = \bar{W} + \frac{S^2}{2} - \frac{t_l^*}{\sqrt{n}} \sqrt{S^2 \left(1 + \frac{S^2}{2}\right)}, \tag{16}$$

$$U_{1-\alpha} = \bar{W} + \frac{S^2}{2} - \frac{t_u^*}{\sqrt{n}} \sqrt{S^2 \left(1 + \frac{S^2}{2}\right)}. \tag{17}$$

So the  $100(1 - \alpha)\%$  confidence intervals for  $\theta$  is  $(\exp(L_{1-\alpha}), \exp.(U_{1-\alpha}))$ .

4. A signed likelihood ratio approach: Wu et al. (2003) used the log-likelihood ratio to construct confidence intervals. The signed log-likelihood ratio  $r$  is defined as

$$r(m) = \text{sgn}(\hat{m} - m) \{2[l(\hat{m}, \hat{\sigma}^2) - l(m, \hat{\sigma}_m^2)]\}^{1/2}. \tag{18}$$

The log-likelihood as a function of  $m = \log(\theta)$  and  $\sigma^2$  is

$$l(m, \sigma^2) = -\frac{n}{2} \log(\sigma^2) + \left(m - \frac{1}{2}\sigma^2\right) \frac{\sum_{i=1}^n Y_i}{\sigma^2} - \frac{1}{2\sigma^2} \sum_{i=1}^n Y_i^2 - \left(m - \frac{1}{2}\sigma^2\right)^2 \frac{n}{2\sigma^2}.$$

The overall MLE is

$$\hat{\sigma}^2 = \bar{w}_2 - \bar{w}_1^2, \quad \hat{m} = \bar{w}_1 + \frac{1}{2}\hat{\sigma}^2,$$

where  $\bar{w}_1 = \frac{1}{n} \sum_{i=1}^n W_i, \bar{w}_2 = \frac{1}{n} \sum_{i=1}^n W_i^2$ . For a fixed  $m$ , the MLE of  $\sigma^2$  is

$$\hat{\sigma}_m^2 = 2 \left[ (m + 1)^2 + \bar{w}_2 - 2m\bar{w}_1 - 2m \right]^{1/2} - 2.$$

Thus, the specified form of signed log-likelihood ratio  $r$  is

$$r(m) = \text{sgn}(\hat{m} - m) \times \left\{ n \log \frac{\hat{\sigma}_m^2}{\sigma^2} + n(\bar{\omega}_1 - m + \hat{\sigma}_m^2/2) \right\}^{1/2} \tag{19}$$

The  $100(1 - \alpha)\%$  confidence intervals would be  $(\exp(\hat{m}_{\alpha/2}), \exp(\hat{m}_{1-\alpha/2}))$ , where  $\hat{m}_\alpha$  is the solution of  $r(m) = z_\alpha$  - the  $\alpha$ th quantile of a standard normal distribution. The equation has no explicit solution, but it can be approached by numerical methods such as Newton-Raphson method, etc. An example of constructing the lower bound is given by (Wu et al. 2003):

- i). Set up accuracy  $\varepsilon$ , differentiation constant  $\delta$ , and initial value  $m_0$ .
- ii). Estimate  $m^*$  as

$$m^* = m_0 + \frac{z_{\alpha/2} - r(m_0)}{[r(m_0 + \delta) - r(m_0 - \delta)]/(2\delta)}.$$

- iii). Substitute  $m_0$  with  $m^*$  if  $|m^* - m_0| > \varepsilon$  and repeat step (ii) again.

The construction of upper bound is basically the same except for replacing  $z_{\alpha/2}$  with  $z_{1-\alpha/2}$ .

5. An adjusted signed log-likelihood ratio approach: Wu et al. (2003) also proposed a modified version of the signed likelihood ratio statistics. They defined an adjusted signed log-likelihood statistics as

$$r^*(m) = r(m) + r^{-1}(m) \log \left[ \frac{u(m)}{r(m)} \right],$$

where  $r(m)$  is defined as in the previous section. The  $u(m)$  here is a function of  $m$  defined as

$$u(m) = \sqrt{n}(\hat{m} - m) \left( \frac{\hat{\sigma}}{\hat{\sigma}_m^3} \right) \left( \frac{1}{2} + \frac{1}{\hat{\sigma}_m^2} \right)^{-1/2}. \tag{20}$$

The  $100(1 - \alpha)\%$  confidence intervals can be constructed in the same fashion of  $r(m)$ 's:  $(\exp(\hat{m}_{\alpha/2}), \exp(\hat{m}_{1-\alpha/2}))$ , where  $\hat{m}_\alpha$  is the solution of  $r^*(m) = z_\alpha$ . The equations are solved with

the same algorithm described in last section, with simply replacing  $r$  with  $r^*$ .

6. A generalized pivot approach: Krishnamoorthy and Mathew (2003) applied the concept of generalized pivotal quantity on lognormal means. The generalized pivotal quantity can be viewed as a new concept of hypothesis test, and it yields the same coverage rate as a standard frequentist hypothesis testing asymptotically. For more details about generalized pivotal quantity, or fiducial quantity, please see the appendix of this chapter. The generalized pivot for  $\ln(\theta)$  is given by

$$T = \bar{W} - \frac{Z}{U/\sqrt{n-1}} \frac{S}{\sqrt{n}} + \frac{1}{2} \frac{S^2}{U^2(n-1)}, \tag{21}$$

where  $Z \sim N(0, 1)$ ,  $U^2 \sim \chi_{n-1}^2$ , and  $Z$  and  $U^2$  are independent. The same approach in parametric bootstrap is used here to estimate the generalized confidence intervals. Suppose  $Z_i^* \sim N(0, 1)$  and  $U_i^{2*} \sim \chi_{n-1}^2, i = 1, \dots, B$ , where  $B$  is a sufficiently large number. Then calculate  $T_i^*$  as in (21), and denote the if as the  $\alpha/2$  empirical quantile and  $t_u^*$  as the  $(1 - \alpha/2)$  empirical quantile. The estimated bounds are  $(t_l^*, t_u^*)$ . So the  $100(1 - \alpha)\%$  generalized confidence interval for  $\theta$  is  $(\exp(t_l^*), \exp(t_u^*))$ . However, as being pointed out by Krishnamoorthy and Mathew (2003), the type I error and the power of such a test might depend on unknown parameters. It would be necessary to simulate type I error probability in order to see whether the test controls type I error.

Simulation results by Zhou and Gao (1997) show that Cox's method has the best performance in moderate to large samples, in terms of both computational simplicity and statistical efficiency. And thus Cox's method is recommended when sample size is sufficiently enough. With small sample size, the parametric bootstrap method provides the most satisfactory confidence interval among the methods examined in Zhou and Gao (1997). However, Krishnamoorthy and Mathew (2003) showed that the generalized pivot approach has better performance than the parametric bootstrap approach in one-side hypothesis testing with small sample size. In a following simulation by Wu et al. (2003), the authors

showed that the adjusted signed log-likelihood ratio-based method provided the most satisfactory coverage probability and average biases. Although the computation of adjusted signed log-likelihood ratio approach is way more complicated than others, it is recommended when the sample size is too small for other methods.

It is worthwhile to notice that all methods above are based on the lognormal assumption, which requires the log-transformed data to be normally distributed. Although the estimators still behave well when the log-transformed data is approximately normally distributed, Briggs et al. (2005) argued that the inference would be invalid and misleading when the sample distribution extremely deviants from the assumed distribution. Hence, checking the normality (with QQ plot, goodness of fit, etc.) of transformed data is always necessary. When the normality assumption is not appropriate, other distributions such as Gamma are available. And it is always possible to trade efficiency for robustness via using nonparametric methods which will be introduced later.

**Nonparametric Methods on Continuous Data**

It is totally possible to estimate  $\theta$  and confidence interval and do hypothesis testing without parametric assumptions. Although the efficiency is often not satisfactory, the central limit theorem granted that the sample mean would converge to a normal distribution.

Denote the sample mean as

$$\hat{\theta}_s - \bar{Y}. \tag{22}$$

Central limit theorem and Slutsky’s theorem grant that

$$\frac{1}{s_n} (\hat{\theta}_s - \theta) \rightarrow N(0, 1),$$

where  $s_n$  is the standard error of  $\bar{X}$ . There are two ways to estimate this standard error, both of which are straightforward.

The first one is to estimate it directly from the sample standard error. In other words, the standard error  $s_b$  is the square root of  $\frac{1}{n(n-1)} \sum_{i=1}^n (Y_i - \bar{Y})$ .

The second one is to use the bootstrap approach proposed by Efron (1981). The algorithm can be summarized as below:

1. Resample  $n$  observations from the original data with equal weight and replacement.
2. Calculate the sample mean from the newly sampled data, denoted as  $\theta_s^i$ .
3. Repeat steps 1 and 2 for  $M$  times, where  $M$  is a sufficiently large number chosen by the investigator.
4. Calculate the standard error  $s_b$  of  $\{\theta_s^i\}$ .

Based on central limit theorem, the confidence interval would be  $(\bar{Y} + S_b Z_{\alpha/2}, \bar{Y} + S_b Z_{1-\alpha/2})$ , where  $Z_q$  is the  $q$ -th quantile of a standard normal distribution.

Hall (1992) proposed a monotone transformation of  $t$ -statistics to correct for skewness effects of a positive skewed distribution without assuming any parametric forms. The original  $t$ -statistic is

$$T = \frac{\sqrt{n}(\bar{Y} - \theta)}{\hat{\tau}},$$

where  $\hat{\tau} = \frac{1}{n} (Y_i - \bar{Y})^2$ . The transformation is

$$g(T) = T + n^{-1/2} \hat{\gamma} (aT^2 + b) + n^{-1} (a\hat{\gamma})^2 T^3 / 3, \tag{23}$$

where  $\hat{\gamma} = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^3 / \hat{\tau}^3$ ,  $a = 1/3$ , and  $b = 1/6$ . It is monotone and invertible. The unique inverse function of  $g$  is

$$T = g^{-1} = n^{1/2} (a\hat{\gamma})^{-1} \left[ (1 + 3a\hat{\gamma} (n^{-1/2}x - n^{-1}b\hat{\gamma}))^{1/3} - 1 \right]. \tag{24}$$

There are several ways to construct the confidence intervals based on the proposed  $g$  function. It was shown in Zhou and Gao (2000) that the bootstrap approach has the best performance. The algorithm is similar to what has been discussed:

1. Resample  $n$  observations from the original data with equal weight and replacement.
2. Calculate  $g(t)$  from the newly sampled data.
3. Repeat steps 1 and 2 for  $M$  times, where  $M$  is a sufficiently large number chosen by the investigator.
4. Denote the sample  $\alpha/2$  and  $1 - \alpha/2$  quantiles as  $g_{\alpha/2}^*$  and  $g_{1-\alpha/2}^*$ .

The resulting  $1 - \alpha$  two-sided confidence intervals are  $\left(\bar{Y} - n^{-1/2} \hat{\tau} g^{-1}(g_{1-\alpha/2}^*), \bar{Y} - n^{-1/2} \hat{\tau} g^{-1}(g_{\alpha/2}^*)\right)$ .

Zhou and Gao (2000) recommended the application of parametric bootstrap version of Hall’s method, which yields the best coverage rate for both upper and lower endpoints in a simulation study of one-sided confidence intervals.

In another simulation study, Zhou (1998) showed that the sample mean has relatively large mean square error even when the sample size is as large as 200, compared to other estimators discussed in the previous section. And the mean square error increases as  $\sigma$  increases, which is equivalent to say as the skewness increases. It is important to notice that the simulation study is conducted on lognormal data, where the lognormal assumption actually holds. This might explain part of the bad performance of sample mean compared to estimators based on lognormal assumption. Yet the efficiency of sample mean on skewed data is still very low.

**Zero-Inflated Data**

As discussed, medical cost data is often accompanied with a considerable amount of observations that have zero cost. The proportion of zero data might sometimes reach 30%. This point mass at zero causes extra difficulty in making statistical inference, but it could be easily fixed with small modifications of the methods used on continuous data. The nonparametric methods described in the previous section, in fact, need no modifications at all and can be used directly in this situation. For instance, the sample mean is a nonparametric estimator of the population mean, and bootstrap would give a confidence interval for it. So they will not be discussed in this section anymore, and the focus will be placed on parametric methods.

A most commonly used parametric model for zero-inflated data is a two-part model. A two-part model assumes that the number of zero observations is a random variable from a binomial distribution  $bin(n, p)$ , where  $n$  is the number of observations and  $p$  is the probability of one subject to have zero medical cost in study period. The nonzero observations, conditioned on the fact that they are nonzero, are treated as the continuous data discussed in previous sections. The conditional distribution is assumed to be a lognormal distribution in this section.

For each group, the distribution of samples is a lognormal distribution with a point mass at zero, which is named as delta distribution by Aitchison (1955). Suppose  $\{Y_1, \dots, Y_n\}$  is a random sample from a delta distribution, then the population mean is

$$\theta = (1 - p) \exp(\mu + \sigma^2),$$

where  $p$  is the probability of the random variable to be zero and  $\mu$  and  $\sigma$  are mean and variance, respectively, of the conditional normal distribution after transformation. Denote the number of zero observations as  $N_0$ , the number of nonzero observations as  $N_1$ . In this section, the parameter of interest is  $\theta$ , and, again, the construction of confidence intervals of  $\theta$  is also discussed.

**Point Estimate**

1. The MVUE fo  $\theta$  is

$$\hat{\theta}_A = (1 - \hat{p}) \exp(\hat{\mu}) g_n \left(\frac{1}{2} \hat{\sigma}^2\right), \tag{25}$$

where

$$\hat{p} = \frac{N_0}{n},$$

$$\hat{\mu} = \frac{1}{N_1} \sum_{i=1}^{N_1} w_i,$$

and

$$\hat{\sigma}^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} (w_i - \hat{\mu})^2.$$

2. A bias-corrected MLE for  $\theta$  is

$$\hat{\theta}_M = (1 - \hat{p}) \exp\left(\hat{\mu} \frac{1}{2} \hat{\sigma}^2\right). \tag{26}$$

Notice that in (26), the unbiased estimator  $\hat{\sigma}^2$  is used instead of the MLE  $\frac{N_1-1}{N_1} \hat{\sigma}^2$ . That is the reason why it is named a bias-corrected MLE.

**Confidence Intervals**

Several methods have been proposed to construct the confidence intervals.

1. The MVUE intervals: Owen and DeRouen (1980) derived a minimum-variance unbiased estimator (MVUE) confidence interval for the population mean of zero-inflated lognormal distribution. The asymptotic variance of  $\hat{\theta}_A$  is

$$V(\hat{\theta}_A) = n^{-1} \exp(2\hat{\mu} + \hat{\sigma}^2) \times t \left\{ \hat{p}(1 - \hat{p}) + \frac{1}{2}(1 - \hat{p})(2\hat{\sigma}^2 + \hat{\sigma}^4) \right\}.$$

So the  $100(1 - \alpha)\%$  confidence intervals of  $\hat{\theta}_A$  can be asymptotically approximated by

$$\left( \hat{\theta}_A - z_{1-\alpha/2} \sqrt{V}, \hat{\theta}_A + z_{\alpha/2} \sqrt{V} \right).$$

2. The ML confidence intervals: Using delta method and property of MLE, a consistent variance estimator of the bias-corrected MLE,  $\log(\hat{\theta}_M)$ , can be written as

$$\hat{S\hat{E}}^2 = \frac{N_0}{nN_1} + \frac{\hat{\sigma}^2}{N_1} + \frac{\hat{\sigma}^4}{2N_1}.$$

So the two-sided  $100(1 - \alpha)\%$  confidence intervals are

$$\left( \hat{\theta}_M \exp(z_{\alpha/2} \hat{S\hat{E}}), \hat{\theta}_M \exp(z_{1-\alpha/2} \hat{S\hat{E}}) \right).$$

3. A bootstrap approach for ML confidence intervals: Similar to the Angus methods in the previous section, an approximate pivotal statistics can be derived:

$$T = \frac{\log(1 - \hat{p}) + \hat{\mu} + \frac{\hat{\sigma}^2}{2} - \log(1 - p) - \mu - \frac{\hat{\sigma}^2}{2}}{\left\{ \frac{\hat{p}}{n(1-\hat{p})} + \frac{\hat{\sigma}^2}{n(1-\hat{p})} + \frac{\hat{\sigma}^4}{2n(1-\hat{p})} \right\}^{0.5}}. \tag{27}$$

It follows the same distribution as the following statistic:

$$T = \frac{\frac{\sqrt{N_1}}{\sigma} \log\left(\frac{N_1}{n(1-p)}\right) + Z + \frac{\sigma\sqrt{N_1}}{2} \left(\frac{\chi^2_{(N_1-1)}}{N_1} - 1\right)}{\left\{ \frac{n-N_1}{n\sigma^2} + \frac{\chi^2_{(N_1-1)}}{N_1} \left(1 + \frac{\sigma^2\chi^2_{(N_1-1)}}{2N_1}\right) \right\}^{0.5}}, \tag{28}$$

where  $Z$  and  $\chi^2_{(N_1-1)}$  are independent random variables with standard normal distribution and  $\chi^2$  distribution with  $N_1 - 1$  degrees of freedom.

The procedure for bootstrap is to (i) generate the number of zero observations,  $N_0$ , from a binomial distribution  $Bin(n, p)$ , (ii) generate  $Z^*$  and  $\chi^{2*}$  from the distributions described above, (iii) calculate the  $T^*$  with (28); and (iv) repeat i through iii for sufficiently many times and get the sample quantiles  $t_{\alpha/2}$  and  $t_{1-\alpha/2}$ .

So the two-sided  $100(1 - \alpha)\%$  confidence intervals are

$$\left( \hat{\theta}_M \exp(t_{\alpha/2} \hat{S\hat{E}}), \hat{\theta}_M \exp(t_{1-\alpha/2} \hat{S\hat{E}}) \right).$$

4. A signed likelihood ratio approach: The ML confidence intervals are based on the asymptotic normality of MLE, which is questionable with small or moderate samples. An alternative would be the likelihood ratio interval. The log-likelihood as a function of  $m = \log(\theta)$ ,  $\mu$ , and  $\sigma^2$  is

$$l(m, \mu, \sigma^2) = N_0 \log \left\{ 1 - \exp\left(m - \mu - \frac{\sigma^2}{2}\right) \right\} + N_1 \left( \theta - \mu - \frac{\sigma^2}{2} \right) - \frac{N_1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N_1} (w_i - \mu)^2.$$

Since there are nuisance parameters in the log-likelihood, the profile likelihood for  $m$  will be used to compute the likelihood ratio statistics. In general, the way to solve this problem is to (i) use iterative algorithm to find the  $f_i$  and  $a^2$  that maximized the log-likelihood given  $m$  and  $\mu + \sigma^2 > m$ ; (ii) define  $l_{prof}(m) = l(m, \hat{\mu}[m], \hat{\sigma}^2[m])$ , and find the  $m$  that maximizes this profile log likelihood; (iii) define likelihood ratio statistics  $W(m) = 2(l_{prof}(\hat{m}) - l_{prof}(m))$ ; and (iv) define  $Z(m) = \text{sgn}(\hat{m} - m) \sqrt{W(m)}$ .

The  $100(1 - \alpha)\%$  confidence intervals would be  $(\exp(\hat{m}_{\alpha/2}), \exp(\hat{m}_{1-\alpha/2}))$ , where  $\hat{m}_\alpha$  is the solution of  $Z(\hat{m}_\alpha) = z_\alpha$ .

5. An adjusted signed log-likelihood ratio approach: Tian and Wu (2006) proposed a modified version of the signed likelihood ratio statistics. They defined an adjusted signed log-likelihood statistics as

$$Z^*(m) = Z(m)Z^{-1}(m)\log\left[\frac{u(m)}{Z(m)}\right],$$

where  $Z(m)$  is defined as in the previous section:  $\text{sgn}(\hat{m} - m)\sqrt{W(m)}$ . The  $u(m)$  here is complicated:

$$u(m) = \frac{A}{B} \frac{C}{D} \tag{29}$$

where  $A, B, C,$  and  $D$  are

$$\begin{aligned} A &= \left(\hat{a}_m + 1 + \frac{\hat{\mu}(m)}{\hat{\sigma}^2(m)}\right) \left[\frac{\hat{\mu}(m) - \hat{\mu}}{2\hat{\sigma}^2\hat{\sigma}^4(m)}\right] \\ &+ \left(\frac{-1}{\hat{\sigma}^2(m)}\right) \left(\frac{1}{2\hat{\sigma}^4(m)}\right) \times \left\{\log\left(\frac{N_1}{N_0\hat{a}_m}\right) - \frac{1}{2}\log\left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2(m)}\right) - \frac{\hat{\mu}^2}{2\hat{\sigma}^2} + \frac{\hat{\mu}^2(m)}{2\hat{\sigma}^2(m)}\right\} \\ &+ \frac{1}{\hat{\sigma}^2(m)} \left[\frac{\hat{a}_m}{2} + \frac{1}{2} + \frac{1}{2\hat{\sigma}^2(m)} - \frac{\hat{\mu}^2(m)}{2\hat{\sigma}^4(m)}\right] \\ &\times \left[\frac{1}{2\hat{\sigma}^2} - \frac{1}{2\hat{\sigma}^2(m)}\right], \\ B &= \frac{-n}{2N_0\hat{\sigma}^6}, \\ C &= \frac{-nN_1^3}{2N_0\hat{\sigma}^6}, \\ D &= \frac{-N_0\hat{b}_m - N_1}{\hat{\sigma}^2(m)} \left[-\frac{N_0\hat{b}_m}{4} + \frac{N_1}{2\hat{\sigma}^4(m)} - \frac{T}{\hat{\sigma}^6(m)} + \frac{2\hat{\mu}(m)\sum_{i=1}^{N_1}W_i}{\hat{\sigma}^6(m)} - \frac{N_1\hat{\mu}^2(m)}{\hat{\sigma}^6(m)}\right] \\ &- \left(\frac{N_0\hat{b}_{mm}}{2} + \frac{N_1\hat{\mu}(m)}{\hat{\sigma}^4(m)} - \frac{\sum_{i=1}^{N_1}W_i}{\hat{\sigma}^4(m)}\right)^2, \end{aligned}$$

where

$$\begin{aligned} \hat{a}_m &= \frac{\exp\left(m - \hat{\mu}(m) - \frac{\hat{\sigma}^2(m)}{2}\right)}{1 - \exp\left(m - \hat{\mu}(m) - \frac{\hat{\sigma}^2(m)}{2}\right)}, \\ \hat{b}_m &= \frac{\hat{a}_m}{1 - \exp\left(m - \hat{\mu}(m) - \frac{\hat{\sigma}^2(m)}{2}\right)}, \\ T &= \sum_{i=1}^{N_1} W_i^2. \end{aligned}$$

The  $100(1 - \alpha)\%$  confidence intervals would be  $(\exp(\hat{m}_{\alpha/2}), \exp(\hat{m}_{1-\alpha/2}))$ , where  $\hat{m}_\alpha$  is the solution of  $Z^*(\hat{m}_\alpha) = z_\alpha$ .

6. A generalized pivot approach: Tian (2005) applied the concept of generalized confidence intervals on the zero-inflated data. Recall that the models are almost the same except for the excess zeros. Tian derived a generalized pivot for  $p$  using the relationship between binomial distribution and beta distribution. The author also provided a computing algorithm for this method:

- i). Compute the transformed sample mean  $\bar{W}$  and sample variance  $S^2$ .
- ii). Generate  $Z \sim N(0,1), U^2 \sim \chi_{N_1-1}^2, T_{p_1} \sim \text{beta}(N_0 + 1, N_1)$ , and  $T_{p_2} \sim \text{beta}(N_0, N_1 + 1)$ . Compute  $T_\theta = \bar{W} - \left[Z/U/\sqrt{N_1 - 1}\right]s/\sqrt{N_1} + \left[S^2/U^2/(N_1 - 1)\right]$ . Then compute  $T_1 = \log(1 - T_{p_1}) + T_\theta$  and  $T_2 = \log(1 - T_{p_2}) + T_\theta$ .
- iii). Repeat ii for sufficiently many times and get a series of  $T_1$ 's and  $T_2$ 's.
- iv). Take the  $\alpha/2$  sample quantile of  $T_1$ 's, denoted as  $L$ , and take the  $(1 - \alpha/2)$  sample quantile of  $T_2$ 's, denoted as  $U$ . The  $100(1 - \alpha)\%$  confidence intervals would be  $(L, U)$ .

The simulation by Zhou and Tu (2000) showed that the bootstrap interval yields the best coverage probability among the first four methods in small to moderate samples, although bias-corrected ML has better accuracy when the skewness is very small. Tian (2005) verified that the generalized confidence intervals provide comparable results as the first four methods. Based on Tian's simulation,

the generalized confidence intervals seem to be anti-conservative, and its performance is actually worse than other methods when the skewness is low. But when the data is highly skewed, say  $\sigma = 10$ , the coverage probability of generalized confidence intervals is closed to the true value. The adjusted signed log-likelihood method has the best performance among all these methods based on the results of Tian and Wu (2006), although no direct comparison has been made between adjusted signed log-likelihood ratio-based intervals and generalized confidence intervals. Another aspect to be considered is the computation difficulty. The likelihood-based methods both are more difficult to compute than other methods, as can be seen from the description of methods.

**Two Sample**

Before proceeding to discuss these methods, there are a few notations needed to be set up.  $\{Y_{j,1}, \dots, Y_{j,n_j}\}, j = 1, 2$  is now two sets of observations from distributions with mean  $\theta_j$  and variance  $\tau_j^2$ , respectively. Define  $W_{i,j} = \log(Y_{i,j}), \forall i \in (1, \dots, n_j), j = 1, 2$ , and denote the variance of  $W_{i,j}$  as  $\sigma_j^2$ , mean as  $\mu_j$ , for  $j = 1, 2$ . Let  $\bar{W}_j$  be defined as  $\sum_{i=1}^{n_j} W_{i,j}$ , and  $S_j^2$ , the sample variance, be  $\frac{1}{n-1} \sum_{i=1}^n (W_{i,j} - \bar{W}_j)^2$ .

The difference between two population means is  $\delta = \theta_1 - \theta_2$ , which is the parameter of interest in this section. With parametric assumption, i.e., lognormal assumption,  $\delta$  can be further specified. Under lognormal assumption,  $\{W_{j,1}, \dots, W_{j,n_j}\}$  come from a normal distribution with mean  $\mu_j$  and variance  $\sigma_j^2, j = 1, 2$ . And the difference of two lognormal means is

$$\begin{aligned} \delta &\equiv \theta_1 - \theta_2 \\ &= \exp\left(\mu_1 + \frac{1}{2}\delta_1^2\right) - \exp\left(\mu_2 + \frac{1}{2}\delta_2^2\right). \end{aligned} \quad (30)$$

**Point Estimate**

1. Mean difference: A straightforward estimator of the mean difference would be the difference of the sample means

$$\bar{\delta} = \bar{Y}_1 - \bar{Y}_2. \quad (31)$$

2. The maximum likelihood estimator: When lognormal assumption is appropriate, the MLE of  $\delta$  is available in the form of

$$\hat{\delta} = \exp\left(\hat{\mu}_1 + \frac{1}{2}\hat{\sigma}_1^2\right) - \exp\left(\hat{\mu}_2 + \frac{1}{2}\hat{\sigma}_2^2\right).$$

The asymptotic variance of MLE will be given in (37).

3. Smooth quantile estimation: Dominici et al. (2005) proposed a new kind of smoothing estimator which needs no parametric assumptions. They called it smooth quantile ratio estimator.

Step 1. Estimate  $\beta$  in

$$\log \frac{y_{1(i)}}{y_{2(i)}} = s(p_i, \beta) + \varepsilon_i, i = 1, \dots, n, \quad (32)$$

where  $s(p_i, \beta) = \sum_{j=0}^{\lambda} B_j(p_i)\beta_j, p_i = i/(n+1)$  and  $B_j(p)$  are orthonormal basis functions with  $B_0(p) = 1$ . If the sample size is imbalanced, say  $n_1 > n_2$ , a tiny modification is needed: replace  $y_2$  by  $q_2$ , the linear interpolant of the order statistics  $y_{2(i)}$  at the grid of points  $p_{1i} = i/(n_1 + 1), i = 1, \dots, n_1$ . The choice of  $s$  is rather flexible, for instance, natural cubic splines, smoothing splines, and polynomials are all available choices. The simulation study by Dominici et al. (2005) showed that the estimates are quite close to each other.

Step 2. Define  $u_1 = (y_{1(1)}, \dots, y_{1(n)}, y_{1(1)}^*, \dots, y_{1(n)}^*)$  and similar with  $u_2$ , where  $y_{1(i)}^* = y_{2(i)} \exp\{s(p_i, \hat{\beta})\}, y_{2(i)}^* = y_{1(i)} \exp\{s(p_i, \hat{\beta})\}$ .

And estimate  $\Delta$  by

$$\hat{\Delta}_{SQ}(u_1, u_2, \lambda) = \bar{u}_1 - \bar{u}_2. \quad (33)$$

Notice that it is symmetric in the two samples. Furthermore, it can be viewed as a linear combination of order statistics, but with weights estimated from the data, and thus it is related to L-estimation.

The authors show that under mild conditions, the proposed estimator is asymptotically normal. In other words,  $\sqrt{n}(\hat{\Delta} - \Delta)$  is asymptotically normal with mean 0 and variance  $\sigma_{\Delta}^2$ . The asymptotic variance is given by



$$\sigma_{\Delta}^2 = \int_{p=0}^1 \int_{q=0}^1 \{\min(p, q) - pq\} \{\lambda_1 \eta_1(p) \eta_1(q) + \lambda_2 \eta_2(p) \eta_2(q)\} dpdq,$$

where

$$\eta_k(p) = \frac{F_1^{-1}(p) + \frac{1}{2} \left[ F_1^{-1}(p) + F_2^{-1}(p) - \int_0^1 \sum_{j=1}^{\lambda} B_j(q) \{F_1^{-1}(q) + F_2^{-1}(q)\} dq \right]}{(-1)^g F_g^{-1}(p) f_g(F_g^{-1}(p))}.$$

The estimation is achieved by substituting all unknown values with their empirical estimates.

In a simulation study, Dominici et al. (2005) showed that  $\hat{\Delta}(\lambda = 2)$  has more robust performance than the MLE of lognormal distribution, and it yields almost the same result when the parametric assumption is met. The choice of  $\lambda$  can also be made by using cross validation. However, the computation of quantile smooth estimation, especially its asymptotic variance, is rather difficult compared to those of MLE.

**Confidence Intervals**

With no parametric assumption, one can use bootstrap or the asymptotic distribution of smooth quantile ratio estimator to construct the confidence intervals for the corresponding estimators. There are various ways to construct the confidence intervals when lognormal assumption is applied.

1. A maximum likelihood approach: The maximum likelihood estimate for  $\delta$  is

$$\hat{\delta} = \exp\left(\hat{\mu}_1 + \frac{1}{2}\hat{\sigma}_1^2\right) - \exp\left(\hat{\mu}_2 + \frac{1}{2}\hat{\sigma}_2^2\right). \quad (34)$$

where

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} W_{ij}; \quad (35)$$

$$\hat{\sigma}_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (W_{ij} - \hat{\mu}_i)^2. \quad (36)$$

It is known that the asymptotic variance of MLE achieves the variance bound given by

$$\hat{v}^2 = h(\hat{\theta})' I^{-1} h(\hat{\theta}), \quad (37)$$

where  $I$  is the Fisher information matrix and  $\hat{I}$  denotes its estimator:

$$\begin{pmatrix} n_1/\hat{\sigma}_1 & 0 & 0 & 0 \\ 0 & n_1/(2\hat{\sigma}_1^2) & 0 & 0 \\ 0 & 0 & n_1/\hat{\sigma}_1 & 0 \\ 0 & 0 & 0 & n_1/(2\hat{\sigma}_1^2) \end{pmatrix}$$

The function  $h$  is defined as the partial derivative of  $\delta$  with respect to  $\varphi = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ :

$$h(\theta) = \frac{\partial \delta}{\partial \theta} = \left( m_1, \frac{1}{2} m_1, -m_2, -\frac{1}{2} m_2 \right)', \quad (38)$$

where  $m_1 = \exp(\mu_1 + \frac{1}{2}\sigma_1^2)$  and  $m_2 = \exp(\mu_2 + \frac{1}{2}\sigma_2^2)$ . The  $100(1 - \alpha)\%$  confidence interval can be given by

$$[\delta - z_{\alpha/2} \hat{v}, \hat{\delta} + z_{\alpha/2} \hat{v}], \quad (39)$$

where  $z$  comes from a standard normal distribution. Since this is an asymptotic property, this CI can be foreseen to have poor performance in small sample settings.

2. A bootstrap approach: A parametric bootstrap method can be employed to replace the role of asymptotic standard normal distribution. The algorithm is summarized below:

- (1) Compute  $\hat{\mu}_i, \hat{\sigma}_i^2$  and  $\hat{\delta}, \hat{v}$  from the samples of interest.
- (2) General  $n_i$  samples from  $N(\hat{\mu}_i, \hat{\sigma}_i^2), i = 1, 2$ .
- (3) Calculate  $\hat{\delta}_j$  and  $\hat{v}_j$  from the bootstrap sample.
- (4) Compute the test statistic  $t_j = (\hat{\delta}_j - \hat{\delta})/\hat{v}_j$ .
- (5) Repeat steps (2) and (4) for  $m$  times.

The  $100(1 - \alpha) \%$  confidence intervals are constructed as in (39) with the corresponding empirical quantiles of  $t$  serving as the role of  $z$ .

3 A signed log-likelihood ratio approach: Rewrite the log-likelihood function as a function of  $\delta$ :

$$\begin{aligned}
 l(\delta, \lambda) &= -n_1 \log \sqrt{2\pi} - n_2 \log \sqrt{2\pi} \\
 &\quad - n_2 \log \sigma_1 - n_2 \log \sigma_2 - \frac{1}{2\sigma_1^2} \sum_{j=1}^{n_1} \\
 &\quad \left( y_{1j} - \left( \log \left\{ \delta + \exp \left( \mu_2 + \frac{1}{2} \sigma_2^2 \right) \right\} - \frac{1}{2} \sigma_1^2 \right) \right)^2 \\
 &\quad - \frac{1}{2\sigma_1^2} \sum_{j=1}^{n_2} (y_{2j} - \mu_2)^2,
 \end{aligned} \tag{40}$$

where  $\lambda$  is the vector of nuisance parameters  $(\mu_2, \sigma_1, \sigma_2)$ . The signed log-likelihood ratio statistic (SLLR) is

$$r(\delta) = \text{sgn}(\hat{\delta} - \delta) (2\{l(\hat{\delta}, \hat{\lambda}) - l(\delta, \hat{\lambda}_\delta)\})^{1/2}, \tag{41}$$

where  $\hat{\delta}$  and  $\hat{\lambda}$  denote the maximum likelihood estimators, and  $\hat{\lambda}_\delta$  denotes constrained maximum likelihood estimators: the MLE of nuisance parameters at a given value of  $\delta$ . The distribution of  $r$  approximates the standard normal to the first order. Thus, the CI is given by

$$\{\delta; -z_{\alpha/2} \leq r(\delta) \leq z_{\alpha/2}\}$$

4. A generalized pivotal approach: Generalized pivotal is a statistics that has a distribution free of unknown parameters and an observed value that does not depend on nuisance parameters. In this case, define the generalized pivotal quantities as

$$T_D = \exp(T_1) - \exp(T_2).$$

Notice that this expression depends on two statistics, namely,  $T_1$  and  $T_2$ . They are defined as

$$T_i = \hat{\mu}_i - \frac{\bar{Y}_i - \mu_i}{S_i/\sqrt{n_i}} \hat{\sigma}_i^2 / \sqrt{n_i} + \frac{1}{2} \frac{\sigma_i^2}{S_i^2} \hat{\sigma}_i^2, \quad i = 1, 2. \tag{42}$$

This is equivalent to

$$\begin{aligned}
 T_i &= \hat{\mu}_i - \frac{Z_i}{U_i/\sqrt{n_i-1}} \frac{\hat{\sigma}_i}{\sqrt{n_i}} + \frac{1}{2} \\
 &\quad \times \frac{\hat{\sigma}_i^2}{U_i^2/(n_i-1)}, \quad i = 1, 2,
 \end{aligned} \tag{43}$$

and  $Z_i \sim N(0, 1), U_i^2 \sim \chi_{n_i-1}^2$ . In order to get a CI with GP, some samples can be drawn from  $Z_i$  and  $U_i^2$  and calculate  $T_D$ s. CI can be constructed with enough sample of  $T_D$ s.

The simulation by Chen and Zhou (2006) showed that the generalized confidence intervals yield the best coverage probability, though its performance in small samples is slightly worse. As an alternative, the ratio of two means is also of some interest. The adjusted signed log-likelihood approach is available in construction of confidence intervals, and it is the best choice in that case. For more details, see Chen and Zhou (2006).

### Hypothesis Testing

The hypothesis to be discussed here is a two-sided hypothesis:

$$H : \delta = 0; \quad v.s. \quad K : \delta \neq 0.$$

A one-side test can be derived from two-side tests easily by taking the upper critical value or the lower critical value.

1. A nonparametric bootstrap approach: Zhou et al. (1997) proposed to use bootstrap to get the p-value of the t-statistics. Unlike the bootstrap method used to construct the confidence interval, this time the method does not require parametric assumption. The algorithm is summarized below:

1. Calculate the combined sample mean:

$$\hat{\nu} = \frac{n_1}{n_1 + n_2} \bar{Y}_1 + \frac{n_2}{n_1 + n_2} \bar{Y}_2.$$

2. Transform the samples so that they share a common mean:

$$T_{i,1} = Y_{i,1} - \bar{Y}_1 + \hat{\nu}, \quad T_{i,2} = Y_{i,2} - \bar{Y}_2 + \hat{\nu}.$$

3. Resample  $n_1$  and  $n_2$  observations with equal weights from  $\{T_{i,1}\}$  and  $\{T_{i,2}\}$  with replacement, respectively. Denote the bootstrap samples as  $\{Z_{i,1}\}$  and  $\{Z_{i,2}\}$ .

4. Compute the bootstrap statistics:

$$t^* = \frac{\bar{Z}_1 - \bar{Z}_2}{\sqrt{\frac{\tau_1^{*2}}{n_1} + \frac{\tau_2^{*2}}{n_2}}},$$

where  $\tau_i^{*2}$  is the sample variance of the bootstrap samples.

5. Repeat steps 3 and 4 for  $B$  times, where  $B$  is a large number chosen by the investigator, and denote the series of test statistics as  $\{t_i^*\}_{i=1}^B$ .

6. Calculate the observed test statistics in the same manner of step 4 with original samples.

7. The p-value is

$$p = \frac{\#\{t_i^* : |t_i^*| > |t_{obs}|\}}{B}$$

2. Z-score test: The test statistic is defined as

$$Z = \frac{\bar{W}_1 - \bar{W}_2 + 0.5(S_2^2 - S_1^2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + 0.5\left(\frac{S_1^4}{n_1 - 1} + \frac{S_2^4}{n_2 - 1}\right)}}. \quad (44)$$

The distribution of  $Z$  is approximately standard normal under the null hypothesis. Thus, the p-value is  $\min\{1 - \Phi(Z), \Phi(Z)\}$ . This test and the following tests all require the lognormal assumption.

3. Score tests: Gupta and Li (2006) derived the score test of two lognormal means. Let  $\hat{\lambda}_0 =$

$(\hat{\mu}_{20}, \hat{\sigma}_{10}, \hat{\sigma}_{10})$  be the constrained MLE under the condition that  $\delta = 0$ . The test statistics  $R$  is

$$R = \left[ \frac{\bar{W}_1 - n_2 \hat{\mu}_{20}}{\hat{\sigma}_{10}^2} - \frac{n_1 \hat{\sigma}_{20}^2}{2 \hat{\sigma}_{10}^2} + \frac{n_1}{2} \right]^2 \left[ \frac{\hat{\sigma}_{20}^4}{2n_2} + \frac{\hat{\sigma}_{20}^1}{n_2} + \frac{\hat{\sigma}_{10}^4}{2n_1} + \frac{\hat{\sigma}_{10}^4}{n_1} \right].$$

The score  $R$  follows a  $\chi_1^2$  distribution when the sample sizes, i.e.,  $n_1$  and  $n_2$ , go to infinite.

4. Generalized p-value: The generalized p-value can be achieved with the TDs that were used to construct the generalized confidence interval. Suppose the null hypothesis is  $\theta_1 - \theta_2 = 0$ , then the generalized p-value is

$$p = \min\{p(T_D \leq 0 | \theta_1 - \theta_2 = 0), p(T_D \geq 0 | \theta_1 - \theta_2 = 0)\},$$

which can be estimated by the empirical distribution of  $T_{Ds}$ .

The z-score test is of great computational simplicity and has a straightforward interpretation. The simulation of Zhou et al. (1997) showed that it has a satisfactory performance when the sample sizes of both groups are large. However, Krishnamoorthy and Mathew (2003) discovered that the distribution of z-scores is skewed when the samples are imbalanced between two groups and when the skewness is high compared to the sample sizes. In that case, the generalized p-value would be a better choice for hypothesis testing. Gupta and Li (2006) argued about the same issue, and they showed that the score tests have a better control over type I error and higher power than z-score tests. It is recommended to use score tests or generalized p-value especially when the samples size are not equal. Again, caution should be taken when interpreting the generalized p-value. All three methods discussed above are based on the parametric assumptions; therefore, they are all subjected to huge errors when the lognormal assumption is violated. In that case, a bootstrap test is more preferable since it makes no parametric assumptions. Zhou and Tu (1999) discussed the comparison of multiple population means with zero-inflated distribution.

### Applications on a Simple Example

Callahan et al. (1997) studied the relationship between depression and the expected cost of diagnostic testing for a patient. A subset of patients who had a chronic medical condition as defined by Ambulatory Care Group system is selected out of the entire dataset. The focus of statistical analysis was on the mean of diagnostic testing cost because it can be used to reconstruct the total cost. The data can be summarized in a 124 by 2 matrix, of which the first column records the cost and the second one contains the indicators of depression. Thirteen patients are diagnosed as depression in this sample (depression =1). Four observations out of them have zero costs. The ratio is 17 out of 111 for the non-depression patients (depression =0).

In order to see how these methods perform on this dataset, three questions are raised:

1. What is the mean cost of those non-depression patients who have positive cost, and what is the corresponding confidence interval?
2. What is the mean cost of those non-depression patients, and what is the corresponding confidence interval?
3. What is the difference in mean cost between the depression group and non-depression group among those who have positive cost? And, of course, the confidence intervals and hypothesis testing.

Although they are made up in this example, these questions are commonly seen in real analysis and would help the performance of various methods.

The estimators based on lognormal assumption provide similar answers, while the sample mean is separated from the others. In terms of confidence

intervals, nonparametric methods have similar performances, but there is an obvious difference between nonparametric and parametric. Nonparametric methods tend to be more conservative and robust. The most conservative parametric CI is Angus’s conservative CI, which has a range of 604 larger than those nonparametric methods except for Hall’s transformation. It is noted that Cox’s method yields the most unconservative result. The estimates of lower bound are more alike than estimated for the upper bound, which might be the result of the right skewness. The lower bound by Hall’s transformation is close to those of parametric methods, but its upper bound is much larger than other estimates.

Results of (Tables 4 and 5) are similar to those on positive data. The sample mean is larger than the other two estimators; nonparametric CIs tend to be more conservative than parametric CIs (Tables 1, 2, 3, 6, 7, 8).

For two-sample inference, there are nine observations with positive costs in the depression group, and the standard error of their costs is 1116.3. This might contribute to the extreme estimate of upper bound by generalized pivotal method. Other than that, the results are consistent. Zero is included in all confidence intervals constructed by different methods. This phenomenon is consistent with the results of hypothesis testing, where all four p-values are not significant under common settings.

### Regression

In some sense, linear regressions can be viewed as a generalization of multiple comparison. Consider a simple linear regression with a binary variable as its covariate; the test on the coefficient is the same as a two-sample t-test on mean difference. When the covariate at hand is continuous, i.e., there are

**Table 1** 95% confidence intervals of the one-sample mean (1)

95% Confidence intervals	Parametric methods					
	Cox	Angus	Parametric bootstrap	SLR	Adjusted SLR	Generalized pivotal
Lower bounds	407.9	406.6	419.2	416.3	418.5	420.1
Upper bounds	731.7	1010.2	759.2	750.6	761.2	767.7

infinite categories, a test of the mean relation of the dependent variable and covariate can be achieved by a linear regression. Both ordinary linear regression and generalized linear model describe the mean relation of dependent variable – the outcome and covariates. In other words, it is an “on-average” type of description of the data. The other kind of regression that is going to be discussed in this section is the quantile regression. As will be explained later, quantile regression is slightly different in interpretation from linear regression.

There are extensive econometric literatures on methodologies and applications of regression on medical costs. The features of cost data are the same as those in last section: skewness, nonnegative values, and nontrivial fraction of zero observations. Clustering and multimodality might also affect the validity of results if not properly adjusted.

The most common way to analyze cost data is log transformation. As discussed in the last

section, log-transformed data often has more symmetric distribution than the original data. And the heteroscedasticity found in cost data can sometimes be mediated by variance-stabilizing transformation including log transformation. Thus, linear regressions can be applied on this transformed data. However, the regression on transformed data can only be interpreted as the mean relationship between the transformed outcome and covariates, which is not of scientific interest. It does not cause any trouble when the relation of interest is multiplicative, for instance, the influence of inflation rate on wages. But when the quantity of interest is, say, the total cost, a regression on the transformed data is not enough to answer the question. Therefore, back-transformation becomes a problem. The smearing estimator by Duan (1983) is dominating in this area.

Another way to deal with skewness and non-constant variance is to implement a generalized linear model (GLM). The relation between the dependent variable and covariates is described by two equations in GLM, which are the link function and mean-variance relationship. The flexibility of link function and variance structure provides a wide range of models that can be described under the setting of GLM. Various methods have been proposed to facilitate researchers to choose the best models that fit the data. Manning et al. (2005) discovered that the GLM and log-transformed OLS can be summarized in one family of models named generalized gamma model.

In most study, the methods described above would not be considered complete without the way to deal with the nontrivial fraction of zeros. The zeros cause a direct problem with log transformation, where log (0) has no meaning. A straightforward, also naive, solution is to add a

**Table 2** 95% confidence intervals of the one-sample mean (2)

95% Confidence intervals	NP methods		
	CLT	NP bootstrap	Hall
Lower bounds	343.5	346.2	420.5
Upper bounds	819	816.2	1692.3

**Table 3** Estimates of the one-sample mean

Point estimate	Sample mean	MLE	UMVUE	cm MSE
	581.3	542.7	540	529

**Table 4** Estimates of the zero-inflated mean

Point estimate	Sample mean	Bias-corrected MLE	MVUE
	492.3	462.7	457.7

**Table 5** 95% confidence intervals of the zero-inflated mean

95% Confidence intervals	Parametric methods					NP methods	
	MLE	MVUE	Parametric bootstrap	SLR	Generalized pivotal	CLT	NP bootstrap
Lower bounds	342.1	317.9	330.6	347.8	344.2	287.3	288.9
Upper bounds	625.8	597.4	613.1	640.8	653.5	697	695.6

**Table 6** Estimates of the mean difference between two continuous samples

Point estimate	Sample mean	MLE
	269.1	331.6

**Table 7** 95% confidence intervals of the mean difference between two continuous samples

95% Confidence intervals	MLE	Parametric bootstrap	SLR	Generalized pivotal
Lower bounds	-491.5	-174.9	-442.1	-195.5
Upper bounds	1154.7	2455.3	2568.8	8613.1

**Table 8** P-value of the hypothesis that mean difference is zero

p-Value	Score test	Z-Score test	Bootstrap test	Generalized p-value
	0.85	0.35	0.43	0.15

small constant to zeros. The constant is often chosen to be the minimum positive values in the sample. As one could easily point out, this method has barely any scientific justifications, and its only purpose is to make the model work. Another method is to describe the distribution of cost as a combination of several distributions, which is referred to as the mixture distribution. A common strategy is to describe whether positive cost would be observed in the first part of a two-part model and then use the regression methods discussed before to analyze the cost conditioning on the observations that have medical costs in the second part. The technical problem that arises in a two-part model is the conditioning variance of the estimators, which will be explained in details later in this section.

**Parameters of Interest**

There are several parameters of the cost data that are of practical interests.

1. The conditional mean  $\mu(x) = E[Y|X = x]$ . This is the expected cost of a patient given one’s covariates. It can also be used to make inference about the total cost of one population.

2. The (conditional) marginal effect  $\frac{\partial \mu(x)}{\partial x_k} = \frac{\partial E[y|x]}{\partial x_k}$ . It is a typical measurement of how a certain covariate  $x_k$  affects the dependent variable  $Y$ . In simple regression, it is called “slope.” However, the concept of slope might not be valid in other framework of regressions, and that is why the marginal effect is brought up. Noted that the slope in linear regression does not depend on other covariates, marginal effects are different in the sense that they actually depend on the value of other covariates. Interpretations of marginal effects must not ignore this property.

3. The average marginal effects  $\theta_1 = \frac{1}{n} \sum_{i=1}^n \frac{\partial \mu(x)}{\partial x_k}$  for fixed  $x$  or  $\theta_2 = E\left[\frac{\partial \mu(x)}{\partial x_k}\right]$  for randomized  $x$ .

Marginal effects are conditioned on other covariates. The average marginal effects are created as unconditional values which take the average over possible values of covariates. Therefore, they are features of the entire population instead of any individuals.

As an example, in linear regression, the mean of  $y$  given  $x$  is simply  $x^T\beta$ , and slope of mean w.r.t.  $x_k$  is  $[\beta_k]$  and  $\theta_1 = \theta_2 = \beta_k$ . There are more summarized quantities for the data, but the methods introduced in this section would only focus on these quantities.

**Linear Regression on Raw Data**

Despite the low efficiency, the least squares estimators of linear models are applicable on medical cost data. The estimators of coefficients remain unbiased and consistent, which means it provides results that are acceptable as long as the sample size is large enough. However, cautions should be taken in estimating the variance of the estimated coefficients. It is quite possible that heteroscedasticity exists in cost data. Statistical inference of coefficients would be invalid, without accounting for the heteroscedasticity. Therefore a robust standard error will be more plausible than the homoscedastic standard errors. Huber/White

estimate of the variance-covariance matrix is highly recommended to construct the robust standard errors of coefficients. A typical linear model can be written as

$$Y = X\beta + \varepsilon,$$

where

$$Y = (y_1, y_2, \dots, y_n)^T,$$

and the design matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{pmatrix},$$

and the residuals:

$$\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T.$$

There are certain assumptions in order to make the linear regression valid. This section would only introduce some crucial assumptions but make no further comments. A systematic analysis and descriptions of linear regressions can be found in various textbooks, for instance, Seber and Lee (2012) and Hayashi (2000).

**Assumption 1: Linearity**

$$EY = X\beta.$$

This assumption is sometimes written as

$$y_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i,$$

where  $\varepsilon$  has mean zero. An important concept to be memorized, that is, the linearity, here refers to the linearity in coefficients instead of covariates.

**Assumption 2: Exogeneity**

$$E(\varepsilon_i | X) = 0, \quad (i = 1, 2, \dots, n).$$

The statement means the expectation of residual is zero conditioning on all the covariates. The exogeneity here is actually strict exogeneity.

There is also an assumption called weak exogeneity, which is weaker than this one. When  $X$  is treated as fixed, the exogeneity holds.

**Assumption 3: No Multicollinearity**

$$\text{rank}(X) = p,$$

or, in other words, none of the row vectors of  $X$  can be written as a linear combination of other rows.

**Assumption 4: Uncorrelation**

$$\text{cor}(\varepsilon_i, \varepsilon_j | X) = 0, \quad i \neq j.$$

If this assumption is violated, it is necessary to use estimators of standard errors that adjust for correlations. This is most often observed in spatial or temporal data. A group of observations that has correlations among its members is called a cluster.

**Assumption 5: Constant Variance or Homoscedasticity**

$$\text{var}(\varepsilon_i | X) = E(\varepsilon_i^2 | X) = \sigma^2, \quad (i = 1, \dots, n).$$

The homoscedasticity assumption might be one of the least important assumptions for linear regression – it is too strong to be true in most real-world problems, and also, there are well-developed methods available to estimate the standard errors when it is violated.

**Assumption 6: Normality**

$$\varepsilon_i | X \sim N(0, \sigma_i), \quad (i = 1, \dots, n).$$

This parametric assumption is important in getting the distribution of the test statistics in hypothesis testing. But most of the tests are still valid as long as sample size is large enough even when the errors are not normally distributed.

Linear regressions describe the mean relations between the outcome and covariates. The central limit theorem ensures that a consistent conclusion would be achieved, which means estimates and inferences from an appropriate linear regression would be correct with infinite samples. However, consistency is only one aspect of data analysis.

Researches are often restricted by sample sizes, in which cases efficiency would be of more importance to investigators than consistency. Unfortunately, one major disadvantage of linear regression on raw data is its lack of efficiency. In other words, this regression method needs a greater sample size to reach the same accuracy than some other methods. Recall that skewness and heavy tailedness of the distribution of outcomes ( $Y$ ) are the main features that are responsible for the low efficiency of ordinary regressions. It is natural to think of transformations on  $Y$  to “correct” these features.

### Transformation on $Y$

The intuition of transformation is straightforward: to achieve a better distribution of data by transforming the outcomes with some monotone functions. The advantage is also clear: an appropriate transformation would increase the efficiency of estimation (Manning and Mullahy 2001, Briggs et al. 2005). As been discussed in section “Parameters of Interest,” an obvious issue of transformation is the change of scale. The inference made on transformed scales might not have scientific meanings. Moreover, it is inappropriate to transform estimates directly back to original scale, resulting in biased and inconsistent estimates. Statistical inferences on transformed scale are very likely to be different from those made on original scale. Thus, the main difficulty in the methods based on transformation is the back-transformation problem.

A general procedure can be summarized into three steps: transformation, regression, and back-transformation.

The first step, transformation, consists of choosing a transform function  $h$  and substitute  $y$  with  $h(y)$ . There are various functions that can serve as the transform functions as long as it is monotone and thus invertible. Box-Cox transformation is considered as a well-defined group of transformations for skewed data. Another variance-stabilizing transformation is also available (Weisberg 2005) For analysis of cost data, log transformation is more preferable than others in practice due to certain practical reasons. For instance, regression analysis on the log-transformed scale reveals the proportional

changes of the outcome. It is sometimes the quantity of interest, say, when investigating the association between wages and inflation rate. If then, the problem of back-transformation is avoided. Yet the inference of total mean is often what investigators of medical cost concern about, which requires back-transformation. Another issue is that variance-stabilizing transformations can normalize the distribution of dependent variable, while they may not stabilize the variance as it should do. Therefore, homoscedasticity might not hold for the transformed data.

The next step is to apply the methods discussed in the last section on the transformed data. It is recommended to employ as few assumptions as possible since there is no a priori knowledge of the transformed data. The inference made on transformed scale might be adequate to answer the questions as mentioned above, and then there is no need for the back-transformation step. Otherwise, the analysis should be continued.

The last step, back-transformation, is the key step in this method. Transformation is a tool to gain efficiency, but the questions of interest are still on the original scale of the cost data. The back-transformation methods are dominated by Duan’s smearing estimators. Duan (1983) proposes a nonparametric estimator that uses the average of the transformed residuals to estimate the expectation of dependent variable on the original scale. We estimate  $EY_0$  by substituting the unknown cdf  $F$  by its empirical estimate  $\hat{F}_n$ :

$$EY_0 = \frac{1}{n} \sum_{i=1}^n h(x_0\beta + \hat{\varepsilon}_i). \quad (45)$$

Further substituting the regression parameter  $\beta$  in (45) by its least squares estimates  $\hat{\beta}$ , the smearing estimator is thus defined as

$$EY_0 = \frac{1}{n} \sum_{i=1}^n h(x_0\hat{\beta} + \hat{\varepsilon}_i). \quad (46)$$

Applications and generalizations of Duan’s method have been proposed in recent years. In the rest of this section, three procedures would be introduced as examples for transformation-based methods. The first one is the widely used



logarithm transformation by Ai and Norton (2000); the rest are robust, yet efficient, nonparametric methods by Welsh and Zhou (2006) and Zhou et al. (2008).

**Example: Log Transformation**

Ai and Norton (2000) derived the forms of standard errors of smearing estimators under log transformations by delta method. Their methods allow the situations where a nonlinear regression has been applied in the second step. Results for linear regression can be easily achieved from the general conclusions.

Although normality assumption might not always hold for transformed data, there is no harm to look at the simplified case when the residuals are assumed to be normally distributed. Write the model as  $\ln(y) = k(x, \beta) + s(x, \gamma)\varepsilon$ , where  $k(x, \beta)$  is any models of the expectation of  $\ln(y)$  given  $x$  and  $\varepsilon$  has mean 0 and unit variance. Imposing normality assumption on  $\varepsilon$  means assuming  $\varepsilon$  follows a standard normal distribution. Notice that the square of  $s(x, \gamma)$  is the variance of the error term  $s(x, \gamma)\varepsilon$ , writing it as a function of  $x$  allows for heteroscedasticity. Both  $k(x, \beta)$  and  $s(x, \gamma)$  need to be specified. For linear models,  $k(x, \beta)$  is defined as  $x'\beta$ . Suppose  $\hat{\beta}$  is the estimate of the linear regression, or nonlinear regression, depending on the form of  $k$ , on transformed data,  $\hat{\varepsilon}_i$  is the residual for  $x_i$ ,  $\hat{\Sigma}_\beta$  which is the heteroscedasticity-consistent covariance matrix. An additional regression is needed in order to get the estimates, which is to regress  $\hat{\varepsilon}_i^2$  on  $s(x_i, \gamma)$ . Denote the estimate of  $\gamma$  from the second regression as  $\hat{\gamma}$  and the heteroscedasticity-consistent covariance matrix as  $\hat{\Sigma}_\gamma$ . Then the estimates of  $y$ 's expectation give  $x$  as

$$\hat{\mu}(x) = \exp(k(x, \hat{\beta}) + 0.5s^2(x, \hat{\gamma})),$$

with variance

$$\omega_1(x) = \left( \frac{\partial \mu(x)}{\partial \beta} \Sigma_\beta \frac{\partial \mu(x)}{\partial \beta'} \right) + \left( \frac{\partial \mu(x)}{\partial \gamma} \Sigma_\gamma \frac{\partial \mu(x)}{\partial \gamma'} \right).$$

The incremental effects of  $k^{th}$  elements of  $x$  is

$$\frac{\partial \hat{\mu}(x)}{\partial x^j} = \hat{\mu}(x) \left[ \frac{\partial h(x, \hat{\beta})}{\partial x^j} + 0.5 \frac{s^2(x, \hat{\gamma})}{\partial x^j} \right],$$

with variance

$$\omega_{2j}(x) = \left( \frac{\partial^2 \mu(x)}{\partial x^j \partial \beta} \Sigma_\beta \frac{\partial^2 \mu(x)}{\partial x^j \partial \beta'} \right) + \left( \frac{\partial^2 \mu(x)}{\partial x^j \partial \gamma} \Sigma_\gamma \frac{\partial^2 \mu(x)}{\partial x^j \partial \gamma'} \right).$$

The sample average incremental effect or the marginal effect is

$$\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \hat{\mu}(x_i)}{\partial x^j},$$

with variance

$$\omega_{3j}(x) = \left( \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \mu(x_i)}{\partial x^j \partial \beta} \right) \Sigma_\beta \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \mu(x_i)}{\partial \beta' \partial x^j} \right) \right) + \left( \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \mu(x_i)}{\partial x^j \partial \gamma} \right) \Sigma_\gamma \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \mu(x_i)}{\partial \gamma' \partial x^j} \right) \right).$$

The quantities needed in the above formulas are listed in the appendix.

Now if the normality assumption is inappropriate, estimators are still available and complicated. A new quantity needs to be defined:  $m_i(x, \beta, \gamma) = \exp(k(x, \beta) + [(\ln(y_i) - k(x_i, \beta))/s(x_i, \gamma)]s(x, \gamma))$ , which is the predicted value of  $\mu(x)$  based on  $x_i$ . The intuitive idea is simple: replace the distribution function with empirical distribution – its empirical estimate. The three estimators are listed below. The estimated variances can be found in Appendix B.

$$\hat{\mu}(x) = \frac{1}{n} \sum_{i=1}^n m_i(x, \hat{\beta}, \hat{\gamma}),$$

$$\frac{\partial \hat{\mu}(x)}{\partial x^j} = \frac{1}{n} \sum_{i=1}^n \frac{\partial m_i(x, \hat{\beta}, \hat{\gamma})}{\partial x^j},$$

$$\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^n \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial m_i(x_k, \hat{\beta}, \hat{\gamma})}{\partial x^j} \right).$$

In this example, the transformation is pre-specified and thus known. Transformation

functions would have higher efficiency when the assumptions are more likely to be true. The following example shows how to estimate the transform function from the data.

**Example: Estimating Transformation Function**

It turns out that the transformation function that satisfies  $h(y) = x'\beta + s(x'\gamma)\varepsilon$  is more restricted than it seems to be. Zhou et al. (2008) showed that the transformation function can be estimated from the data once such linear model is specified. Now, the model is defined as  $h(y) = x'\beta + s(x'\gamma)\varepsilon$ . Notice that now it is a linear regression with heteroscedasticity, but the variance is a known function of the scalar  $x'\gamma$ . Denote  $x'\beta$  as  $z_1$  and  $x'\gamma$  as  $z_2$ , the density function of  $z_1, z_2$  as  $p(z_1, z_2)$ , and the cumulative distribution of  $y$  given  $z_1, z_2$  as  $G(y|z_1, z_2)$ , Zhou et al. (2008) derived the relation between  $h$  and  $g_1 = \partial G/\partial z_1$ .

$$h(y) = - \int_{y_0}^y \frac{\sum_{i=1}^n p(u, Z_{1i}, Z_{2i})}{\sum_{i=1}^n g_1(u, Z_{1i}, Z_{2i})} du.$$

They proposed to estimate it with kernel density estimate of unknown density function and distribution function:

$$h_n(y) = - \int_{y_0}^y \frac{\sum_{i=1}^n p_n(u|Z_{1i}, Z_{2i})p_n(Z_{1i}, Z_{2i})}{\sum_{i=1}^n g_{1n}(u|Z_{1i}, Z_{2i})p_n(Z_{1i}, Z_{2i})} du,$$

where

$$p_n(z_1, z_2) = \frac{1}{nb_1b_2} \times \sum_{i=1}^n K_1\left(\frac{Z_{1i} - z_1}{b_1}\right) K_2\left(\frac{Z_{2i} - z_2}{b_2}\right),$$

$$g_{1n}(y|z_1, z_2) = \partial G_n(y|z_1, z_2)/\partial z_1,$$

$$p_n(z_1, z_2) = \frac{1}{nb_0b_1b_2p_n(z_1, z_2)} \sum_{i=1}^n K_0\left(\frac{Y_i - y}{b_0}\right) K_1\left(\frac{Z_{1i} - z_1}{b_1}\right) K_2\left(\frac{Z_{2i} - z_2}{b_2}\right).$$

$K_1, K_2$ , and  $K_3$  are kernel functions with bandwidth 61, 62, and 63, respectively. The unknown parameter can be approached by using estimating equations:

$$\sum_{i=1}^n \frac{[h(y_i) - X'_i\beta]X_i}{s^2(X'_i\gamma)} = 0,$$

$$\sum_{i=1}^n [(h(y_i) - X'_i\beta)^2 - s^2(X'_i\gamma)X_i] = 0,$$

and

$$\beta_n = \sum_{i=1}^n \frac{X_i X'_i}{s^2(X'_i\gamma)}^{-1} \sum_{i=1}^n \frac{X_i h(y_i)}{s^2(X'_i\gamma)}$$

The conditional mean on the original scale ( $n(x)$ ) can be easily estimated by the smearing estimator:

$$\hat{u}(x) = \frac{1}{n} \sum_{i=1}^n \hat{h}^{-1}\left(x'\hat{\beta} + s(x'\hat{\gamma}) \frac{\hat{h}(y_i) - X'_i\hat{\beta}}{s(x'\hat{\gamma})}\right).$$

Zhou et al. (2008) proved that the estimator  $\hat{\mu}(x)$  converges to the true value at the rate  $n^{-1/2}$ , and, as a nonparametric method, it is suitable for any distribution of  $y$ . For more details and the estimate of variance, please see Zhou et al. (2008).

**Example: Nonparametric Retransformation**

Welsh and Zhou (2006) proposed a method that can estimate the back-transformed mean and its standard error for any transformation functions. The model is assumed to be  $h(y) = x'\beta_0 + g_i(\beta_0, \gamma_0)\varepsilon$ , where  $g_i$  can be a function of  $x_i$  and  $\varepsilon_i$  are

independent and identically distributed random variables.  $\psi \sim (\beta^T, \gamma^T)^T$  is estimated from estimating equations. Then denote  $\eta_i = x^T \beta_0 + g(\psi_0)$   $e_i(\psi)$ , where  $e_i(\psi) = \frac{h(y_i) - x_i^T \beta_0}{g_i(\psi_0)}$  and the estimated mean on the original scale is  $\hat{m} = \sum_{i=1}^n h^{-1}(\eta_i(\hat{\psi}))$ ,

which is also a smearing estimator. The idea of this method is to estimate the empirical distribution of residuals  $e_i$  instead of making assumptions. The corresponding standards are estimated with the help of the properties of empirical process. In the original paper, Welsh and Zhou (2006) generalized this method to the situation when there are observations with zero costs.

The idea of transformation method is to transform the data so that it has a “better” distribution, which is often more symmetric and less heavy tailed and can be better fitted with a linear model. By doing this, one can gain efficiency from transformation and assumptions. A natural alternative is to abandon the requirement about symmetry. For instance, a log-transformed linear model can be interpreted as a lognormal model as well. In the next section, this kind of models – generalized linear model – and applications of them will be discussed.

## Transformation on E[Y]

Linear model can be viewed as a parametric model based on the normality assumption, where the mean of normal distribution is assumed to have a linear relationship with the coefficients. If the model is correct, the dependent variable is normally distributed – therefore symmetric and without heavy tail. A natural generalization of this traditional linear model is to expand the family of distributions to account for possible skewness and heavy tail, which is called the generalized linear model by McCullagh and Nelder (1989). The GLM is first introduced to the area of medical cost analysis by Blough et al. (1999).

Let  $\mu$  be  $E(Y)$ , where  $Y$  is a  $n \times 1$  vector.  $Y_i$ ,  $i = 1, 2, \dots, n$  are i.i.d. from a common distribution

$f$ . Assumed that such relationship exists:  $g(\mu) = X\beta$  with  $g$  being a monotone increasing function and the variance-covariance matrix of  $Y$  is a function of  $\mu : V(\mu)$ , which is determined by the density function  $f$ . The function  $g$  is usually called as the link function, and  $\text{var}(Y) = V(\mu)$  is called the mean-variance relationship or variance function. The unknown parameter can be estimated by maximum likelihood estimator since a parametric form of  $f$  is available. For short, a GLM describes the relation between a function of the expectation of  $Y$  and covariates; variation is addressed by the mean-variance relationship and/or the assumed distribution.

One important advantage of GLM is that it can handle various types of data. For instance, discrete data can be described by the Poisson distribution with a log link function. For binary data, it can be analyzed by a Bernoulli distribution with logit link, which is known as a logistic regression.

Recall that in linear model, the normality assumption is the least important assumption because of central limit theorems. The same thing happens here. The parametric assumption is not necessarily required in setting a GLM model, although it is still popular because of its direct interpretations. In the previous setup, one needs to specify the actual distribution of the dependent variable and then use it to derive the score function. But, in fact, one only needs the mean-variance relationship and use it to construct an estimating equation which has the same properties as the score function. The estimators from corresponding estimating equations are still consistent. Therefore, the procedure reduced to specify (1) the link function and (2) the mean-variance relationship. Notice that the first term is the first moment of dependent variable and the second term is about the second moment. That is why econometricians also call GLM and generalized moment methods.

With parametric assumptions, the MLE might have explicit solutions. Otherwise, the estimators can be solved by solving the following estimating equations with numerical method:

$$\sum_{i=1}^N \frac{\partial \mu(x_i; \beta)}{\partial \beta} V^{-1}(\mu(x_i; \beta))(y_i - \mu(x_i; \beta)) = 0, \tag{47}$$

where  $\mu(x_i; \beta) = g^{-1}(x_i' \beta)$ . If the model is specified correctly, the asymptotic variance of the estimator will be the inverse of Fisher information up to some constant. Or, one can use the sandwich estimator as a robust estimator. A commonly used test for the coefficients is the Wald test.

The interpretations of the regression must be taken care of. A GLM describes the relationship between covariates and a function of  $Y$ 's expectation. Logistic regression, for example, shows the linear relationship between the covariates and the odds ratio. In medical cost data, the situation is simpler since the most widely used GLM model in analyzing cost data is a gamma distribution with a log link. Or without the parametric assumption, one can employ a log link and  $V(y) = \phi \mu^2$ , which is a feature often observed in most medical cost data (Blough et al. 1999, Manning and Mullahy 2001).

**Flexible Link Function**

Basu and Rathouz proposed a method that enables investigators to choose the link function and variance function from a certain family and thus provide an option when there is no a priori knowledge of the link function and the mean-variance relationship. They define a parametric family of link function indexed by  $\lambda$ :

$$h(y, \lambda) = \begin{cases} (\mu_i^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(\mu_i), & \text{if } \lambda = 0 \end{cases}$$

This family of functions is a modification of Box-Cox transformation (1). Similarly, the authors define two families  $h(\mu, \theta_1, \theta_2)$ : PV and QV.

The power variance family:

$$g(\mu_i; \theta_1, \theta_2) = \theta_1 \mu_i^{\theta_2};$$

Quadratic variance family:

$$g(\mu_i; \theta_1, \theta_2) = \theta_1 \mu_i + \theta_2 \mu_i^2.$$

Denote the parameters as  $\gamma = (\beta^T, \lambda, \theta_1, \theta_2)^T$ . Then the estimating equations are

$$G_{\beta_j}^i = (Y_i - \mu_i) V_i^{-1} (\partial \mu_i / \partial \beta_j);$$

$$G_{\lambda}^i = (Y_i - \mu_i) V_i^{-1} (\partial \mu_i / \partial \lambda);$$

$$G_{\theta_1}^i = [(Y_i - \mu_i)^2 - V_i] V_i^{-2} (\partial \mu_i / \partial \theta_1);$$

$$G_{\theta_2}^i = [(Y_i - \mu_i)^2 - V_i] V_i^{-2} (\partial \mu_i / \partial \theta_2).$$

And they can be combined in a vector form. Let

$$G_{\gamma}^i = (G_{\beta_1}^i, G_{\beta_2}^i, \dots, G_{\beta_p}^i, G_{\lambda}^i, G_{\theta_1}^i, G_{\theta_2}^i)^T.$$

The estimating equation is then

$$\sum_{i=1}^n G_{\gamma}^i = 0.$$

The additional indexes  $\lambda$  and  $\theta_1, \theta_2$  can be incorporated into the generalized estimating equations: the variance can be estimated by the sandwich estimator. The marginal effect of  $x^j$  is

$$\frac{\partial \hat{\mu}(x)}{\partial x^j} = \frac{1}{n} \sum_{i=1}^n \hat{\beta}_j [\mu(X_i^T \hat{\beta} \lambda)]^{1-\lambda}.$$

The authors showed that estimating link function results in some loss of efficiency, but it is partially recovered by estimation of the variance structure. They also recommended the use of power variance family when treating continuous outcomes and quadratic power family for discrete family.

The next example gives a general distribution that is able to cover many other distributions.

**A Generalized Gamma Model**

Manning et al. (2005) proposed a generalized gamma model(GGM) to analyze skewed and

heavy-tailed data. One key feature of this GGM is that it provides more flexibility than some other models: lognormal models, gamma models, and Weibull models are all special cases of GGM. The density function of GGM is

$$f(y; k, \mu, \sigma) = \frac{\gamma^\gamma}{\sigma y \sqrt{\gamma} \Gamma(\gamma)} \exp[z\sqrt{\gamma} - \mu], \quad (48)$$

where  $\gamma = |\kappa|^{-2}$ ,  $z = \text{sign}(k) \log(y) - \mu/\sigma$  and  $\mu = \gamma \exp(|\kappa| z)$ . The parameter  $\mu$  is replaced by  $X_i^T \beta$ . The expected value of  $y$  condition on  $x$  is given by

$$E(y|x) = \exp\left[x^T \beta + \left(\frac{\sigma}{\kappa}\right) \log(\kappa^2) + \log\left(\Gamma\left(\frac{1}{\kappa^2} + \frac{\sigma}{\kappa}\right) - \log\left(\Gamma\left(\frac{1}{\kappa^2}\right)\right)\right],$$

As shown in Manning et al. (2005), (48) is a lognormal model when  $\kappa$  is close to zero or a gamma distribution when  $\sigma = \mathcal{K} > 0$ . In other words, the value of the parameters of GGM can distinguish those special models from each other. A natural benefit of this kind of setting is that a model selection problem can be restated as a hypothesis testing on the parameters. Or in another aspect, it provides a systematic way to evaluate the appropriateness of those models.

In their paper, Manning et al. (2005) compared three versions of GMMs – featured by the way to deal with heteroscedasticity – against some existing model including back-transformed linear regression of  $\ln(y)$  on  $x$ , a GLM with log link and gamma distribution, and a maximum likelihood estimator of Weibull model. Results showed that the GGM would choose the right model properly, yet the heteroscedasticity in  $x$  has to be accounted for. Also, GGM can better approximate the distribution of the data than other parametric models due to its flexibility.

### Two-Part Models

The models discussed above are all based on positive and continuous data. But in real-life research, there is always a considerable fraction of observations that have zero cost. One can choose to

keep or drop all zero-cost observations depending on the research interest. The most commonly used modification is to construct a two-part model.

The intuition behind the two-part model is to describe separately the event that cost occurs and how much the cost is when it occurs. The outcome variable in the first part is a binary variable  $\delta_i$ , where 0 stands for no cost occurs and 1 stands for positive cost. Most of methods that are available for binary outcomes are applicable here, and logistic regression is a typical method that one would use. In the second part, all observations left have positive costs and that turns the problem to what have been talked about.

It seems a little bit complicated, but by the short argument below, it will be clear why a two-part model will simplify the problem. Suppose there is a parametric distribution for the second part. The likelihood function:

$$\begin{aligned} L_n &= \prod_{i=1}^n p(\delta_i|x_i) f(y_i|\delta_i = 1, x_i)^{\delta_i} \\ &= \prod_{i=1}^n p(\delta_i|x_i) \prod_{\delta_i=1} f(y_i|\delta_i = 1, x_i) \end{aligned} \quad (49)$$

If the conditional density function  $f$  does not depend on  $\delta$ , then the likelihood function can be maximized separately. Recall that all models in previous sections have nothing to do with  $\delta$ ; they can serve as the conditional density function here. Therefore, all one needs to do is to analyze the first part and the second part separately and then combine the result into one. The estimated mean of the population will be

$$\hat{y} = \hat{p} \times \hat{\mu},$$

where  $\hat{\mu}$  is the estimated mean of the cost in the second part and  $\hat{p}$  is the probability that cost occurs. Blough et al. (1999) estimated the variance of  $\hat{y}$  by

$$\begin{aligned} \text{Var}(\hat{y}) &= \text{Var}(\hat{p}\hat{\mu}) \\ &= \hat{p}^2 \text{Var}(\hat{\mu}) + \hat{\mu}^2 \text{Var}(\hat{p}), \end{aligned} \quad (50)$$

which is an approximation of the true variance. An alternative to use this equation is to generate

the variance by bootstrap methods. The parameters used in the first part are not necessary to be the same as those in the second part. Interpretations of the coefficients are different from the previous section since the inference on the second part is conditioning on the event that cost occurs.

As mentioned early, two-part models are quite popular in the analysis of medical cost. An example can be found in Blough et al. (1999) where they used a logistic regression for the first part and a GLM with log link for the second part. If one chooses to transform  $y_i$ , the back-transformation problem for a two-part model had been studied by Welsh and Zhou (2006).

### Mixtures of Distributions

With a point mass at zero, observations gathering around zero can also be viewed as multimodality, which can be explained by that the distribution is actually a mixture of several distributions. In fact, the two-part model is a special case of mixture models. A mixture model is helpful in classifying the observations into high-cost groups and low-cost groups. Say the true distribution of medical cost in a certain population is a mixture of several normal distributions with different means due to some unknown features of patients. Then the unknown features can be treated as a latent variable that would help in telling which normal curve the patient is in. Expectation maximization (EM) algorithm would give the estimates of the coefficients of interests. More details about mixture models can be found in McLachlan and Peel (2000).

There are other methods to deal with zero-cost observations, which include adding a constant to each sample and forced the data to be positive. However, some methods have hardly any realistic meaning but only serve as a way to address the zero-cost observations. An advantage of the two-part model is that it makes some sense in terms of real-life interpretations.

### Quantile Regression

All of the methods introduced above focus on the relation between covariates and the mean of

outcomes. The mean is one quantity that can summarize the property of the conditional distribution of outcome variables. Of course there are more summary quantities, for instance, the median, 25%, and quantile, 75%, all of which can present the distribution in some sense. It is noted that the quantiles are better estimators than the mean for skewed or heavy-tailed data. However, the quantity of interest in this analysis is the total medical costs, which is directly related with mean but not a single quantile. In order to estimate the total medical cost or the mean, a series of quantiles should be estimated so that an empirical estimate of distribution can be achieved. The regression of quantiles is called the quantile regression. Koenker and Hallock (2001) said that “Quantile regression seeks to extend these ideas to the estimation of conditional quantile functions – models in which quantiles of the conditional distribution of the response variable are expressed as functions of observed covariates.”

The quantile regression can be viewed as a generalized median regression. In a median regression, the output of regression would describe the relation between the median, or 50% quantile, and the covariates. Median regression seeks to minimize the difference between the estimated values and the real values, in contrast to mean regression. Or in other words, median regression estimators minimize the sum of absolute value of the difference:

$$\min \sum_{i=1}^n |y_i - X'_i \beta|. \quad (51)$$

Now let  $\tau$  ranges from 0 to 100%, a regression on the  $\tau$  th quantile is

$$\min \sum_{i=1}^n \rho_\tau(y_i - X'_i \beta), \quad (52)$$

where  $\rho_\tau$  is defined as  $\rho_\tau(u) = u(\tau - I(u < 0))$ . When  $\tau$  is set to be 0.5,  $\rho$  is equivalent to taking the absolute value up to a scalar. And thus (52) is exactly the same as (51). This optimization can be easily solved by many algorithms. The

implementation of quantile regression can be achieved through written software such as *quantreg* in R by Koenker (2009). A nature estimator of the conditional mean is the average of all conditional quantiles. The marginal effects are now specified with respect to each quantile.

A detailed example of quantile regression analysis can be found in Koenker and Hallock (2001). In general, investigators can set a series of  $r$ , say from 10% to 90% increased by 10%. That would give ten regression results, each of which stands for the relation of covariates and the corresponding quantile. A major advantage of quantile regression over linear regression is that it reveals the different behavior of covariates on outcomes. Regression on the mean averages this effects and report only the averaged value. It is very possible that a feature would behave differently on subjects with relatively low costs and those with relatively high costs as shown in the example in Koenker and Hallock (2001). A possible explanation is that there are unknown features even after some features are controlled; those features would affect the costs and have interaction with the controlled features. This concept is similar to mixture models, where there are unknown features that define different models. But quantile regression does not attempt to figure out the classifier; it simply performs the regression on different quantiles.

---

## Prediction

As have been studied in previous sections, various methods and models can be employed to discover and quantify the association between covariates and medical cost in the target population. Naturally, one would be interested in whether it is possible to predict the future medical cost for an individual, or a group of people, given certain information. It is worthy of noting that prediction is a very broad subject where methods arise from various disciplines, which is beyond the scope of this chapter. In this section, a brief overview of prediction methods and some important concepts

about prediction are introduced, leaving the details to be explored by readers.

## Some Basic Concepts of Prediction Models

The primary question of interest is how to accurately predict the response, in this case the medical cost, given other individual information (predictors) and previous knowledge (the observed sample and maybe the theoretical model). A secondary question is how to estimate the accuracy, i.e., the prediction error, of the proposed method. This type of prediction is called a supervised learning in the sense that there is a response (or outcome) that can be used to judge how well the method does. Usually it is achieved by specifying a loss function which penalizes the method based on the deviation from the true response, e.g., the square error and absolute error. The regression methods described in the last section can be counted as methods of supervised learning, where most of them use square error loss and quantile regression uses several versions of absolute error loss. Notice that an additional assumption is needed in order to make the prediction valid: the sample been predicted should be from the same population from where the observed sample is drawn. There are a bunch of other methods available, to name some, principle component analysis (PCA), support vector machine (SVM), neural networks, random forest, and so on. A general and broad introduction of the methods can be found in Friedman et al. (2001).

As for the measure of accuracy of prediction, several measures are available, for example, root mean squared error (RMSE) and mean squared prediction error (MSPE). RMSE is defined as the squared root of mean squared error, which can be estimated by the mean of squared difference between the fitted values and true values. MSPE is the mean of squared difference between the predicted values and the true values. The difference between MSPE and MSE is that the model used to generate predicted values is fitted by another dataset, while the MSE is calculated with the model fitted by the same dataset. In

other words, it requires two independent datasets to estimate MSPE but only one for MSE. The dataset used to fit the model is called the training set, and the other one is called test set. MSE is almost always smaller than MSPE. Theoretically, MSPE is a better measure of accuracy than RMSE. However, estimating MSPE requires two independent datasets, which might be a luxury for study with small sample size. Meanwhile, depending on MSE, it might result in overfitting the current dataset, and thus the model is not valid for generalization onto other datasets.

Recall that the purpose of prediction is to predict the response with the highest accuracy, so the next question is how to choose the best out of all these models, which is called model selection in literatures. The basic idea is to estimate the measures, each model achieved on the study dataset, and choose the one with the best performance. One question that researchers often encounter is how to decide what predictors and how many of them should be included in the model. Say the measure of accuracy is MSPE. Ideally, there should be a sufficiently large training set to fit the model and a test set that could give a good estimate of MSPE. However, this might not be the case in real-world study. There are different approaches to overcome the limitation of sample size and generate an acceptable estimate of MSPE, like pseudo out-of-sample forecast and cross validation. Take the cross validation, for example, a  $k$ -fold cross validation will randomly divide the sample into  $k$  subsamples. One subsample will be kept as the test set and the other  $(k-1)$  subsamples are used to fit the model. One can take the average of the  $k$ -fitted models as the single fitted model. The average of the  $k$  MSPE is then used as a quantity that summarizes how this model performs and also an estimate for the MSPE. The model that has the lowest average MSPE will then be chosen. A common mistake in doing cross validation is to somehow use the whole dataset in fitting the model, for instance, using the whole dataset to choose predictors and then fitted the model using these predictors by “ $k$ -fold cross validation.” The MSPE calculated in that manner would be smaller than the true value, and it cannot be served as an estimate of the true MSPE. Also,

since the way it is generated is similar to that of MSE, it might also result in overfitting when using it as a measure to choose the best model.

Even with cross validation, overfitting is still a problem. Throwing more predictors into the model will result in smaller MSPE in most cases. The small MSPE presents as a problem since it is possible that the fitted model has been modified to describe and only describe this observed sample, or training set, and thus the model is limited in being generalized to other samples in the population. Therefore, it is a trade-off between the ability of generalization and the accuracy.

### **Difference from Regression Analysis**

At the first sight, prediction and statistical inference are similar to each other in the context of regressions: there is an observed sample, with several predictors (or covariates) and an outcome variable; one builds a model to describe the association between predictors and outcomes so that the mean, quantiles, or the distribution of the outcome can be explained by a function of predictors. However, the focus of these two analyses is different. For statistical inference, the target is to describe the relationship between the covariates and outcomes in the population from which the sample is drawn. For prediction, the major interest lies in the accuracy of the predicted value, regardless of whether the model makes sense or not. For instance, it is okay to look at the fitted model and say certain predictors' prediction ability is high, but one should not overinterpret relationships discovered in a prediction model. And also, additional assumptions are needed if the regression model is used for prediction. The most important assumption is that new sample should be from the same population where the model is fitted, so that it is legit to use the model fitted on the observed sample to make prediction. Another thing is that the conditional expectation of response give predictors has different interpretations under different setting. In regression analysis, it is the average response for those who have the given levels of predictors, the uncertainty of which is estimated by the standard error. For prediction model, it is



the predicted expectation of response given the level of predictors, the uncertainty of which is estimated by MSPE. Generally speaking, the prediction error is larger than the standard error.

## Appendix

### Concept of General Pivots

The concept of generalized pivot is first introduced by Tsui and Weerahandi (1989). Weerahandi (1993) compared the properties of frequentist confidence intervals and generalized confidence intervals to give an intuitive understanding:

Property 1: Consider a particular situation of interval estimation of a parameter  $\theta$ . If the same experiment is repeated a large number of times (depending on the required accuracy of the desired coverage) to obtain new sets of observations  $x$ , then the confidence intervals by conventional definition will correctly include the true value of the parameter 95% of the time.

Property 2: After a large number of independent situations of setting 95% confidence intervals for certain parameters of interest, the investigator will have correctly included the true values of the parameters in the corresponding intervals 95% of the time.

Property 1 is the property of classic frequentist confidence intervals and it implies Property 2. However, it is not always possible to find the confidence intervals that satisfy Property 1, a well-known example of which is the Behrens-Fisher problem. Weerahandi (1993) argued that Property 2 is of direct practical importance because the statistical inference is no longer an issue if indeed repeated samples can be obtained from the same experiment. The confidence intervals that have Property 2 are thus called generalized confidence intervals.

In order to construct a confidence interval, a quantity call pivotal quantity is essential. The discussion of generalized confidence intervals is

actually a discussion of the generalized pivotal quantity. Hannig et al. (2006) refined the definition given by Weerahandi (1993) and discovered that a subclass of generalized pivotal quantity is of interests and good properties. This subclass of generalized pivotal quantity is named the fiducial generalized pivotal quantity due to its close connection with Fisher (1935) fiducial argument.

**Definition 1** A function of  $(\mathbb{S}, \mathbb{S}^*, \xi)$  for a parameter  $\theta$ , denoted as  $P_\theta(\mathbb{S}, \mathbb{S}^*, \xi)$ , is called a fiducial generalized pivotal quantity (FGPQ) if it satisfies the following two conditions (FGPQ1). The conditional distribution of  $P_\theta(\mathbb{S}, \mathbb{S}^*, \xi)$ , conditional on  $\mathbb{S} = s$ , is free of  $\xi$  (FGPQ2). For every allowable  $s \in \mathbb{R}^k, P_\theta(s, s^*, \xi) = 0$ .

Hannig et al. (2006) proved that, under mild conditions, the coverage probability of a generalized confidence interval is correct as sample size goes to infinite. The authors also gave a structural method to construct the fiducial generalized pivotal quantity. It is briefly described here:

**Definition 2** Let  $\mathbb{S} = (S_1, \dots, S_k) \in \mathbb{S} \subset \mathbb{R}^k$  be a  $k$ -dimensional statistic whose distribution depends on a  $p$ -dimensional parameter  $\xi \in \Xi$ . Suppose there exist mappings  $f_1, \dots, f_k$ , with  $f_j: \mathbb{R}^k \times \mathbb{R}^p \rightarrow \mathbb{R}$ , such that, if  $E_i = f_i(\mathbb{S}; \xi)$ , for  $i = 1, \dots, k$ ; then  $\mathbb{E} = (E_1, \dots, E_k)$  has a joint distribution that is free of  $\xi$ . We say that  $f(\mathbb{S}, \xi)$  is a pivotal quantity for  $\xi$  where  $f = (f_1, \dots, f_k)$ .

**Definition 3** Let  $f(\mathbb{S}, \xi)$  be a pivotal quantity for  $\xi$  as described in Definition 2. For each  $s \in \mathbb{S}$ , define  $\varepsilon(s) = f(s, \Xi)$ . Suppose the mapping  $f(s, \cdot): \Xi \rightarrow \varepsilon(s)$  is invertible for every  $s \in \mathbb{S}$ . We then say that  $f(\mathbb{S}, \xi)$  is an invertible pivotal quantity for  $\xi$ . In this case we write  $g(s, \cdot) = (g_1(s, \cdot), \dots, g_k(s, \cdot))$  for the inverse mapping so that whenever  $e = f(s, \xi)$ , we have  $g(s, e) = \xi$ .

**Theorem 1** Let  $\mathbb{S} = (S_1, \dots, S_k) \in \mathbb{S} \subset \mathbb{R}^k$  be a  $k$ -dimensional statistic whose distribution depends on a  $p$ -dimensional parameter  $\xi \in \Xi$

Suppose there exist mappings  $f_1, \dots, f_k$ , with  $f_j : \mathbb{R}^k \times \mathbb{R}^p \rightarrow \mathbb{R}$ , such that  $f = (f_1, \dots, f_k)$  is an invertible pivotal quantity with inverse mapping  $g(s, \cdot)$ . Define

$$\begin{aligned} \mathcal{R}_\theta &= \mathcal{R}_\theta(\mathbb{S}, \mathbb{S}^*, \xi) \\ &= \pi(g_1(\mathbb{S}, f(\mathbb{S}^*, \xi)), \dots, g_k(\mathbb{S}, f(\mathbb{S}^*, \xi))) \\ &= \pi(g_1(\mathbb{S}, \mathbb{E}^*), \dots, g_k(\mathbb{S}, \mathbb{E}^*)) \end{aligned} \tag{53}$$

where  $\mathbb{E}^* = f(\mathbb{S}^*, \xi)$  is an independent copy of  $\mathbb{E}$ . Then  $P_\theta$  is a FGPO for  $\theta = \pi(\xi)$ . When  $\theta$  is a scalar parameter, an equal-tailed two-sided  $(1 - \alpha)$  100% GCI for  $\theta$  is given by  $P_{\theta, \alpha/2} \leq \theta \leq P_{\theta, 1-\alpha/2}$ . Here  $P_{\theta, \gamma} = P_{\theta, \gamma}(s)$  denotes the  $100\gamma^{\text{th}}$  percentile of the distribution of  $P_\theta$  conditional on  $\mathbb{S} = s$ . One-sided generalized confidence bounds are obtained in an obvious manner.

This method is only valid in problems where complete statistics exist. For the incomplete cases, the authors gave two generalizations of this method. For more details, please see Hannig et al. (2006).

**Variances and Estimators for Back-Transformations**

$$\begin{aligned} \frac{\partial \hat{\mu}(x)}{\partial \beta} &= \hat{\mu}(x) \left[ \frac{\partial h(x, \hat{\beta})}{\partial \beta} \right], \\ \frac{\partial \hat{\mu}(x)}{\partial \gamma} &= \hat{\mu}(x) \left[ 0.5 \frac{\partial s^2(x, \hat{\gamma})}{\partial \gamma} \right], \\ \frac{\partial \hat{\mu}(x)}{\partial x^i \partial \beta} &= \frac{\partial \hat{\mu}(x)}{\partial \beta} \left[ \frac{\partial h(x, \hat{\beta})}{\partial x^i} + 0.5 \frac{\partial s^2(x, \hat{\gamma})}{\partial x^i} \right] \\ &\quad + \hat{\mu}(x) \left[ \frac{\partial^2 h(x, \hat{\beta})}{\partial x^i \partial \beta} \right], \\ \frac{\partial \hat{\mu}(x)}{\partial x^i \partial \gamma} &= \frac{\partial \hat{\mu}(x)}{\partial \gamma} \left[ \frac{\partial h(x, \hat{\beta})}{\partial x^i} + 0.5 \frac{\partial s^2(x, \hat{\gamma})}{\partial x^i} \right] \\ &\quad + 0.5 \hat{\mu}(x) \left[ \frac{\partial^2 s^2(x, \hat{\gamma})}{\partial x^i \partial \gamma} \right]. \end{aligned}$$

The variances derived from delta methods are

$$\begin{aligned} \omega_1(x) &= \left( \frac{\partial \mu(x)}{\partial \beta} \sum_\beta \frac{\partial \mu(x)}{\partial \beta'} \right) + \left( \frac{\partial \mu(x)}{\partial \gamma} \sum_\gamma \frac{\partial \mu(x)}{\partial \gamma'} \right) \\ &\quad + 2 \left( \frac{\partial \mu(x)}{\partial \beta} \sum_\beta \frac{\partial \mu(x)}{\partial \gamma} \right) + 2 \frac{\partial \mu(x)}{\partial \beta} \sum_{2D\beta} \\ &\quad + 2 \frac{\partial \mu(x)}{\partial \gamma} \sum_{1D\gamma} + \sum_{1DD}, \\ \omega_{2j}(x) &= \left( \frac{\partial^2 \mu(x)}{\partial x^i \partial \beta} \sum_\beta \frac{\partial^2 \mu(x)}{\partial \beta' \partial x^j} \right) + \left( \frac{\partial^2 \mu(x)}{\partial x^i \partial \gamma} \sum_\gamma \frac{\partial^2 \mu(x)}{\partial \gamma' \partial x^j} \right) \\ &\quad + 2 \left( \frac{\partial^2 \mu(x)}{\partial x^i \partial \beta} \sum_{\beta, \gamma} \frac{\partial^2 \mu(x)}{\partial \gamma' \partial x^j} \right) + 2 \frac{\partial^2 \mu(x)}{\partial x^i \partial \beta} \sum_{2D\beta} \\ &\quad + 2 \frac{\partial^2 \mu(x)}{\partial x^i \partial \gamma} \sum_{2D\gamma} + \sum_{2DD}, \\ \omega_{3j}(x) &= \left( \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \mu(x_i)}{\partial x^i \partial \beta} \right) \Sigma_\beta \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \mu(x_i)}{\partial \beta' \partial x^j} \right) \right) \\ &\quad + \left( \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \mu(x_i)}{\partial x^i \partial \gamma} \right) \Sigma_\gamma \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \mu(x_i)}{\partial \gamma' \partial x^j} \right) \right) \\ &\quad + 2 \left( \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \mu(x_i)}{\partial x^i \partial \beta} \right) \Sigma_{\beta\gamma} \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \mu(x_i)}{\partial \gamma' \partial x^j} \right) \right) \\ &\quad + 2 \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \mu(x_i)}{\partial x^i \partial \beta} \Sigma_{3D\beta} \right) \\ &\quad + 2 \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \mu(x_i)}{\partial x^i \partial \gamma} \Sigma_{3D\gamma} \right) + \Sigma_{3DD}. \end{aligned} \tag{9}$$

The derivatives:

$$\begin{aligned} \frac{\partial \hat{\mu}(x)}{\partial \beta} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial m_i(x, \hat{\beta}, \hat{\gamma})}{\partial \beta}, \\ \frac{\partial \hat{\mu}(x)}{\partial \gamma} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial m_i(x, \hat{\beta}, \hat{\gamma})}{\partial \gamma}, \\ \frac{\partial^2 \hat{\mu}(x)}{\partial x^i \partial \beta} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial m_i^2(x, \hat{\beta}, \hat{\gamma})}{\partial x^i \partial \beta}, \\ \frac{\partial^2 \hat{\mu}(x)}{\partial x^i \partial \gamma} &= \frac{1}{n} \sum_{i=1}^n \frac{\partial m_i^2(x, \hat{\beta}, \hat{\gamma})}{\partial x^i \partial \gamma}, \\ \hat{\Sigma}_{\beta\gamma} &= \left( \sum_{i=1}^n \frac{\partial k(x_i, \hat{\beta})}{\partial \beta} \frac{\partial k(x_i, \hat{\beta})}{\partial \beta'} \right)^{-1} \\ &\quad \left( \frac{\partial k(x_i, \hat{\beta})}{\partial \beta} \hat{\varepsilon}_i \hat{\eta}_i \frac{\partial s^2(x_i, \hat{\gamma})}{\partial \gamma} \right) \left( \sum_{i=1}^n \frac{\partial s^2(x_i, \hat{\gamma})}{\partial \gamma} \frac{\partial s(x_i, \hat{\gamma})}{\partial \gamma'} \right)^{-1} \end{aligned}$$

Estimated coefficient	Dependent variable	Independent variable
$\hat{\Sigma}_{1D\beta}$	$m_i(x_i, \hat{\beta}, \hat{\gamma}) \times \hat{\epsilon}_i$	$\partial k(x_i, \hat{\beta}) / \partial \beta$
$\hat{\Sigma}_{1D\gamma}$	$m_i(x_i, \hat{\beta}, \hat{\gamma}) \times \hat{\eta}_i$	$\partial s^2(x_i, \hat{\gamma}) / \partial \gamma$
$\hat{\Sigma}_{2D\beta}$	$(\partial m_i(x_i, \hat{\beta}, \hat{\gamma}) / \partial x^j) \times \hat{\epsilon}_i$	$\partial k(x_i, \hat{\beta}) / \partial \beta$
$\hat{\Sigma}_{2D\gamma}$	$(\partial m_i(x_i, \hat{\beta}, \hat{\gamma}) / \partial x^j) \times \hat{\eta}_i$	$\partial s^2(x_i, \hat{\gamma}) / \partial \gamma$
$\hat{\Sigma}_{3D\beta}$	$(n^{-1} \sum_i \partial m_i(x_i, \hat{\beta}, \hat{\gamma}) / \partial x^j) \times \hat{\epsilon}_i$	$\partial k(x_i, \hat{\beta}) / \partial \beta$
$\hat{\Sigma}_{3D\gamma}$	$(n^{-1} \sum_i \partial m_i(x_i, \hat{\beta}, \hat{\gamma}) / \partial x^j) \times \hat{\eta}_i$	$\partial s^2(x_i, \hat{\gamma}) / \partial \gamma$
Estimated variance		Sample variance of
$\hat{\gamma}_{DD}$	$m_i(x_i, \hat{\beta}, \hat{\gamma})$	
$\hat{\gamma}_{DD}$	$\partial m_i(x_i, \hat{\beta}, \hat{\gamma}) / \partial x^j$	
$\hat{\gamma}_{DD}$	$n^{-i} \sum_i \partial m_i(x_i, \hat{\beta}, \hat{\gamma}) / \partial x^j$	

**References**

Ai C, Norton EC. Standard errors for the retransformation problem with heteroscedasticity. *J Health Econ.* 2000;19(5):697–718.

Aitchison J. On the distribution of a positive random variable having a discrete probability mass at the origin. *J Am Stat Assoc.* 1955;50(271):901–8.

Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. *J Health Econ.* 1999;18(2):153–71.

Box GEP. Science and statistics. *J Am Stat Assoc.* 1976;71(356):791–9.

Briggs A, Nixon R, Dixon S, Thompson S. Parametric modelling of cost data: some simulation evidence. *Health Econ.* 2005;14(4):421–8.

Callahan CM, Kesterson JG, Tierney WM, et al. Association of symptoms of depression with diagnostic test charges among older adults. *Ann Intern Med.* 1997;126(6):426.

Yea-Hung Chen and Xiao-Hua Zhou. Interval estimates for the ratio and difference of two lognormal means. *Stat Med.* 25(23):4099–4113, 2006. ISSN 1097-0258. <https://doi.org/10.1002/sim.2504>.

Dominici F, Cope L, Naiman DQ, Zeger SL. Smooth quantile ratio estimation. *Biometrika.* 2005;92(3):543–57.

Duan N. Smearing estimate: a nonparametric retransformation method. *J Am Stat Assoc.* 1983;78(383):605–10. ISSN 01621459. URL <http://www.jstor.org/stable/2288126>

Efron B. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika.* 1981;68(3):589–99.

Fisher RA. The fiducial argument in statistical inference. *Ann Hum Genet.* 1935;6(4):391–8.

Friedman J, Hastie T, Tibshirani R. The elements of statistical learning, volume 1. Springer Series in Statistics. 2001.

Gupta RC, Li X. Statistical inference for the common mean of two log-normal distributions and some applications in reliability. *Comput stat data anal.* 2006;50(11):3141–64.

Hall P. On the removal of skewness by transformation. *J R Stat Soc Ser B Methodol.* 1992;54(1):221–8.

Hannig J, Iyer H, Patterson P. Fiducial generalized confidence intervals. *J Am Stat Assoc.* 2006;101(473):254–69. <https://doi.org/10.1198/016214505000000736>.

Hayashi F. *Econometrics*, vol. volume 1. Princeton: Princeton University Press; 2000.

Koenker R. Quantreg: quantile regression. R package version, 4. 2009.

Koenker R, Hallock KF. Quantile regression. *J Econ Perspect.* 2001;15(4):143–56.

Krishnamoorthy K, Mathew T. Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. *J stat plann infer.* 2003;115(1):103–21.

Land CE. An evaluation of approximate confidence interval estimation methods for lognormal means. *Technometrics.* 1972;14(1):145–58.

Manning WG, Mullahy J. Estimating log models: to transform or not to transform? *J Health Econ.* 2001;20(4):461–94.

Manning WG, Basu A, Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. *J Health Econ.* 2005;24(3):465–88.

Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. *J Health Econ.* 1998;17(3):283–95. ISSN 0167-6296. URL <http://ukpmc.ac.uk/abstract/MED/10180919>

McCullagh P, Nelder JA. *Generalized linear models*. Boca Raton: Chapman & Hall/CRC; 1989.

McLachlan GJ, Peel D. *Finite mixture models*, vol. volume 299. Hoboken: Wiley-Interscience; 2000.

Owen WJ, DeRouen TA. Estimation of the mean for lognormal data containing zeroes and left-censored values, with applications to the measurement of worker exposure to air contaminants. *Biometrics.* 1980;36(4):707–19. ISSN 0006341X. URL <http://www.jstor.org/stable/2556125>

Seber GAF, Lee AJ. *Linear regression analysis*, vol. volume 936. Hoboken: Wiley; 2012.

Tian L, Wu J. Confidence intervals for the mean of lognormal data with excess zeros. *Biom J.* 2006;48(1):149–56.

Lili Tian. Inferences on the mean of zero-inflated lognormal data: the generalized variable approach. *Stat Med.* 24(20):3223–3232, 2005. ISSN 1097-0258. <https://doi.org/10.1002/sim.2169>.

Tsui K-W, Weerahandi S. Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *J Am Stat Assoc.* 1989;84(406):602–7. ISSN 01621459. URL <http://www.jstor.org/stable/2289949>

- Weerahandi S. Generalized confidence intervals. *J Am Stat Assoc.* 1993;88(423):899–905. ISSN 01621459. URL <http://www.jstor.org/stable/2290779>
- Weisberg S. *Applied linear regression*, volume 528. Wiley; 2005.
- Welsh AH, Zhou XH. Estimating the retransformed mean in a heteroscedastic two-part model. *J stat plann infer.* 2006;136(3):860–81.
- Wu J, Wong ACM, Jiang G. Likelihood-based confidence intervals for a log-normal mean. *Stat Med.* 2003;22(11):1849–60.
- Zhou XH. Estimation of the log-normal mean. *Stat Med.* 1998;17(19):2251–64.
- Zhou XH, Gao S. Confidence intervals for the log-normal mean. *Stat Med.* 1997;16(7):783–90.
- Zhou XH, Gao S. One-sided confidence intervals for means of positively skewed distributions. *Am Stat.* 2000:100–4.
- Zhou XH, Tu W. Comparison of several independent population means when their samples contain log-normal and possibly zero observations. *Biometrics.* 1999;55(2):645–51.
- Zhou XH, Tu W. Confidence intervals for the mean of diagnostic test charge data containing zeros. *Biometrics.* 2000;56(4):1118–25.
- Zhou XH, Lin H, Johnson E. Non-parametric heteroscedastic transformation regression models for skewed data with an application to health care costs. *J R Stat Soc Ser B Stat Methodol.* 2008;70(5):1029–47.
- Zhou X-H, Gao S, Hui SL. Methods for comparing the means of two independent log-normal samples. *Biometrics.* 1997;53(3):1129–35. ISSN 0006341X. URL <http://www.jstor.org/stable/2533570>



# Instrumental Variable Analysis

# 21

Michael Baiocchi, Jing Cheng, and Dylan S. Small

## Contents

<b>Introduction</b> .....	480
Example: Neonatal Intensive Care Units .....	481
The Fundamentals .....	481
Methods to Address Selection Bias .....	484
Instrumental Variables: NICU Example Revisited .....	485
<b>Sources of Instruments in Health Services Research Studies</b> .....	487
<b>IV Assumptions and Estimation for Binary IV and Binary Treatment</b> .....	490
Framework and Notation .....	490
Two-Stage Least Squares (Wald) Estimator .....	492
More Efficient Estimation .....	493
Estimation with Observed Covariates .....	495
<b>Understanding the Treatment Effect That IV Estimates</b> .....	495
Relationship Between Average Treatment Effect for Compliers and Average Treatment Effect for the Whole Population .....	495
Characterizing the Compliers .....	496
Understanding the IV Estimate When Compliance Status Is Not Deterministic .....	496
<b>Assessing the IV Assumptions and Sensitivity Analysis for Violations of Assumptions</b> .....	498
Assessing the IV Assumptions .....	498
Sensitivity Analysis .....	502

---

M. Baiocchi (✉)  
Department of Statistics, Stanford University, Stanford,  
CA, USA  
e-mail: [baiocchi@stanford.edu](mailto:baiocchi@stanford.edu)

J. Cheng  
Department of Preventive and Restorative Dental Sciences,  
University of California, San Francisco School of  
Dentistry, San Francisco, CA, USA  
e-mail: [jing.cheng@ucsf.edu](mailto:jing.cheng@ucsf.edu)

D. S. Small  
University of Pennsylvania, Philadelphia, PA, USA  
e-mail: [dsmall@wharton.upenn.edu](mailto:dsmall@wharton.upenn.edu)

<b>Weak Instruments</b> .....	504
<b>Binary Outcomes</b> .....	505
Two-Stage Residual Inclusion .....	506
Bivariate Probit Models .....	506
Matching-Based Estimator: Effect Ratio .....	507
<b>Multinomial, Survival and Distributional Outcomes</b> .....	508
Multinomial Outcome .....	508
Survival Outcome .....	508
Effect of Treatment on Distribution of Outcomes .....	509
<b>Study Design IV and Multiple IVs</b> .....	512
Study Design IV: Near-Far Matching .....	512
Multilevel and Continuous IVs .....	514
Multiple IVs .....	514
<b>Multilevel and Continuously Valued</b>	
<b>Treatments</b> .....	515
<b>Extended Instrumental Variable Method for When Proposed IV Has a Direct Effect</b> .....	517
<b>Software</b> .....	517
<b>References</b> .....	518

## Abstract

A goal of many health services research studies is to determine the causal effect of a treatment or intervention on health outcomes. Often, it is not ethically or practically possible to conduct a perfectly randomized experiment, and instead an observational study must be used. A major difficulty with observational studies is that there might be unmeasured confounding, i.e., unmeasured ways in which the treatment and control groups differ before treatment that affect the outcome. Instrumental variable analysis is a method for controlling for unmeasured confounding. Instrumental variable analysis requires the measurement of a valid instrumental variable, which is a variable that is independent of the unmeasured confounding and encourages a subject to take one treatment level versus another, while having no effect on the outcome beyond its encouragement of a certain treatment level. This chapter discusses the types of causal effects that can be estimated by instrumental variable analysis, the assumptions needed for instrumental variable analysis to provide valid estimates of causal effects and sensitivity analysis for those assumptions, methods of estimation of

causal effects using instrumental variables, and sources of instrumental variables in health services research studies.

## Introduction

The goal of health services research is to provide actionable information for policymakers. Modern policy decision makers are driven by data-backed arguments regarding what might change as a result of an intervention. As analysts, this requires specific attention to determining the causal impact between a given intervention and future outcomes. In order to justify a change in the way medicine is practiced, correlation is not sufficient; detecting and quantifying causal connections is necessary.

Medicine has relied on randomized controlled studies as the gold standard for detecting and quantifying causal connections between an intervention and future outcomes. Randomization offers a clear mechanism for limiting the number of alternate possible explanations for what generates the differences between the treated and control groups. The demand for causal evidence in medicine far exceeds the ability to practically control, finance,

and/or conduct randomized studies. Observational data offers a sensible alternative source of data for developing evidence about the implications of different medical interventions. However, for studies using observational data to be considered as a reliable source for evidence of causal effects, great care is needed to design studies in a way that limits the number of alternative explanations for observed differences in outcomes between intervention and control. This chapter will examine instrumental variables as a framework for designing high-quality observational studies. A few of the common pitfalls to be aware of will be discussed.

### Example: Neonatal Intensive Care Units

The development of medical care for premature infants (preemies) has been a spectacular success for modern medicine. This care is offered within neonatal intensive care units (NICUs) of varying intensity of care. Higher-intensity NICUs (those classified as various grades of level 3 by the American Academy of Pediatrics) have more sophisticated medical machinery and highly skilled doctors who specialize in the treatment of tiny preemies.

While establishing value requires addressing questions of both costs and outcomes, our example will focus on estimating the difference in rates of death between the higher level NICUs and the lower level NICUs. Using data from Pennsylvania from the years 1995 to 2005, a simple comparison of death rates at high-level facilities to low-level facilities shows a higher-death rate at high-level facilities, 2.26% compared to 1.25%. This higher-death rate at high-level facilities is surprising only if one assumes preemies were randomly assigned to either a high- or low-level NICU, regardless of how sick they were. In fact, as in most health applications, the sickest patients were routed to the highest level of intensity. As a result, one cannot necessarily attribute the variation in the outcome to variation in the treatment intensity. Fortunately, our data provide a detailed assessment of baseline severity with 45 covariates including variables such as gestational age, birth

weight, congenital disorder indicators, parity, and information about the mother's socioeconomic status. Yet even with this level of detail, our data cannot characterize the full set of clinical factors that a physician or family considers when deciding whether to route a preemie to a high-intensity care unit. As shall be discussed, these missing attributes will cause us considerable problems.

What one wants is not the naïve comparison of rates of death, that is, the percentage of preemies who died at the different types of NICUs, but the difference in probabilities of death for each preemie given whether the preemie was to be delivered at a low-level facility or a high-level facility. This is the causal effect of treatment. This concept is formalized below.

## The Fundamentals

### The Potential Outcome Framework

The literature has made great use of the potential outcome framework (as described in Neyman 1990; Rubin 1974; Holland 1988) as a systematic, mathematical description of the cause-and-effect relationship between variables. Let us assume there are three variables of interest: the outcome of interest  $Y$ , the treatment variable  $D$  (the  $D$  comes from the notion of “dose” rather than a “treatment”), and  $\mathbf{X}$  as a vector of covariates. For most of this chapter it will be assumed that there are only two treatment levels (e.g., the new intervention under consideration vs. the old intervention), though this assumption is only for simplicity's sake and treatments with more than two levels are permissible. These two levels will be referred to using the generic terms “treatment” and “control,” without much discussion of what those two words mean aside from saying that they serve as contrasting interventions to one another. In the potential outcome framework, the notion is that each individual has two possible outcomes – one which is observed if the person were to take the treatment and one if the person were to take the control. In practice only one of these outcomes can be observed because taking the treatment often precludes taking the control and vice versa. Let subject  $i$  taking the treatment be denoted as

$D_i = 1$  and subject  $i$  taking the control as  $D_i = 0$ . To formally denote the outcome subject  $i$  would experience under the treatment and control, write  $Y(D_i = 1)$  and  $Y(D_i = 0)$ , respectively. To simplify the notation, let  $Y_i^1$  and  $Y_i^0$  denote the potential outcome under treatment and the potential outcome under control, respectively. In this chapter,  $Y$  will be thought of as a scalar, though it is possible to develop a framework where  $Y$  is a vector of outcomes. Excellent resources exist for reading up on the potential outcome framework (e.g., Rosenbaum 2002; Pearl 2009; Hernán and Robins 2013).

The ultimate, often unattainable, quantity of interest, namely, the individual level treatment effect, can be described as

$$\Delta_i = Y_i^1 - Y_i^0$$

Thus,  $\Delta_i$  will tell us the difference in outcome, for subject  $i$ , between taking the treatment and control. If this quantity could be observed, then the benefit from intervention would be known explicitly. But, in practice only one or the other of the potential outcomes is observed. To see this, write the observed outcome, denoted  $Y_i^{obs}$  for the  $i$ th individual, as a function of the potential outcomes (Neyman 1990; Rubin 1974):

$$Y_i^{obs} = D_i * Y_i^1 - (1 - D_i) * Y_i^0$$

Observing one of the potential outcomes precludes observing the other. In all but the most contrived settings, this problem is intractable. Both the treatment and control outcomes cannot be observed. So other parameters of interest must be turned to.

### Parameters of Interest

Suppose we, as the analysts, have collected characteristics of the subjects in our study. It is important to stress that these baseline characteristics should be based on the state of the subject prior to the intervention to avoid the potential to bias the treatment effect (see Cox (1958, Sect. 4.2) and Rosenbaum (2002)). For example, say a new drug is being tested for its ability to lower the risk of

heart attack. High blood pressure is known to correlate with higher risk of heart attack, so it is tempting to control for this covariate. Controlling for blood pressure is likely to improve the precision of the estimate if a pretreatment blood pressure measure is used. It would be a mistake to use a posttreatment measurement of blood pressure as a control because this measurement may be affected by the drug and would thus result in an attenuated estimated causal effect. Intuitively, this is because the estimation procedure is limiting comparison in outcome not just between people who took the drug and who didn't but between people who took the drug and then had a certain level of blood pressure to people didn't take the drug and had the same level of blood pressure. The impact from the drug may have already happened via the lowering of the blood pressure.

Let's denote these measured pretreatment characteristics as  $\mathbf{X}_i$  for the  $i$ th subject. Further, the subjects are likely to have characteristics which were not recorded. Let's denote these unobserved characteristics as  $\mathbf{U}_i$  for the  $i$ th subject. To keep things simple, assume that the covariates are linearly related to the outcomes like so

$$\begin{aligned} Y_i^1 &= \mathbf{X}_i^T \boldsymbol{\beta}^1 + \mathbf{U}_i^T \boldsymbol{\alpha}^1 \\ Y_i^0 &= \mathbf{X}_i^T \boldsymbol{\beta}^0 + \mathbf{U}_i^T \boldsymbol{\alpha}^0 \end{aligned}$$

Note that the coefficients need to be indexed by the treatment level in order to account for interactions between the treatment level and the covariates. Also, it may appear strange putting coefficients on the unobserved variables, but this is required at the bare minimum to make the dimensions agree. In practice, let us write  $\varepsilon_i^1$  in place of the clunkier  $\mathbf{U}_i^T \boldsymbol{\alpha}^1$ , but this is a move of convenience rather than discipline. There is likely not just one scalar, unobservable covariate omitted from our dataset, so it is more realistic to write  $\mathbf{U}_i^T \boldsymbol{\alpha}^1$ . Note that this means something a bit magical is happening when an author proposes a parametric distribution for  $\varepsilon_i^1$ .

Combining our equations for the observed outcome and the linear models, the observed outcome can be decomposed in terms of covariates, both observed and unobserved, as well as the treatment:



$$Y_i^{obs} = \mathbf{X}_i^T \boldsymbol{\beta}^0 + D_i [(\mathbf{X}_i^T \boldsymbol{\beta}^1 - \mathbf{X}_i^T \boldsymbol{\beta}^0) + (\mathbf{U}_i^T \boldsymbol{\alpha}^1 - \mathbf{U}_i^T \boldsymbol{\alpha}^0)] + \mathbf{U}_i^T \boldsymbol{\alpha}^0$$

It is standard in econometrics to think of the above model as a regression, where the coefficient on the treatment variable comes from two sources of variation: the first source is the variation due to the observed covariates  $(\mathbf{X}_i^T \boldsymbol{\beta}^1 - \mathbf{X}_i^T \boldsymbol{\beta}^0)$  and the second is the variation due to the unobserved covariates,  $(\mathbf{U}_i^T \boldsymbol{\alpha}^1 - \mathbf{U}_i^T \boldsymbol{\alpha}^0)$ , where  $D_i$  may be correlated with  $U_i$ . It is common to interpret the first source of variation as the gains for the average person with covariate levels  $\mathbf{X}_i$  and the second source of variation to be referred to as idiosyncratic gains for subject  $i$ . The idiosyncratic gains are the part of this model which allows persons  $i$  and  $j$  to differ in gains from treatment even when  $\mathbf{X}_i = \mathbf{X}_j$ .

**Selection Bias**

One of the biggest problems with observational studies is that there is selection bias. Loosely speaking, selection bias arises from how the subjects are sorted (or sort themselves) into the treatment or control groups. The intuition here is the treatment group was different from the control group even before the intervention, and the two groups would probably have had different outcomes even if there had been no intervention at all. Selection bias can occur in a couple of different ways, but one way to write it is

$$f(\mathbf{X}, \mathbf{U} | D = 1) \neq f(\mathbf{X}, \mathbf{U} | D = 0)$$

that is, the joint distribution of the covariates for those who received the treatment is different than for those who received the control. If this is true, that there is selection bias, then

$$E[Y^1 - Y^0 | \mathbf{X}] \neq E[Y(1) | \mathbf{X}, D = 1] - E[Y(0) | \mathbf{X}, D = 0]$$

This is problematic because the left-hand side of this equation is our unobservable quantity of interest, but the right-hand side is made up of directly observable quantities. But it seems like the above equation is used in other settings, namely, experimentation. Why is that acceptable?

In an experiment, because of randomization, it is known that

$$(X, U) \perp\!\!\!\perp D,$$

where  $\perp\!\!\!\perp$  denotes independence. And it follows that

$$E[Y^1 - Y^0 | \mathbf{X}] = E[Y^1 | \mathbf{X}, D = 1] - E[Y^0 | \mathbf{X}, D = 0]$$

Though it is often a dubious claim, many of the standard observational study techniques require an assumption which essentially says that the only selection between treated and control groups is on levels of the observed covariates, i.e.,  $U \perp\!\!\!\perp D | \mathbf{X}$ . This is sometimes referred to as overt selection bias. Typically, if overt selection bias is the only form of bias, then either conditioning on observed covariates (e.g., by using a regression) or matching is enough to address overt bias. One particular assumption that is invoked quite often in the current health literature is the absence of omitted variables (i.e., only overt bias).

Hidden bias exists when there are imbalances in the unobserved covariates. Let's use the observed outcome formula again, rewriting it like so:

$$Y_i^{obs} = \mathbf{X}_i^T \boldsymbol{\beta}^0 + D_i E[\Delta | \mathbf{X}] + \mathbf{U}_i^T \boldsymbol{\alpha}^0 + D_i (\mathbf{U}_i^T \boldsymbol{\alpha}^1 - \mathbf{U}_i^T \boldsymbol{\alpha}^0)$$

A least squares regression of  $Y$  on  $D$  based on the model above will tend to produce biased estimates for  $E[\Delta | \mathbf{X}]$  when  $D$  is correlated with either  $\mathbf{U}_i^T \boldsymbol{\alpha}^0$  or  $(\mathbf{U}_i^T \boldsymbol{\alpha}^1 - \mathbf{U}_i^T \boldsymbol{\alpha}^0)$ . This can arise from unobserved covariates which influence both potential outcomes and selection into treatment. This bias is referred to as *hidden bias*. If the average treatment effects given  $\mathbf{X}$ ,  $E[\Delta | \mathbf{X}]$ , and the hidden biases given  $\mathbf{X}$ ,  $E[\mathbf{U}_i^T \boldsymbol{\alpha}^1 | \mathbf{X}, D = 1] - E[\mathbf{U}_i^T \boldsymbol{\alpha}^0 | \mathbf{X}, D = 0]$ , are the same for all  $\mathbf{X}$ , then the regression estimate of  $E[\Delta | \mathbf{X}]$  is biased by

$$E[\mathbf{U}_i^T \boldsymbol{\alpha}^1 | \mathbf{X}, D = 1] - E[\mathbf{U}_i^T \boldsymbol{\alpha}^0 | \mathbf{X}, D = 0]$$

## Methods to Address Selection Bias

In a randomized experiment setting, inference on the causal effect of treatment on the outcome requires no further assumption than the method for randomizing subjects into the treatment or control (Fisher 1949). The randomization guarantees independence of assigned treatment from the covariates. This independence is for all covariates, both observed and unobserved. By observed covariates we mean those covariates which appear in the analyst's data set and unobserved all of those that don't. If the sample is large enough, then this independence means that the treatment group will almost surely have quite a similar covariate distribution as the control group. Therefore, any variation noted in the outcome is more readily attributed to the variation in the treatment level rather than variation in the covariates.

The primary challenge to observational studies is that selection into treatment is not randomly assigned. Usually there are covariates, both observed and unobserved, which determine who receives treatment and who receives control. In such a case, variation in the outcome is not easily attributable to treatment levels because covariates are different between the different levels as well. There are techniques which were created to address this selection bias. These methods can be classified (roughly) into two groups: (1) those methods which address only the observed selection bias and (2) those methods which attempt to address selection bias on both the observed as well as unobserved covariates. Falling into the first category are techniques like regression, Bayesian hierarchical modeling, propensity score matching, and inverse probability weighting. The second category includes methods like instrumental variables, regression discontinuity, and difference in differences.

### Methods to Address Overt Selection Bias

Only through special justification should methods which address only overt bias be considered valid. Usually, this justification takes the form of an assumption. Informally, this assumption can be thought of as saying: selection into the treatment is occurring only on variables

which are observed. Formally, this assumption is often written as

$$(Y^0, Y^1) \perp\!\!\!\perp D \mid \mathbf{X}, \\ 0 < pr(D = 1 \mid \mathbf{X}) < 1$$

where  $\perp\!\!\!\perp$  denotes the conditional independence between the treatment and the joint distribution of the counterfactual outcomes given  $\mathbf{X}$ . Two random variables are conditionally independent given a third variable if and only if they are independent in their conditional distribution given the conditioning variable. The above assumption, essentially saying that all needed covariates are measured, has a few different names: strongly ignorable treatment assignment (Rosenbaum and Rubin 1983), selection on observables (Heckman and Robb 1985), conditional independence, no hidden bias (only overt bias due to  $\mathbf{X}$ ), no unmeasured confounders (in the epidemiology literature), or the absence of omitted variable bias (in the econometrics literature).

To assume strongly ignorable treatment assignment in a medical application is to go a bit against intuition. In practice, the analyst has access to some subset of the recorded information from the patients' interaction with the health system. Currently, most analysts do not have access to many measurements the medical decision makers have (e.g., results of labs, biometric information), so they are forced to use transactional information (e.g., insurance claims) which are good for indicating the presence of a condition but not necessarily the severity. It is possible that as electronic health records become more readily available, this problem will diminish, but currently this should be a source of great skepticism for methods relying on the assumption of no unobserved bias. But the issue does not stop here. The health analyst should be aware that medical practitioners are keen observers and intuitively adept at identifying issues which may go either unrecorded or may even be unquantifiable (e.g., practitioners will regularly refer to the frailty of a patient, which seems to be a generally understood yet unmeasurable quality of a patient). Given the additional information the medical decision

makers have, and their desire to choose an optimal outcome, medical decision makers are actively working against the reasonableness of strongly ignorable treatment assignment.

It is unfortunate that methods which were designed only to address overt bias have become the default tools of choice in the literature. Given the complexity of health decision, it strains credibility that all variables which influence treatment and outcome are recorded and available to the analyst. The default for health analysts (and critically minded reviewers) should be to assume unobserved selection is occurring and to look for ways of mitigating it.

### **Instrumental Variables: A Framework to Address Overt Bias and Bias Due to Omitted Variables**

Regression, propensity score matching, and any methods predicated on only overt bias do not address selection on unobserved covariates. It is important to be aware of this because a well-informed researcher needs to judge if available covariates are enough to make a compelling argument for the absence of omitted variables. This is often a dubious claim because (1) a clever reviewer will find several variables missing from your data set and/or (2) there are intangible variables that are difficult, or perhaps inconceivable, to measure. Instrumental variable (IV) techniques are one way of addressing unobserved selection bias.

It is important to note IV techniques do not come for free, without hefty assumptions. It is important to consider these assumptions carefully before deciding to use an IV analysis.

An instrumental variable (IV) design takes advantage of randomness which occurs in the treatment assignment to help address imbalances in the unobserved variables. An instrument is a haphazard nudge toward acceptance of a treatment that affects outcomes only to the extent that it affects acceptance of the treatment. In settings in which treatment assignment is mostly deliberate and not random, there may nevertheless exist some essentially random nudges to accept treatment, so that use of an instrument might extract bits of random treatment assignment from a

setting that is otherwise quite biased in its treatment assignments.

There have been many different formulations of IV, reflecting the diverse academic traditions that use IV. Though IVs existed in the literature for quite some time, Angrist et al. (1996) used the potential outcome framework to bring greater clarity to the math of IV. For the health analyst, perhaps Holland (1988) offers the most intuitive introduction to IVs, framing IV as a randomized trial with noncompliance. The frameworks for IV discussed in both Angrist et al. (1996) and Holland (1988) enhance the classic econometric presentation of IVs where the focus is on correlation with the error term. Health analysts will likely find these introductions most engaging.

To illustrate IVs, consider the NICU example from earlier.

### **Instrumental Variables: NICU Example Revisited**

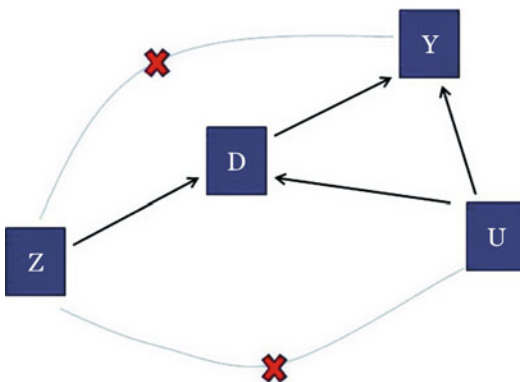
Neonatal intensive care units (NICUs) have been established to deliver high-intensity care for premature infants (those infants born before 37 weeks of gestation). Considering all of the preemies that were delivered in Pennsylvania between 1995 and 2005, 2.26% of the preemies delivered at high-level NICUs died, while only 1.25% of the preemies who were delivered at low-level NICUs died. No one believes the difference in outcomes reported above is solely attributable to the difference in level of intensity of treatment. People believe it is due to difference in covariates. Based on the observable covariates, this is plausible because the preemies delivered at high-level NICUs weighed almost 250 g less than the preemies which were delivered at low-level NICUs (2,454 at high-level NICUs vs. 2,693 at low-level NICUs). Similarly preemies delivered at high-level NICUs were born a week earlier than their counterparts at low-level NICUs on average (34.5 vs. 35.5 weeks).

Unfortunately, full medical records were not available for this study. Only birth and death certificates and a form UB-92 that hospitals provided were available. It is quite likely that not all

necessary covariates in our dataset are available, so assuming only overt bias is likely to lead to biased estimates. To attempt to deal with this problem, Baiocchi et al. (2010) and Lorch et al. (2012a) used an instrumental variable approach. They used distance to treatment facility as an instrument, because travel time largely determines the likelihood that mother will deliver at a given facility but appears to be largely uncorrelated with the level of severity a preemie experiences.

To help visualize the problem, look at Fig. 1 below. This is an example of a directed acyclic graph (Pearl 2009). The arrows denote causal relationships. Read the relationship between variables  $D$  and  $Y$  like so: changing the value of  $D$  causes  $Y$  to change. In our example,  $Y$  represents mortality. The variable  $D$  indicates whether or not a baby attended a high-level NICU. Our goal is to understand the arrow connecting  $D$  to  $Y$ . In order to keep the current example simple, assume there are no observed covariates (which would be denoted using an  $X$  in Fig. 1). In general, IV techniques are able to adjust for variation in observed covariates (see section “Estimation with Observed Covariates”).

The  $U$  variable causes consternation as it represents the unobserved level of severity of the preemie, and it is causally linked to both mortality (sicker babies are more likely to die) and to which treatment the preemies selects (sicker babies are more likely to be delivered in high-level NICUs). Because  $U$  is unobserved directly, it cannot be



**Fig. 1** Directed acyclic graph for the relationship between an instrumental variable  $Z$ , a treatment  $D$ , unmeasured confounders  $U$ , and an outcome  $Y$

precisely adjusted for using statistical methods such as propensity scores or regression. If the story stopped with just  $D$ ,  $Y$ , and  $U$ , then the effect of  $D$  on  $Y$  could not be estimated.

Instrumental variable estimation makes use of an uncomplicated form of variation in the system. What is needed is a variable, typically called an instrument (represented by  $Z$  in Fig. 1) that has very special characteristics. It takes some practice to understand exactly what constitutes a good instrumental variable.

Consider excess travel time as a possible instrument. Excess travel time is defined as the time it takes to travel from the mother’s residence to the nearest high-level NICU minus the time it takes to travel to the nearest low-level NICU. If the mother lives closest to a high-level NICU, then excess travel time will take on negative values. If she lives closest to a low-level NICU, excess travel time will be positive.

There are three key features a variable must have in order to qualify as an instrument (see section “IV Assumptions and Estimation for Binary IV and Binary Treatment” for mathematical details on these features and additional assumptions for IV methods). The first feature (represented by the directed arrow from  $Z$  to  $D$  in Fig. 1) is that the instrument causes a change in the treatment assignment. When a woman becomes pregnant she has a high probability of establishing a relationship with the proximal NICU, regardless of the level, because she is not anticipating having a preemie. Proximity as a leading determinate in choosing a facility has been discussed in Phibbs et al. (1993). By selecting where to live, mothers assign themselves to be more or less likely to deliver in a high-level NICU. The fact that changes in the instrument are associated with changes in the treatment is verifiable from the data.

The second feature (represented by the crossed out arrow from  $Z$  to  $U$ ) is that the instrument is not associated with variation in unobserved variables  $U$  that also affect the outcome. That is,  $Z$  is not connected to the unobserved confounding that was a worry to begin with. In our example, this would mean unobserved severity is not caused by variation in geography. Since high-level NICUs

tend to be in urban areas and low-level NICUs tend to be the only type in rural areas, this assumption would be dubious if there were high-level of pollutants in urban areas (think of Manchester, England circa the Industrial Revolution) or if there were more pollutants in the drinking water in rural areas than in urban areas. The pollutants may have an impact on the unobserved levels of severity. The assumption that the instrument is not associated with variation in the unobserved variables, while most certainly an assumption, can at least be corroborated by looking at the values of variables that are perhaps related to the unobserved variables of concern (see section “Assessing the IV Assumptions”).

The third feature (represented by the crossed out line from  $Z$  to  $Y$  in Fig. 1) is that the instrument does not cause the outcome variable to change directly. That is, it is only through its impact on the treatment that the instrumental variable affects the outcome. In our case, presumably a nearby hospital with a high-level NICU affects mortality only if the baby receives care at that hospital. That is, proximity to a high-level NICU in and of itself does not change the probability of death for a premie, except through the increased probability of the premie being delivered at the high-level NICU. This is often referred to as the exclusion restriction and can be a slippery concept to get a hold of. See Angrist et al. (1996) for discussion of the exclusion restriction. In our case it seems quite reasonable.

---

## Sources of Instruments in Health Services Research Studies

In this section, common types of IVs that have been used in health services research studies will be described, and issues to consider in assessing their validity will be discussed. One way to study the effect of a treatment when that treatment cannot be controlled is to conduct a randomized encouragement trial. In such a trial, some subjects are randomly chosen to get extra encouragement to take the treatment and the rest of the subjects receive no extra encouragement (Holland 1988). For example, Permutt and Hebel (1989) studied

the effect of maternal smoking during pregnancy on an infant's birthweight using a randomized encouragement trial in which some mothers received extra encouragement to stop smoking through a master's level staff person providing information, support, practical guidance, and behavioral strategies (Sexton and Hebel 1984). For a randomized encouragement trial, the randomized encouragement assignment (1 if encouraged, 0 if not encouraged) is a potential IV. The randomized encouragement is independent of unmeasured confounders because it is randomly assigned by the investigators and will be associated with the treatment if the encouragement is effective. The only potential concern with the randomized encouragement being a valid IV is that the randomized encouragement might have a direct effect on the outcome not through the treatment. For example, in the randomized encouragement trial to encourage expectant mothers to stop smoking, the encouragement could have a direct effect if the staff person providing the encouragement also encouraged expectant mothers to stop drinking alcohol during pregnancy. To minimize a potential direct effect of the encouragement (Sexton and Hebel 1984) asked the staff person providing encouragement to avoid recommendations or information concerning other habits that might affect birthweight such as alcohol or caffeine consumption and also prohibited discussion of maternal nutrition or weight gain.

When comparing two treatments, one of which is only provided by specialty care providers and one of which is provided by more general providers, the distance a person lives from the nearest specialty care provider has often been used as an IV. Proximity to a specialty care provider particularly enhances the chance of being treated by the specialty care provider for acute conditions. For less acute conditions, patients/providers have more time to decide and plan where to be treated, and proximity may have less of an influence on treatment selection. For treatments that are stigmatized such as substance abuse treatment, proximity could have a negative effect on the chance of being treatment. A classic example of the use of distance as an IV is McClellan et al.'s study of the

effect of cardiac catheterization for patients suffering a heart attack (McClellan et al. 1994); the IV used in the study was the differential distance the patient lives from the nearest hospital that performs cardiac catheterization to the nearest hospital that does not perform cardiac catheterization. Another example is the study of the effect of high-level versus low-level NICUS (Lorch et al. 2012a) that was discussed in section “[Instrumental Variables: NICU Example Revisited](#).” Because distance to a specialty care provider is often associated with socioeconomic characteristics, it will typically be necessary to control for socioeconomic characteristics in order for distance to potentially be independent of unmeasured confounders. The possibility that distance might have a direct effect because the time it takes to receive treatment affects outcomes needs to be considered in assessing whether distance is a valid IV.

A general strategy for finding an IV for comparing two treatments  $A$  and  $B$  is to look for naturally occurring variation in medical practice patterns at the level of geographic region, hospital or individual physician, and then use whether the region/hospital/individual physician has a high or low use of treatment  $A$  as the IV. Brookhart and Schneeweiss (2007) termed these IVs “preference-based instruments” because they are derived from the assumption that different providers or groups of providers have different preferences or treatment algorithms dictating how medications or medical procedures are used. Examples of studies using preference-based IVs are Brooks et al. (2004) that studied the effect of surgery plus irradiation versus mastectomy for breast cancer patients using geographic region as the IV (Johnston 2000) that studied the effect of surgery versus endovascular therapy for patients with a ruptured cerebral aneurysm using hospital as the IV and Brookhart et al. (2006) that studied the benefits and risks of selective cyclooxygenase 2 inhibitors versus non-selective nonsteroidal antiinflammatory drugs for treating gastrointestinal problems using individual physician as the IV. For proposed preference-based IVs, it is important to consider that the patient mix may differ between the different groups of

providers with different preferences, which would make the preference-based IV invalid unless patient mix is fully controlled for. It is useful to look at whether measured patient risk factors differ between groups of providers with different preferences. If there are measured differences, there are likely to be unmeasured differences as well; see section “[Assessing the IV Assumptions and Sensitivity Analysis for Violations of Assumptions](#)” for further discussion. Also, for proposed preference-based IVs, it is important to consider whether the IV has a direct effect; a direct effect could arise if the group of providers that prefers treatment  $A$  treats patients differently in ways other than the treatment under study compared to the providers who prefer treatment  $B$ . For example, Newman et al. (2012)s studied the efficacy of phototherapy for newborns with hyperbilirubinemia and considered the frequency of phototherapy use at the newborn’s birth hospital as an IV. However, chart reviews revealed that hospitals that use more phototherapy also have a greater use of infant formula; use of infant formula is also thought to be an effective treatment for hyperbilirubinemia. Consequently, the proposed preference-based IV has a direct effect (going to a hospital with higher use of phototherapy also means a newborn is more likely to receive infant formula even if the newborn does not receive phototherapy) and is not valid. The issue of whether a proposed preference-based IV has a direct effect can be studied by looking at whether the IV is associated with concomitant treatments like use of infant formula (Brookhart and Schneeweiss 2007). A related way in which a proposed preference-based IV can have a direct effect is that the group of providers who prefer treatment  $A$  may have more skill than the group of providers who prefer treatment  $B$ . Also, providers who prefer treatment  $A$  may deliver treatment  $A$  better than those providers who prefer treatment  $B$  because they have more practice with it, for example, doctors who perform surgery more often may perform better surgeries. Korn and Baumrind (1998) discuss a way to assess whether there are provider skill effects by collecting data from providers on whether or not they would have treated a

different provider's patient with treatment  $A$  or  $B$  based on the patient's pretreatment records.

Another common source for an IV is calendar time. Variations in the use of one treatment versus another could result from changes in guidelines, changes in formularies or reimbursement policies, changes in physician preference (e.g., due to marketing activities by drug makers), release of new effectiveness or safety information, or the arrival of new treatments to the market (Brookhart et al. 2010). For example, Shetty et al. (2009) studied the effect of hormone replacement therapy (HRT) on cardiovascular health among postmenopausal women using calendar time as an IV. HRT was widely used among postmenopausal women until 2002; observational studies had suggested that HRT reduced cardiovascular risk, but the Women's Health Initiative randomized trial reported opposite results in 2002, which caused HRT use to drop sharply. A proposed IV based on calendar time could be associated with confounders that change in time such as the characteristics of patients who enter the cohort, changes in other medical practices, and changes in medical coding systems (Brookhart et al. 2010). The most compelling type of IV based on calendar time is one where a dramatic change in practice occurs in a relatively short period of time (Brookhart et al. 2010).

Another general source for potential IVs is genetic variants which affect treatment variables. For example, Voight et al. (2012) studied the effect of HDL cholesterol on myocardial infarction using as an IV the genetic variant LIPG 396Ser allele for which carriers have higher levels of HDL cholesterol but similar levels of other lipid and non-lipid risk factors compared with noncarriers. Another example is Wehby et al. (2011) that studied the effect of maternal smoking on orofacial clefts in their babies using genetic variants that increase the probability that a mother smokes as IVs. The approach of using genetic variants as an IV is called *Mendelian randomization* because it makes use of the random assignment of genetic variants conditional on parents' genes discovered by Mendel. Although genetic variants are randomly assigned conditional

on a parent's genes, genetic variants need to satisfy additional assumption to be valid IVs:

1. *Not associated with unmeasured confounders through population stratification.* Most Mendelian randomization analyses do not condition on parents' genes, creating the potential of the proposed genetic variant IV being association with unmeasured confounders through population stratification. Population stratification is a condition where there are subpopulations, some of which are more likely to have the genetic variant, and some of which are more likely to have the outcome through mechanisms other than the treatment being studied. For example, consider studying the effect of alcohol consumption on hypertension. Consider using the ALDH2 null variant, which is associated with alcohol consumption, as an IV (individuals who are homozygous for the ALDH2 null variant have severe adverse reactions to alcohol consumption and tend to drink very little (Lawlor et al. 2008)). The ALDH2 null variant is much more common in people with Asian ancestry than other types of ancestry (Goedde et al. 1992). Suppose ancestry was not fully measured. If ancestry is associated with hypertension through means other than differences in the ALDH2 null variant (e.g., through different ancestries tending to have different diets), then ALDH2 would not be a valid IV because it would be associated with an unmeasured confounder.
2. *Not associated with unmeasured confounders through genetic linkage.* Genetic linkage is the tendency of genes that are located near to each other on a chromosome to be inherited together because the genes are unlikely to be separated during the crossing over of the mother's and father's DNA (Sham 1998). Consider using a gene  $A$  as an IV where gene  $A$  is genetically linked to a gene  $B$  that has a causal effect on the outcome through a pathway other than the treatment being studied. If gene  $B$  is not measured and controlled for, then gene  $A$  is not a valid IV because it is associated with the unmeasured confounder gene  $B$ .

3. *No direct effect through pleiotropy.* Pleiotropy refers to a gene having multiple functions. If the genetic variant being used as an IV affects the outcome through a function other than affecting the treatment being studied, this would mean the genetic variant has a direct effect. For example, consider the use of the APOE genotype as an IV for studying the causal effect of low-density lipoprotein cholesterol (LDLc) on myocardial infarction (MI) risk. The *d2* variant of the APOE gene is associated with lower levels of LDLc but is also associated with higher levels of high-density lipoprotein cholesterol, less efficient transfer of very low-density lipoproteins and chylomicrons from the blood to the liver, greater postprandial lipemia, and an increased risk of type III hyperlipoproteinemia (the last three of which are thought to increase MI risk) (Lawlor et al. 2008). Thus, the gene APOE is pleiotropic, affecting myocardial infarction risk through different pathways, making it unsuitable as an IV to examine the causal effect of any one of these pathways on MI risk.

Didelez and Sheehan (2007) and Lawlor et al. (2008) provide good reviews of Mendelian randomization methods.

Another source of IVs for health services research studies is timing of admission variables. For example, Ho et al. (2000) used day of the week of hospital admission as an IV for waiting time for surgery to study the effects of waiting time on length of stay and inpatient mortality among patients admitted to the hospital with a hip fracture. Day of the week of admission is associated with waiting time for surgery because many surgeons only do non-emergency operations on weekdays, and therefore patients admitted on weekends may have to wait longer for surgery. In order for weekday versus weekend admission to be a valid IV, patients admitted on weekdays versus weekends must not differ on unmeasured characteristics (i.e., the IV is independent of unmeasured confounders) and other aspects of hospital care that affect the patients' outcomes besides surgery

must be comparable on weekdays versus weekends (i.e., the IV has no direct effect). Another example of a timing of admission variable used as an IV is hour of birth as an IV for a newborn's length of stay in the hospital (Goyal et al. [in press](#); Malkin et al. 2000).

An additional general source of potential IVs for health services research studies is insurance plans which may vary in the amount of reimbursement they provide for different treatments. For example, Cole et al. (2006) used drug co-payment amount as an IV to study the effect of  $\beta$ -blocker adherence on clinical outcomes and health-care expenditures after a hospitalization for heart failure. In order for variations in insurance plan like drug co-payment amount to be a valid IV, insurance plans must have comparable patients after controlling for measured confounders (i.e., the IV is independent of unmeasured confounders), and insurance plans must not have an effect on the outcome of interest other than through influencing the treatment being studied (i.e., the IV has no direct effect).

---

## IV Assumptions and Estimation for Binary IV and Binary Treatment

In this section, the simplest setting of a binary instrument and a binary treatment will be considered. The main ideas in instrumental variable methods are most easily understood in this setting, and the ideas will be expanded to more complicated settings later.

### Framework and Notation

The Neyman-Rubin potential outcome framework will be used to describe causal effects (Neyman 1990; Rubin 1974). Let  $Z_i$  denote the IV for subject  $i$ , where  $Z_i = 0$  or 1 for a binary IV. Level 1 of the IV is assumed to mean the subject was encouraged to take level 1 of the treatment, where the treatment has levels 0 and 1. Let  $D_i^z$  be the potential treatment received for subject  $i$  if she were assigned level



$z$  of the IV –  $D_i^1$  is the treatment that subject  $i$  would receive if she were assigned level 1 of the IV and  $D_i^0$  is treatment that  $i$  would receive if she were assigned level 0 of the IV. The observed treatment received for subject  $i$  is  $D_i \equiv D_i^{Z_i}$ . Let  $Y_i^{z,d}$  be the potential outcome for subject  $i$  if she were assigned level  $z$  of the IV and level  $d$  of the treatment – there are four such potential outcomes  $Y_i^{1,1}, Y_i^{1,0}, Y_i^{0,1}$ , and  $Y_i^{0,0}$ . However, only one of them will be observed in practice. The observed outcome for subject  $i$  is  $Y_i \equiv Y_i^{Z_i, D_i^{Z_i}}$ . Let  $\mathbf{X}_i$  denote observed covariates for subject  $i$ .

Angrist et al. (1996) considered an IV to be a variable satisfying the following five assumptions:

1. *IV is correlated with treatment received*  $E(D_i^1 | \mathbf{X}_i) > E(D_i^0 | \mathbf{X}_i)$ .
2. *IV is independent of unmeasured confounders (conditional on covariates).*

$Z_i$  is independent of  $(D_i^1, D_i^0, Y_i^{1,1}, Y_i^{0,1}, Y_i^{0,0}) | \mathbf{X}_i$ .

3. *Exclusion restriction (ER).* This assumption says that the IV affects outcomes only through its effect on treatment received:  $Y_i^{z,d} = Y_i^{z,d}$ . Under the ER, write  $Y_i^d = Y_i^{z,d}$  for any  $z$ , that is,  $Y_i^1$  is the potential outcome for subject  $i$  if she were to receive level 1 of the treatment (regardless of her level of the IV), and  $Y_i^0$  is the potential outcome if she were to receive level 0 of the treatment. This assumption is called the no direct effect assumption.
4. *Monotonicity assumption.* This assumption says that there are no subjects who are “defiers,” who would only take level 1 of the treatment if not encouraged to do so, that is, no subjects with  $D_i^1 = 0, D_i^0 = 1$ .

5. *Stable unit treatment value assumption (SUTVA).* This assumption says that the treatment affects only the subject taking the treatment and the treatment effect is stable through time (see Angrist et al. 1996; Rubin 1990 for details). The first part of this assumption that the treatment affects only the subject taking the treatment is called the no interference assumption.

The first three assumptions are the assumptions depicted in Fig. 1.

The fourth assumption, monotonicity, plays a role in interpreting the standard IV estimate as a causal effect for a certain subpopulation. A subject in a study with binary IV and treatment can be classified into one of four latent compliance classes based on the joint values of potential treatment received (Angrist et al. 1996):  $C_i =$  never taker (nt) if  $(D_i^0, D_i^1) = (0, 0)$ , complier (co) if  $(D_i^0, D_i^1) = (0, 1)$ , always taker (at) if  $(D_i^0, D_i^1) = (1, 1)$ , and defier (de) if  $(D_i^0, D_i^1) = (1, 0)$ . Table 1 shows the relationship between observed groups and latent compliance classes. Under the monotonicity assumption, the set of defiers will be empty. The never takers and always takers do not change their treatment status when the instrument changes, so under the ER assumption, the potential treatment and potential outcome under either level of the IV ( $Z_i = 1$  or  $0$ ) is the same. Consequently, the IV is not helpful for learning about the treatment effect for always takers or never takers. Compliers are subjects who change their treatment status with the instrument, that is, the subjects would take the treatment if they were encouraged to take it by the IV but would not otherwise take the treatment. Because these subjects change their treatment with the level of the IV, the IV is helpful for learning about their treatment effects. The average causal effect for this subgroup,

**Table 1** The relation between observed groups and latent compliance classes

$Z_i$	$D_i$		$C_i$	
1	1	Complier	or	Always taker
1	0	Never taker	or	
0	0	Never taker	or	Complier
0	1	Always taker	or	

$E(Y_i^1 - Y_i^0 | C_i = co)$ , is called the complier average causal effect (CACE) or the local average treatment effect (LATE). It provides the information on the average causal effect of receiving the treatment

for compliers. When monotonicity does not hold, the standard IV estimator Eq. 3 discussed in section “Two Stage Least Squares (Wald) Estimator” estimates the quantity (Angrist et al. 1996).

$$\frac{E(Y_i^1 - Y_i^0 | C_i = co) \times \frac{P(C_i = co)}{P(C_i = co) + P(C_i = de)} - E(Y_i^1 - Y_i^0 | C_i = de) \times \frac{P(C_i = de)}{P(C_i = co) + P(C_i = de)}}{P(C_i = co) - P(C_i = de)} \tag{1}$$

Equation 1 could potentially be negative even if the treatment has a positive effect for all subjects (Angrist et al. 1996). However, the IV method estimate of the CACE is not generally sensitive to small violations of the monotonicity assumption (Angrist et al. 1996). Additionally, if the treatment has the same effect for compliers and defiers, the monotonicity assumption is not needed as Eq. 1 equals the CACE,  $E(Y_i^1 - Y_i^0 | C_i = co)$  (Robins and Greenland 1996). For further discussion of understanding the treatment effect that the IV method estimates, see section “Understanding the Treatment Effect That IV Estimates.”

The fifth IV assumption, SUTVA, also plays a role in interpreting what the standard IV method estimate Eq. 3 estimates. Consider in particular the no interference assumption part of SUTVA that subject  $A$  receiving the treatment affects only subject  $A$  and not other subjects. In the NICU study, the no interference assumption is reasonable – if preemie  $A$  is treated at a high-level NICU, this does not affect preemie  $B$ ’s outcome. If there were crowding effects (e.g., treating additional babies at a hospital decreases the quality of care for babies already under care that hospital), this assumption might not be true. SUTVA is also not appropriate for situations like estimating the effect of a vaccine on an individual because herd immunization would lead to causal links between different people (Hudgens and Halloran 2008). When no interference fails to hold, the IV method is roughly estimating the difference between the effect of the treatment and the spillover effect of some

units being treated on those units left untreated (see Sobel 2006 for a precise formulation and details).

In economics, a latent index model is often considered for causal inference about the effect of a binary treatment based on a structural equation model or two-stage linear model, for example,

$$\begin{aligned} D_i^* &= \alpha_0 + \alpha_1 Z_i + \varepsilon_{i1} \\ Y_i &= \beta_0 + \beta_1 D_i + \varepsilon_{i2} \end{aligned}$$

where

$$D_i = \begin{cases} 1 & \text{if } D_i^* > 0 \\ 0 & \text{if } D_i^* \leq 0 \end{cases}$$

$$Z_i \perp\!\!\!\perp \varepsilon_{i1}, \varepsilon_{i2}$$

Vytlacil (2002) shows that a nonparametric version of the latent index model is equivalent to the Assumptions 1–5 above that Angrist et al. (1996) use to define an IV.

### Two-Stage Least Squares (Wald) Estimator

Let us first consider IV estimation when there are no observed covariates  $\mathbf{X}$ . For binary IV and treatment variable, Angrist et al. (1996) show that under the framework and assumptions in section “Two Stage Least Squares (Wald) Estimator,” the CACE is nonparametrically identified by

$$\begin{aligned}
 E(Y_i^{-1} - Y_i^0 | C_i = co) &= \frac{E(Y_i | Z_i = 1) - E(Y_i | Z_i = 0)}{E(D_i | Z_i = 1) - E(D_i | Z_i = 0)}, \\
 \end{aligned} \tag{2}$$

which is the intention-to-treat (ITT) effect divided by the proportion of compliers.

The standard IV estimator or two-stage least squares estimator (2SLS) is the ratio of sample covariances (Durbin 1954):

$$\begin{aligned}
 \widehat{CACE}_{2SLS} &= \frac{c\hat{ov}(Y_i, Z_i)}{c\hat{ov}(D_i, Z_i)} \\
 &= \frac{\hat{E}(Y_i | Z_i = 1) - \hat{E}(Y_i | Z_i = 0)}{\hat{E}(D_i | Z_i = 1) - \hat{E}(D_i | Z_i = 0)} \\
 &\text{for binary IV and treatment.} \\
 \end{aligned} \tag{3}$$

The 2SLS estimator  $\widehat{CACE}_{2SLS}$ , sometimes called the Wald estimator, is the sample analogue of Eq. 2 and consistently estimates the CACE. The asymptotic standard error for  $\widehat{CACE}_{2SLS}$  is given in Imbens and Angrist (1994), Theorem 3.

The 2SLS estimator Eq. 3 can be used when information on  $Y$ ,  $Z$ , and  $D$  are not available in a single data set, but one data set has  $Y$  and  $Z$  and the other data set has  $D$  and  $Z$ ; this is called two-sample instrumental variable estimation (Angrist and Krueger 1992; Inoue and Solon 2010). For example, Kaushal (2007) studied the effect of food stamps on body mass index (BMI) in immigrant families using differences in state responses to a change in federal laws on immigrant eligibility for the food stamp program as an IV. The National Health Interview Study was used to estimate the effect of state lived in on BMI, and the Current Population Survey was used to estimate the effect of state lived in on food stamp program participation because neither data set contained all three variables.

**More Efficient Estimation**

Let  $\mu^{c1} = E(Y_i^1 | C_i = co)$ ,  $\mu^{c0} = E(Y_i^0 | C_i = co)$ ,  $\mu^a = E(Y_i | C_i = at)$ , and  $\mu^n = E(Y_i | C_i = nt)$ ,

and  $\pi_a$ ,  $\pi_c$ , and  $\pi_n$  denote the proportion of always takers, compliers, and never takers, respectively. Note that by Assumptions 1–5 and the mixture structure of the outcomes of the four observed groups shown in Table 1,

$$\begin{aligned}
 E(Y | Z_i = 1, D_i = 1) &= \frac{\pi_c}{\pi_c + \pi_a} \mu^{c1} + \frac{\pi_a}{\pi_c + \pi_a} \mu^a \\
 E(Y | Z_i = 1, D_i = 0) &= \mu^n \\
 \end{aligned} \tag{4}$$

$$\begin{aligned}
 E(Y | Z_i = 0, D_i = 0) &= \frac{\pi_c}{\pi_c + \pi_n} \mu^{c0} + \frac{\pi_n}{\pi_c + \pi_n} \mu^n \\
 E(Y | Z_i = 0, D_i = 1) &= \mu^a \\
 \end{aligned} \tag{5}$$

where the quantities on the left-hand side are expectations of observed outcomes and on the right-hand side are functions of expected potential outcomes and proportions for latent compliance classes. The 2SLS or standard IV estimator is to use the data in the  $(Z_i = 1, D_i = 0)$  group to get  $\hat{\mu}^n$  and then plug it into Eq. 5 to get  $\hat{\mu}^{c0}$  and use the data in the  $(Z_i = 0, D_i = 1)$  group for  $\hat{\mu}^a$  and then plug it into Eq. 4 to get  $\hat{\mu}^{c1}$ . However, the data information in the mixture groups  $(Z_i = 1, D_i = 1)$  and  $(Z_i = 0, D_i = 0)$  is not used in the 2SLS estimator Eq. 3 even though it can be useful for estimating the average potential outcomes. Similarly the 2SLS estimator uses only the information in the treatment group  $(Z_i = 1)$  to estimate  $\pi_n$  and only the information in the control group  $(Z_i = 0)$  to estimate  $\pi_a$ , but the mixture structure (see Table 1) implies that there is additional information in the control group for estimating  $\pi_n$  and additional information in the treatment group for estimating  $\pi_a$ .

Imbens and Rubin (1997a, 1997b) proposed two approaches using mixture modeling to estimate the CACE. One approach assumes a parametric distribution (normal) for the outcomes and then estimates the CACE by maximum likelihood using the EM algorithm. This estimator provides considerable efficiency gains over the 2SLS estimator when the parametric assumptions hold. However, when the parametric assumptions are wrong, this estimator can be inconsistent, whereas the 2SLS estimator is consistent; see Table 4 of (Cheng

et al. 2009b) for finite-sample results. Imbens and Rubin’s other approach to using mixture modeling to estimate the CACE is to approximate the density of the outcome distribution for each compliance class under each randomization group as a piecewise constant function and then estimate the CACE by maximum likelihood (Imbens and Rubin 1997a). This approach is in principle nonparametric as the number of constant pieces in each density function can be increased with the sample size. However, Imbens and Rubin (1997a) do not provide a systematic approach for choosing the number of and locations of the pieces.

To take into account the mixture structure in outcome distribution, Cheng et al. (2009b) developed a systematic and easily implementable approach for inference about the CACE using empirical likelihood (Owen 2002). Empirical likelihood profiles a general multinomial likelihood with support on the observed data points and therefore is an easily constructed random approximation to unknown distributions. Maximum empirical likelihood estimators have good properties. The maximum empirical likelihood estimator for the CACE is robust to parametric distribution assumptions since the empirical likelihood for a parameter such as the CACE is the nonparametric profile likelihood for the parameter.

To explain the methodology of Cheng et al. (2009b), consider a single consent randomized encouragement trial as an example. A single consent trial is a trial in which the group that does not receive encouragement to take the treatment has no access to the treatment so that the set of always takers and defiers is empty in the trial (Zelen 1979). Let the first  $n_0$  subjects be from the not encouraged group and the next  $N - n_0$  subjects be from the encouraged group. Then the empirical likelihood  $L_E$  of the parameters  $(\pi_c, \mu^n, \mu^{c^1}, \mu^{c^0})$  is

$$L_E(\pi_c, \mu^n, \mu^{c^1}, \mu^{c^0}) = \max \left( \prod_{i=1}^{n_0} q_i \right) \left( \prod_{i=n_0+1}^N q_i \right),$$

subject to

$$\begin{aligned} \sum_{i=1}^{n_0} q_i &= 1, \quad \sum_{i=n_0+1}^N q_i = 1, \quad q_i \geq 0, \quad i = 1, \dots, N, \\ \sum_{i=n_0+1}^N q_i D_i &= \pi_c, \quad \sum_{i=n_0+1}^N q_i Y_i D_i = \mu^{c^1} \pi_c, \\ \sum_{i=n_0+1}^N q_i Y_i (1 - D_i) &= \mu^n (1 - \pi_c), \end{aligned}$$

There exist  $p_i^{c^0}, p_i^n, i = 1, \dots, n_0$  such that

$$\begin{aligned} \pi_c p_i^{c^0} + (1 - \pi_c) p_i^n &= q_i, \\ \sum_{i=1}^{n_0} p_i^{c^0} = \sum_{i=1}^{n_0} p_i^n &= 1, \quad p_i^{c^0}, p_i^n \geq 0, \quad i = 1, \dots, n_0, \\ \sum_{i=1}^{n_0} p_i^n (Y_i - \mu^n) &= 0, \\ \sum_{i=1}^{n_0} p_i^{c^0} (Y_i - \mu^{c^0}) &= 0. \end{aligned}$$

where  $p_i^{c^0}$  and  $p_i^n$  are the population probabilities that a randomly chosen complier assigned to the no encouragement group and a randomly chosen never taker assigned to the no encouragement group have the same outcome as subject  $i$ , respectively.

By maximizing the empirical likelihood with the EM algorithm as described in Cheng et al. (2009b), the maximum empirical likelihood estimator for the CACE is obtained. Cheng et al. (2009b) show that the estimator provides substantial efficiency gains over the 2SLS estimator in finite samples. Cheng et al. (2009b) also extend their methodology to general encouragement trials in which there are always takers.

In addition to the inference on CACE, Cheng et al. (2009a) developed a semiparametric IV method based on the empirical likelihood approach for distributional treatment effects for compliers and other general functions of the compliers’ outcome distribution. They showed that their estimators are substantially more efficient than the standard IV estimator for treatment effects on outcome distributions (see section “Effect of Treatment on Distribution of Outcomes” for more details).

## Estimation with Observed Covariates

As discussed above, various methods have been proposed to use IVs to overcome the problem of selection bias in estimating the effect of a treatment on outcomes without covariates. However, in practice, instruments may be valid only after conditioning on covariates. For example, in the NICU study of section “[Instrumental Variables: NICU Example Revisited](#),” race is associated with the proposed IV excess travel time and race is also thought to be associated with infant mortality through mechanisms other than level of NICU delivery such as maternal age, previous Caesarean section, inadequate prenatal care, and chronic medical conditions (Lorch et al. 2012b). Consequently, in order for excess travel time to be independent of unmeasured confounders conditional on measured covariates, it is important that race be included as a measured covariate. To incorporate covariates into the two-stage least squares estimator, regress  $D_i$  on  $\mathbf{X}_i$  and  $Z_i$  in the first stage to obtain  $\hat{D}_i$  and then regress  $Y_i$  on  $\hat{D}_i$  and  $\mathbf{X}_i$  in the second stage. Denote the coefficient on  $\hat{D}_i$  in the second-stage regression by  $\hat{\lambda}^{2SLS}$ . The estimator  $\hat{\lambda}^{2SLS}$  estimates some kind of covariate-averaged CACE as we shall discuss (Angrist and Pischke 2009). Let  $(\lambda, \phi)$  be the minimum mean squared error linear approximation to the average response function for compliers  $E(Y|\mathbf{X}, D, C = co)$ , that is,  $(\lambda, \phi) = \arg \min_{\lambda^*, \phi^*} E[(Y - \phi^{*T}\mathbf{X} - \lambda^*D)^2 | C = co]$  (where  $\mathbf{X}$  is assumed to contain the intercept). Specifically, if the complier average causal effect given  $\mathbf{X}$  is the same for all  $\mathbf{X}$  and the effect of  $\mathbf{X}$  on the outcomes for compliers is linear (i.e.,  $E(Y|\mathbf{X}, D, C = co) = \phi^{*T}\mathbf{X} + \lambda^*D$ ), then  $\lambda$  equals the CACE. The estimator  $\hat{\lambda}^{2SLS}$  is a consistent (i.e., asymptotically unbiased) estimator of  $\lambda$ . Thus, if the complier average causal effect given  $\mathbf{X}$  is the same for all  $\mathbf{X}$  and the effect of  $\mathbf{X}$  on the outcomes for compliers is linear,  $\hat{\lambda}^{2SLS}$  is a consistent estimator of the CACE. The standard error for  $\hat{\lambda}^{2SLS}$  is not the standard error from the second-stage regression but needs to account for

the sampling uncertainty in using  $\hat{D}_i$  as an estimate of  $E(D_i|\mathbf{X}_i, Z_i)$  (see White 1984; Davidson and MacKinnon 1993; Freedman 2009, Chap. 9.8). Other methods besides two-stage least squares for incorporating measured covariates into the IV model are discussed in Little and Yau (1998), Hirano et al. (2000), Angrist and Imbens (1995), Abadie (2003) Tan (2006), O’Malley et al. (2011), Cheng et al. (2009b), and Okui et al. (2012), among others. Little and Yau (1998) and Hirano et al. (2000) introduce covariates in the IV model of Imbens and Angrist (1994) with distributional assumptions and functional form restrictions. Angrist and Imbens (1995) consider settings under fully saturated specifications with discrete covariates. Without distributional assumptions or functional form restrictions, Abadie (2003) develops closed forms for average potential outcomes for compliers under treatment and control with covariates. Cheng et al. (2009b) discuss incorporating covariates with the empirical likelihood approach of section “[More Efficient Estimation](#).”

---

## Understanding the Treatment Effect That IV Estimates

### Relationship Between Average Treatment Effect for Compliers and Average Treatment Effect for the Whole Population

As discussed in section “[TV Assumptions and Estimation for Binary IV and Binary Treatment](#),” the IV method estimates the CACE, the average treatment effect for the compliers ( $E[Y^1 - Y^0 | C = co]$ ). The average treatment effect in the population is, under the monotonicity assumption, a weighted average of the average treatment effect for the compliers, the average treatment effect for the never takers, and the average treatment effect for the always takers:

$$\begin{aligned}
 E[Y^1 - Y^0] &= P(C = co)E[Y^1 - Y^0 | C = co] \\
 &\quad + P(C = at)E[Y^1 - Y^0 | C = at] \\
 &\quad + P(C = nt)E[Y^1 - Y^0 | C = nt].
 \end{aligned}$$

The IV method provides no direct information on the average treatment effect for always takers ( $E[Y^1 - Y^0 | C = at]$ ) or the average treatment effect for never takers ( $E[Y^1 - Y^0 | C = nt]$ ). However, the IV method can provide useful bounds on the average treatment effect for the whole population if a researcher is able to put bounds on the difference between the average treatment effect for compliers and the average treatment effects for never takers and always takers based on subject matter knowledge. For example, suppose a researcher is willing to assume that this difference is no more than  $b$ , then

$$\begin{aligned}
 &E[Y^1 - Y^0 | C = co] - b[1 - P(C = co)] \\
 &\leq E[Y^1 - Y^0] \leq E[Y^1 - Y^0 | C = co] \\
 &\quad + b[1 - P(C = co)],
 \end{aligned} \tag{6}$$

where the quantities on the left and right-hand sides of Eq. 6 other than  $b$  can be estimated as discussed in section “[IV Assumptions and Estimation for Binary IV and Binary Treatment](#).” For binary or other bounded outcomes, the boundedness of the outcomes can be used to tighten bounds on the average treatment effect for the whole population or other treatment effects (Balke and Pearl 1997; Cheng and Small 2006). Qualitative assumptions, such as that the average treatment effect is larger for always takers than compliers, can also be used to tighten the bounds (e.g., Cheng and Small 2006; Bhattacharya et al. 2008; Siddique 2009).

## Characterizing the Compliers

The IV method estimates the average treatment effect for the subpopulation of compliers. Who are these compliers and how do they compare to noncompliers? To understand this better, it is useful to characterize the compliers in terms of their distribution of observed covariates (Angrist and

Pischke 2009; Brookhart and Schneeweiss 2007). The mean of a covariate  $X_i$  among the compliers is

$$E[X_i | C = co] = \frac{E[\kappa_i X_i]}{E[\kappa_i]}, \tag{7}$$

where

$$\kappa_i = 1 - \frac{D_i(1 - Z_i)}{1 - P(Z_i = 1 | X_i)} - \frac{(1 - D_i)Z_i}{P(Z_i = 1 | X_i)}$$

(Abadie 2003). The prevalence ratio of a binary characteristic  $X$  among compliers compared to the full population is  $P(X = 1 | C = co) / P(X = 1)$ . Table 2 shows the mean of various characteristics  $X$  among compliers versus the full population and also shows the prevalence ratio (where the sample estimates of  $P(Z_i = 1 | X_i)$ ,  $E[\kappa_i X_i]$  and  $E[\kappa_i]$  are plugged into Eq. 7). Babies whose mothers are college graduates are slightly underrepresented (prevalence ratio = 0.87), and African-Americans are slightly overrepresented (prevalence ratio = 1.14) among compliers. Very low birthweight (<1500 g) and very premature babies (gestational age  $\leq 32$  weeks) are substantially underrepresented among compliers, with prevalence ratios around one-third; these babies are more likely to be always takers, that is, delivered at high-level NICUs regardless of mother’s travel time. Babies whose mothers’ have comorbidities such as diabetes or hypertension are slightly underrepresented among compliers. Overall, Table 2 suggests that higher risk babies are underrepresented among the compliers. If the effect of high-level NICUs is greater for higher risk babies, then the IV estimate will underestimate the average effect of high-level NICUs for the whole population.

## Understanding the IV Estimate When Compliance Status Is Not Deterministic

For an encouragement that is uniformly delivered, such as patients who made an appointment at a psychiatric outpatient clinic are sent a letter encouraging them to attend the appointment (Kitcheman et al. 2008), it is clear that a subject is either a complier, always taker, never taker, or

**Table 2** Complier characteristics for NICU study. The second column shows the estimated proportion of compliers with a characteristic  $X$ , the third column shows the estimated proportion of the full population with the

characteristic  $X$ , and the fourth column shows the estimated ratio of compliers with  $X$  compared to the full population with  $X$

Characteristic $X$	Prevalence of $X$ among compliers	Prevalence of $X$ in full population	Prevalence ratio of $X$ among compliers to full population
Mother College Graduate	0.23	0.26	0.87
African-American	0.17	0.15	1.14
Birthweight < 1,500 g	0.03	0.09	0.33
Gestational age $\leq$ 32 weeks	0.04	0.13	0.34
Gestational diabetes	0.05	0.05	0.91
Diabetes mellitus	0.02	0.02	0.77
Pregnancy-induced hypertension	0.08	0.10	0.82
Chronic hypertension	0.02	0.02	0.89

defier with respect to the encouragement. However, sometimes encouragements that are not uniformly delivered are used as IVs. For example, in the NICU study, consider the IV of whether the mother's excess travel time to the nearest high-level NICU is more than 10 min. If a mother whose excess travel time to the nearest high-level NICU was more than 10 min moved to a new home with an excess travel time less than 10 min, whether the mother would deliver her baby at a high-level NICU might depend on additional aspects of the move, such as the location and availability of public transportation at her new home (Joffe 2011) and the exact travel time to the nearest high-level NICU at her new home. Consequently, a mother may not be able to be deterministically classified as a complier or not a complier – she may be a complier with respect to certain moves but not others. Another example of nondeterministic compliance is that when physician preference for one drug versus another is used as the IV (e.g.,  $Z = 1$  if a patient's physician prescribes drug  $A$  more often drug  $B$ ), whether a patient receives drug  $A$  may depend on how strongly the physician prefers drug  $A$  (Brookhart and Schneeweiss 2007; Hernán and Robins 2006). Another situation in which nondeterministic compliance status can arise is that the IV may not itself be an encouragement intervention but a proxy for an encouragement

intervention. Consider the case of Mendelian randomization, in which the IV is often a single nucleotide polymorphism (SNP) that might be part of a gene  $A$ . The SNP may be a marker for a gene  $B$  on the same chromosome that actually affects the level of the exposure  $D$ . The encouragement intervention is receiving the gene  $B$  that actually affects the level of the exposure  $D$ , and the SNP is just a proxy for this encouragement. Consequently, even if a subject's exposure level would change as a result of a change in gene  $B$ , whether the subject is a complier with respect to a change in the SNP depends on whether the change in the SNP leads to a change in the gene  $B$ , which is randomly determined through the process of recombination (Joffe 2011).

Brookhart and Schneeweiss (2007) provide a framework for understanding how to interpret the IV estimate when compliance status is not deterministic. Suppose that the study population can be decomposed into a set of  $\kappa + 1$  mutually exclusive groups of patients based on clinical, lifestyle, and other characteristics such that within each group of patients, whether a subject receives treatment is independent of the effect of the treatment. All of the common causes of the potential treatment received  $D^1$ ,  $D^0$ , and the potential outcomes  $Y^1$ ,  $Y^0$  should be included in the characteristics used to define these groups. For example, if there are  $L$  binary common causes of  $(D^1, D^0, Y^1, Y^0)$ ,

then the subgroups can be the  $\kappa + 1 = 2^L$  possible values of these common causes. Denote patient membership in these groups by the set of indicators  $\mathbf{S} = \{S_1, S_2, \dots, S_\kappa\}$ . Consider the following model for the expected potential outcome:

$$E(Y^d | \mathbf{S}) = \alpha_0 + \alpha_1 d + \alpha_2^T \mathbf{S} + \alpha_3^T \mathbf{S} d$$

The average effect of treatment in the population is  $\alpha_1 + \alpha_3^T E[\mathbf{S}]$ , and the average effect of treatment in subgroup  $j$  is  $\alpha_1 + \alpha_{3,j}$ . Under the IV assumptions 1–3 and 5 in section “[Framework and Notation](#),” that is, all the assumptions except monotonicity, the IV estimator estimates the following quantity:

$$\begin{aligned} & \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)} \\ &= \alpha_1 + \sum_{j=1}^{\kappa} \alpha_{3,j} E[S_j] w_j, \end{aligned} \quad (8)$$

where

$$w_j = \frac{E(D|Z=1, S_j=1) - E(D|Z=0, S_j=1)}{E(D|Z=1) - E(D|Z=0)}.$$

The IV estimator Eq. 8 is a “weighted average” of treatment effects in different subgroups, where the subgroups in which the instrument has a stronger effect on the treatment get more weight. Note that when the compliance class is deterministic, then the subgroups can be defined as the compliance classes and Eq. 8 just says that the IV estimator is the average treatment effect for compliers. In the NICU study, where compliance class may not be deterministic, Table 2 suggests that babies in lower-risk groups, for example, not very low birthweight or not very low gestational age, are weighted more heavily in the IV estimator. If there are subgroups for whom the instrument has no effect on their treatment level, then that subgroup gets zero weight. For example, mothers or babies with severe preexisting conditions may virtually always be delivered at a high-level NICU, so that the IV of excess travel time has no effect on their treatment level

(Lorch et al. 2012a). If there are subgroups for whom the encouraging level of the instrument makes them less likely to receive the treatment, then this subgroup would get “negative weight” and Eq. 8 is not a true weighted average, potentially leading the IV estimator to have the opposite sign of the effect of the treatment. For example, Brookhart and Schneeweiss (2007) discussed studying the safety of metformin for treating type II diabetes versus other antihyperglycemic drugs among patients with liver disease using physician preference as the IV ( $Z = 1$  if a physician is more likely to prescribe metformin than other antihyperglycemic drugs). Metformin is contraindicated in patients with decreased liver disease, as it can cause lactic acidosis, a potentially fatal side effect. Brookhart and Schneeweiss (2007) speculated that physicians who infrequently use metformin will be less likely to understand its contraindications and would therefore be more likely to misuse it. If this hypothesis is true, then for estimating the effect of metformin on lactic acidosis, the IV estimator could mistakenly make metformin appear to prevent lactic acidosis, as patients of physicians with  $Z = 1$  are at lower risk of being inappropriately treated with metformin. When the compliance class is deterministic, a subgroup getting negative weight means that there are defiers, violating the monotonicity assumption.

---

## Assessing the IV Assumptions and Sensitivity Analysis for Violations of Assumptions

### Assessing the IV Assumptions

This section will discuss assessing the two key IV assumptions: (1) the IV is independent of unmeasured confounders; (2) the IV affects outcome only through treatment received (the exclusion restriction).

One way of assessing whether the proposed IV is independent of unmeasured confounders conditional on measured confounders is to look at whether the proposed IV is associated with measured confounders. Although measured confounders can be controlled for, if the measured



confounder is only a proxy for the true confounder, then an association between the proposed IV and the measured confounder suggests that there will be an association between the IV and the unmeasured part of the true confounder. If there are two or more sources of confounding, then it is useful to examine if the observable part of one source of confounding is associated with the IV after controlling for the other sources of confounding. These ideas will be illustrated using the NICU study described in section “[Instrumental Variables: NICU Example Revisited](#).” Table 3 shows the imbalance of measured covariates across levels of the IV. The racial composition is very different between the near ( $Z = 1$ ) and far ( $Z = 0$ ) babies, with near babies being much more likely to be African-American. Since race has a substantial association with neonatal outcomes

(Demissie et al. 2001; Lorch et al. 2012b), it is sensible to examine the association of other measured confounders with the IV after controlling for race. Table 4 shows the association of the IV with measured confounders for whites. The clinical measured confounders such as low birthweight, gestational age  $\leq 32$  weeks, and maternal comorbidities (diabetes and hypertension) are generally similar between near and far babies although there are some significant associations. This similarity between the clinical status of near and far babies and mothers after controlling for race provides some support that the IV is approximately, although not exactly, valid for whites. However, whether the mother is a college graduate differs substantially between white near and far mothers, suggesting that there may be residual confounding due to

**Table 3** Imbalance of measured covariates across levels of the instrument for the NICU data. The prevalence difference ratio is the ratio of the imbalance of the measured covariates across levels of the instrument to the imbalance across levels of the treatment. The estimated proportion of

compliers is  $P(D = 1|Z = 1) - P(D = 1|Z = 0) = 0.447$  so that a prevalence difference ratio less than 0.447 for an  $X$  indicates that there would less bias in the IV method from failing to adjust for  $X$  than from ordinary least squares that failed to adjust for  $X$

Characteristic $X$	$P(X \text{near})$ (%)	$P(X \text{far})$ (%)	$p$ -value	Prevalence difference ratio
Birthweight < 1,500 g	9.4	7.7	<0.01	0.02
Mother College Graduate	25.9	26.1	0.26	-0.04
African-American	25.6	4.6	<0.01	0.64
Gestational age $\leq 32$ weeks	14.3	11.7	<0.01	0.23
Gestational diabetes	5.2	5.2	0.47	0.12
Diabetes mellitus	1.8	1.9	0.07	-0.16
Pregnancy-induced hypertension	10.6	10.1	<0.01	0.13
Chronic hypertension	1.9	1.3	<0.01	0.61

**Table 4** Imbalance of measured covariates across levels of the instrument for babies born to white mothers in the NICU data. The prevalence difference ratio is the ratio of the imbalance of the measured covariates across levels of the instrument to the imbalance across levels of the treatment. The estimated proportion of compliers is  $P(D = 1|$

$Z = 1, \text{white}) - P(D = 1|Z = 0, \text{white}) = 0.418$  so that a prevalence difference ratio less than 0.418 for an  $X$  indicates that there would less bias in the IV method from failing to adjust for  $X$  than from ordinary least squares that failed to adjust for  $X$

Characteristic $X$	$P(X \text{near})$ (%)	$P(X \text{far})$ (%)	$p$ -value	Prevalence difference ratio
Birthweight < 1,500 g	7.5	7.2	0.07	0.04
Mother College Graduate	34.4	26.8	<0.01	0.72
Gestational age $\leq 32$ weeks	11.8	11.1	<0.01	0.07
Gestational diabetes	5.6	5.3	0.02	0.34
Diabetes mellitus	1.8	1.9	0.08	-0.17
Pregnancy-induced hypertension	10.6	10.1	<0.01	0.05
Chronic hypertension	1.6	1.3	<0.01	0.43

**Table 5** Imbalance of measured covariates across levels of the instrument for babies born to African-American mothers in the NICU data. The prevalence difference ratio is the ratio of the imbalance of the measured covariates across levels of the instrument to the imbalance across levels of the treatment. The estimated proportion of

Characteristic $X$	$P(X \text{near})$ (%)	$P(X \text{far})$ (%)	$p$ -value	Prevalence difference ratio
Birthweight < 1,500 g	13.5	11.9	<0.01	0.41
Mother College Graduate	8.0	10.7	<0.01	1.60
Gestational age $\leq 32$ weeks	19.3	16.6	<0.01	0.48
Gestational diabetes	4.2	4.3	0.67	-0.70
Diabetes mellitus	1.9	2.6	<0.01	-1.35
Pregnancy-induced hypertension	11.8	10.0	<0.01	0.69
Chronic hypertension	2.8	2.4	0.12	0.34

compliers is  $P(D = 1|Z = 1, \text{African-American}) - P(D = 1|Z = 0, \text{African-American}) = 0.503$  so that a prevalence difference ratio less than 0.503 for an  $X$  indicates that there would less bias in the IV method from failing to adjust for  $X$  than from ordinary least squares that failed to adjust for  $X$

socioeconomic status. Table 5 shows the association of the IV with measured confounders for African-Americans. For African-Americans, there are more substantial associations than for whites between near/far status and the important clinical status variables low birthweight and gestational age  $\leq 32$  weeks, raising more concern about whether the IV is approximately valid for African-Americans.

The last column of Tables 3, 4, and 5 shows the prevalence difference ratio, a measure of how biased an IV analysis would be from failing to adjust from the confounder as compared to an ordinary least squares analysis (Brookhart and Schneeweiss 2007). The below discussion of the prevalence difference ratio is drawn from Brookhart and Schneeweiss (2007). Denote the confounder by  $U$ . Consider the following model for the potential outcome:

$$Y^d = \alpha_0 + \alpha_1 d + \alpha_2 U + \epsilon_d, \tag{9}$$

where  $E(\epsilon_d|U) = 0$ . The average treatment effect is  $E[Y^1 - Y^0] = \alpha_1$ . The observed data is

$$Y = \alpha_0 + \alpha_1 D + \alpha_2 U + \epsilon_0 + D(\epsilon_1 - \epsilon_0).$$

Assume that  $E(\epsilon_d|D, U) = 0$  for  $d = 0$  or  $1$ . This assumption means that if  $U$  were controlled for, the parameters of Eq. 9 could be consistently estimated by least squares. By iterated expectations,  $E[\epsilon_0 + D(\epsilon_1 - \epsilon_0)|D] = 0$ .

Therefore,

$$E(Y|D = 1) - E(Y|D = 0) = \alpha_1 + \alpha_2(E[U|D = 1] - E[U|D = 0]),$$

so that an ordinary least squares analysis that did not adjust for  $U$  would be biased by  $\alpha_2(E[U|D = 1] - E[U|D = 0])$ . To evaluate the IV estimand, consider the further assumption that  $E[\epsilon_0|Z] = 0$  so that the proposed IV can be related to the observed outcome only through its effect on  $D$  or association with  $U$ ; also assume that  $E(\epsilon_1 - \epsilon_0|C)$  is the same for all compliance classes  $C$  so that the complier average causal effect is equal to the overall average causal effect  $\alpha_1$ . These assumptions together say that if  $U$  were controlled for, the IV estimator would consistently estimate the average treatment effect  $\alpha_1$ . Under these assumptions, the probability limit of the IV estimator that does not control for  $U$  can be written as

$$\begin{aligned} & \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]} \\ &= \alpha_1 + \alpha_2 \frac{E(U|Z = 1) - E(U|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)}. \end{aligned}$$

The asymptotic bias of the IV estimator is thus

$$\text{Bias}(\hat{\beta}_1^{IV}) = \alpha_2 \frac{E(U|Z = 1) - E(U|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)}. \tag{10}$$

The term  $E(U|Z = 1) - E(U|Z = 0)$  is the difference in the prevalence of the risk factor  $U$  between levels of the IV. The total bias in the IV estimator is this difference multiplied by the excess risk of the outcome among patients with  $U = 1$  divided by the strength of the IV. For the IV estimator to have less asymptotic bias than ordinary least squares (OLS), the following condition must hold (Brookhart and Schneeweiss 2007)

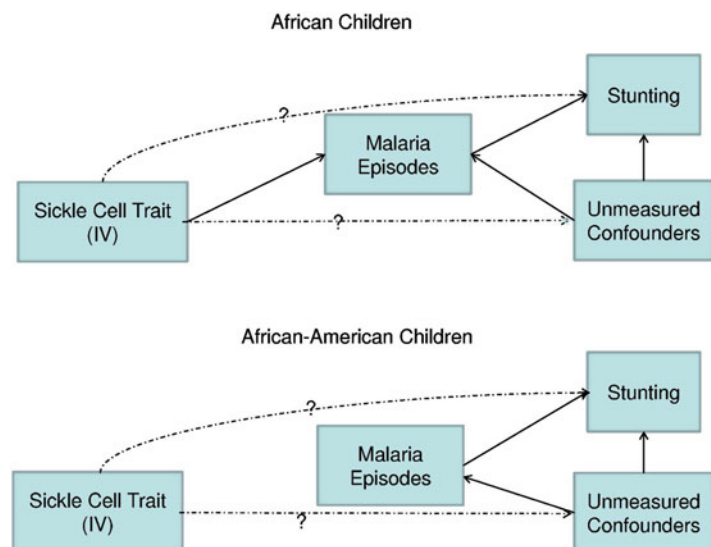
$$\frac{E[U|Z = 1] - E[U|Z = 0]}{E[D|Z = 1] - E[D|Z = 0]} < E[D|Z = 1] - E[D|Z = 0]. \quad (11)$$

In other words, the difference in the prevalence of  $U$  between levels of  $Z$  relative to the difference in the prevalence of  $U$  between levels of  $D$  must be less than the strength of the IV (Brookhart and Schneeweiss 2007). The left-hand side of Eq. 11 is called the prevalence difference ratio (PDR). In order for us to think that the IV analysis is likely to be less biased than OLS, the PDR should be less than the strength of the IV ( $E[D|Z = 1] - E[D|Z = 0]$ ), particularly for those variables clearly related to the outcome. Table 4 shows that the PDRs are generally less than the strength of the IV (0.418) for whites, but the PDRs are often greater than the strength of the IV (0.503) for African-Americans, suggesting that the IV

analysis reduces bias for whites compared to OLS but not for African-Americans.

A way of testing whether the two key IV assumptions (i.e., (i) the IV is independent of unmeasured confounders conditional on the measured confounders and (ii) the IV affects outcomes only through treatment received) hold is to find a subpopulation for whom the link between the IV and treatment received is thought to be broken and then test whether the IV is associated with the outcome in this subpopulation. The only way in which the IV could be associated with the outcome in such a subpopulation is if the IV was associated with unmeasured confounders or directly affected the outcome through a pathway other than treatment received. Figure 2 shows an example. Kang et al. (2013) study the effect of children in Africa getting malaria on their becoming stunted (having a height that is two standard deviations below the expected height for the child’s age) and consider the sickle cell trait as a possible IV. The sickle cell trait is that a person inherits a copy of the hemoglobin variant HbS from one parent and normal hemoglobin from the other. While inheriting two copies of HbS results in sickle cell disease and substantially shortened life expectancy, inheriting only one copy (the sickle cell trait) is protective against malaria and is thought to have little detrimental effect on health (Aidoo et al. 2002). To test

**Fig. 2** Causal diagrams for the effect of the sickle cell trait (the IV) and malaria episodes (the treatment) on stunting (the outcome) in African children and African-American children. If the sickle cell trait is a valid IV, then the dashed lines should be absent and the sickle cell trait will have no effect on stunting among African-American children



whether the sickle cell trait indeed does not affect stunting in ways other than reducing malaria and is not associated with unmeasured confounders, Kang et al. (2013) considered whether the sickle cell trait is associated with stunting among African-American children; the sickle cell trait has high prevalence among African-Americans but does not affect malaria because malaria is not present in the United States. Rehan (1981) and Kramer et al. (1978) found no evidence that sickle cell trait is associated with growth and development in African-American children. This provides evidence that the dashed lines in Fig. 2 are indeed absent, which would mean that the proposed IV of the sickle cell trait does indeed satisfy the two key IV assumptions of being independent of unmeasured confounders and affecting outcomes only through treatment received. Angrist and Krueger (1991) also employed this strategy of finding a subpopulation for whom the link between the IV and treatment received is broken to test their IV of quarter of birth for studying the effect of education on earnings. The reason that quarter of birth is associated with education is that for students who plan to drop out of school as soon as they have reached the age at which they are no longer compelled to be in school (e.g., age 16), quarter of birth affects how much education these students will get before they drop out because children start school at different ages depending on their quarter of birth. However, for students who plan to go to college, quarter of birth does not affect their amount of schooling. Consequently, Angrist and Krueger (1991) looked at whether there was an absence of an association between quarter of birth and earnings among students who went to college to test the IV assumptions.

Newcomers to IV methods often think that the validity of the IV can be tested by regressing the outcome on treatment received, the IV and measured confounders, and testing whether the coefficient on the IV is significant. However, this is not a valid test as even if the IV assumptions hold, the coefficient on the IV would typically be nonzero. One way to see this is that if there are no measured confounders, the test amounts to testing whether (i)  $E[Y|Z = 1, D = 1] - E[Y|Z = 0, D = 1] = 0$  and

(ii)  $E[Y|Z = 1, D = 0] - E[Y|Z = 0, D = 0] = 0$ . These are the differences between (i) the average potential outcome of the group of always takers and compliers together when these subjects are encouraged to receive treatment and receive treatment versus those of always takers alone when they are not encouraged to receive treatment but do receive treatment and (ii) the average potential outcome of never takers when encouraged to receive treatment but do not receive treatment versus those of the group of never takers and compliers when they are not encouraged to receive treatment and do not receive treatment. If the IV assumptions hold that the IV is not associated with unmeasured confounders and has no direct effect on the outcome other than treatment received, then (i) is equal to zero if and only if the average potential outcome of compliers and always takers are the same when both groups receive treatment and (ii) is equal to zero if and only if the average potential outcomes of compliers and never takers are the same when both groups do not receive treatment. Typically, the average potential outcome of compliers and always takers (compliers and never takers) will not be the same when both groups receive (do not receive) treatment even if the IV assumptions hold.

## Sensitivity Analysis

A sensitivity analysis seeks to quantify how sensitive conclusions from an IV analysis are to plausible violations of key assumptions. Sensitivity analysis methods for IV analyses have been developed by Angrist et al. (1996), Brookhart and Schneeweiss (2007), Small (2007), Small and Rosenbaum (2008), and Baiocchi et al. (2010), among others. Here an approach will be presented to sensitivity analysis for violations of the assumption that the IV is independent of unmeasured confounders. Assume that the concern is that the IV may be related to an unmeasured confounder  $U$  which has mean 0 and variance 1 and is independent of the measured confounders  $\mathbf{X}$  ( $U$  can always be taken to be the residual of the unmeasured confounder given

the measured confounders to make this assumption plausible). Consider the following model:

$$\begin{aligned}
 Y_i^d &= \alpha + \beta d + \gamma^T \mathbf{X}_i + \delta U_i + e_i \\
 U_i &= \rho + \eta Z_i + v_i \\
 E(v_i | \mathbf{X}_i, Z_i) &= 0, E(e_i | \mathbf{X}_i, Z_i) = 0.
 \end{aligned}
 \tag{12}$$

$\beta$  is the causal effect of increasing  $D$  by one unit. The sensitivity parameters are  $\delta$ , the effect of a one standard deviation increase in the unmeasured confounder on the mean of the potential outcome under no treatment, and  $\eta$ , how much higher the mean of the unmeasured confounder  $U_i$  is in standard deviation units for  $Z_i = 1$  versus  $Z_i = 0$ . Model (12) says that  $Z_i$  would be a valid IV if both the measured confounders  $\mathbf{X}_i$  and the unmeasured confounder  $U_i$  were controlled for. Under model (12), the following holds

$$\begin{aligned}
 Y_i &= \alpha + \beta D_i + \gamma^T \mathbf{X}_i + \delta U_i + e_i \\
 Y_i - \delta \eta Z_i &= \alpha + \delta \rho + \beta D_i + \gamma^T \mathbf{X}_i + e_i + \delta v_i \\
 E(v_i | \mathbf{X}_i, Z_i) &= 0, E(e_i | \mathbf{X}_i, Z_i) = 0.
 \end{aligned}$$

Consequently, a consistent estimate of and inferences for  $\beta$  can be obtained by carrying out a two-stage least squares analysis with  $Y_i - \delta \eta Z_i$  as the outcome variable,  $D_i$  as the treatment variable,  $\mathbf{X}_i$  as the measured confounders, and  $Z_i$  as the IV. Table 6 shows a sensitivity analysis for the NICU study. If there was an unmeasured confounder  $U$  that decreased the death rate by 0.1% for a one standard deviation increase in  $U$  and was 0.5 standard deviations higher on average in subjects with  $Z = 1$  versus  $Z = 0$ , then there would still

be strong evidence that high-level NICUs reduce mortality substantially (lower end of 95% CI: 0.14% reduction). However, if there was an unmeasured confounder  $U$  that decreased the death rate by 0.5% for a one standard deviation increase in  $U$  and was 0.5 standard deviations higher in subjects with  $Z = 1$  versus  $Z = 0$ , then there would no longer be strong evidence that high-level NICUs reduce mortality substantially. It can be useful to calibrate the effect of a potential unmeasured confounder  $U$  to that of a measured confounder. For example, an increase in gestational age from 30 to 33 weeks, which is a one standard deviation increase in gestational age, is associated with a reduction in the death rate of 2.2%, and the mean gestational age is 0.093 standard deviations smaller among near ( $Z = 1$ ) versus far ( $Z = 0$ ) babies. For a comparable  $U$  that reduced the death rate by 2.2% for a one standard deviation increase in  $U$  and was 0.093 standard deviations smaller in babies with  $Z = 1$  versus  $Z = 0$ , there would still be strong evidence that high-level NICUs reduce mortality substantially (see the last row of Table 6).

A sensitivity analysis for violations of the assumption that the IV has no direct effect on the outcome can be carried out as follows. Suppose that the IV has a direct effect of  $\lambda$  but the IV is independent of unmeasured confounders, that is,

$$\begin{aligned}
 Y_i^{z,d} &= \alpha + \beta d + \gamma^T \mathbf{X}_i + \lambda z + e_i \\
 E(e_i | \mathbf{X}_i, Z_i) &= 0,
 \end{aligned}
 \tag{13}$$

Then, a consistent estimate of and inferences for  $\beta$  can be obtained by carrying out a two-stage least squares analysis with  $Y_i - \lambda Z_i$  as the outcome

**Table 6** Estimates and 95% confidence intervals for  $\beta$ , the risk difference effect of a premature baby being delivered in a high-level NICU, for different values of the sensitivity parameters  $\delta$ , the effect of a one standard

deviation increase in the unmeasured confounder on the mean of the potential outcome under no treatment, and  $\eta$ , how much higher the mean of the unmeasured confounder  $U_i$  is in standard deviation units for  $Z_i = 1$  versus  $Z_i = 0$

$\delta$	$\eta$	$\hat{\beta}$	95% CI for $\beta$
0	0	-0.0059	(-0.0091, -0.0027)
-0.001	0.5	-0.0046	(-0.0079, -0.0014)
-0.005	0.5	0.0004	(-0.0029, 0.0036)
0.001	0.5	-0.0071	(-0.0104, -0.0039)
0.005	0.5	0.0121	(-0.0154, -0.0089)
-0.022	-0.093	-0.0110	(-0.0142, -0.0078)

variable,  $D_i$  as the treatment variable,  $X_i$  as the measured confounders, and  $Z_i$  as the IV. When a proposed IV  $Z$  is thought to be independent of unmeasured confounders but there is concern that  $Z$  might have a direct effect on the outcome, Joffe et al. (2008) proposed an extended instrumental variable strategy for obtaining an unbiased estimate of the causal effect of treatment that requires having a covariate  $W$  which interacts with  $Z$  in affecting treatment but for which the direct effect of  $Z$  does not depend on  $W$ . This method is described in section “[Extended Instrumental Variable Method for When Proposed IV Has a Direct Effect.](#)”

## Weak Instruments

The *strength* of an IV refers to how strongly the IV is associated with the treatment after controlling for the measured confounders  $X$ . An IV is *weak* if this association is weak. When the IV is encouragement (vs. no such encouragement) to accept a treatment, the IV is weak if the encouragement only has a slight impact on acceptance of the treatment. The strength of the IV can be measured by the proportion of compliers or the partial  $r^2$  when adding the IV to the first-stage model for the treatment after already including the measured confounders  $X$  (Bound et al. 1995; Shea 1997).

Studies that use weak IVs face three problems:

1. *High variance.* The IV method is estimating the complier average causal effect (CACE), and the only subjects that are contributing information about the CACE are the compliers. Thus, the weaker the IV is (i.e., the smaller the proportion of compliers), the larger is the variance of the IV estimate. One might think that for a sample of size  $N$ , the variance of the IV estimate would be equivalent to the variance from having a sample of  $N \times P$  ( $C = co$ ) known compliers. However, the situation is actually worse because additional variability is contributed from the always takers and never takers having different sample means in the encouraged and unencouraged groups, even though the population means are the
2. *Misleading inferences from two-stage least squares.* When the IV is weak enough, confidence intervals formed using the asymptotic standard errors for two-stage least squares, that is, Eq. 14, may be misleading. Beginning with Bound et al. (1995), it has been recognized that the most common method of inference with instrumental variables, two-stage least squares, gives highly misleading inferences when the instrument is weak even when the instrument is perfectly valid. The two-stage least squares estimate can have substantial finite sample bias toward the ordinary least squares estimate and the asymptotic variance understates the actual variance. To see this, consider including a random number as an IV (the random number is not a valid IV because it is not correlated with the treatment received). Although the random number is theoretically unrelated to the unmeasured confounding variables, it will have some chance association with the unmeasured

same. Under the assumption that the variance of the outcomes for the always takers, never takers, compliers under treatment, and compliers under control is the same  $\sigma^2$  for each group, the asymptotic variance of  $\sqrt{N}(\widehat{CACE}_{2SLS} - CACE)$ , where  $\widehat{CACE}_{2SLS}$  is the two-stage least squares estimator Eq. 3, is

$$\frac{\sigma^2 \text{Var}(Z)}{\text{Cov}(D, Z)} = \frac{\sigma^2}{[P(D = 1 | Z = 1) - P(D = 1 | Z = 0)]^2}, \quad (14)$$

(Imbens and Angrist 1994). Thus, for a sample of size  $N$ , the variance of the IV estimate is equivalent to the variance from having a sample of  $N P$  ( $C = co$ )<sup>2</sup> known compliers. For example, for a sample size of 10,000 with 20% compliers, the variance of the IV estimate is equivalent to that from a sample of 400 known compliers as could be obtained from a randomized trial of size 400 with perfect compliance. Thus, weak IVs can drastically reduce the effective sample size, resulting in high variance and potentially low power.

confounders in a sample, and thus, some confounding will get transferred to the predicted value of the treatment. This will result in some unmeasured confounding getting transferred to the second-stage estimate of the treatment effect. Stock et al. (2002) studied what strength of IV is needed to ensure that two-stage least squares provides reliable inferences. They suggested looking at the first-stage partial  $F$  statistic for testing that the coefficient on the IV(s) is zero. For one IV, if this first stage partial  $F$  statistic is less than about 10, the two-stage least squares inferences are misleading in the sense that the type I error rate of a nominal 0.05 level is actually greater than 0.15. If more than one IV is used, then the first-stage partial  $F$  statistic needs to be larger to avoid misleading inferences, greater than 12 for two IVs, greater than 16 for five IVs, and greater than 21 for ten IVs.

A number of methods have been developed that provide accurate inferences when the IV is weak. One method is to use the permutation inference developed in Imbens and Rosenbaum (2005) and illustrated in Small and Rosenbaum (2008). Another method developed by Moreira (1990) is to consider the conditional distribution of the likelihood ratio statistic, conditioning on the value of nuisance parameters. This method is implemented in a Stata program CLRv2.

3. *Highly sensitive to bias from unmeasured confounders.* Recall formula (10) for the bias in the IV estimator when the proposed IV is associated with an unmeasured confounder  $U$ . The numerator measures the association between the IV and the unmeasured confounder (multiplied by how much the unmeasured confounder affects the outcome). The denominator is the proportion of compliers and reflects the strength of the IV. Thus, when the IV is weak (i.e., the proportion of compliers is small), the effect of the IV being invalid from being associated with an unmeasured confounder is greatly exacerbated, and even a minor association between the IV and an unmeasured confounder can lead to substantial bias if the IV is weak (Bound et al. 1995; Small and Rosenbaum 2008).

In summary, when the IV is weak, the IV estimate may have high variance, and if it is weak enough (i.e., partial  $F$  statistic less than 10), it is important to use inference methods other than two-stage least squares to provide accurate inferences. These inference methods may inform us that the confidence interval for the treatment effect is very wide, but it is possible that even when the IV is weak, if the treatment effect is large enough and the sample size is big enough, there may still be a statistically significant treatment effect assuming the IV is valid. The third problem with weak IVs is that they are very sensitive to bias from being slightly invalid, that is, being slightly correlated with unmeasured confounders. This problem does not go away with a larger sample size. A slightly biased but strong IV may be preferable to a less biased but weak IV (Small and Rosenbaum 2008).

---

## Binary Outcomes

Often in health services research, the outcomes of interest take values which are not continuous and thus are not amenable to common techniques such as two-stage least squares (2SLS). In this section, methods appropriate for binary outcomes will be discussed. In the next section methods, appropriate for other noncontinuous outcomes settings will be introduced. For good general reviews of estimating IV effects in the binary outcome case, see Clarke and Windmeijer (2012), Vansteelandt et al. (2011), and Angrist (2001) (along with associated comments).

In 2SLS, one regression is run predicting the treatment, and then the estimated value of the treatment from this model is used and put into a second regression of the outcome on the covariates and the predicted treatment. This type of estimator, where the predictions from one model are substituted into a second model, is often referred to as a two-stage predictor substitution (2SPS).

When first encountering situations with binary outcomes, most analysts will recognize the regular 2SLS is problematic because it will not respect boundary conditions (i.e., the functional form imposes no constraints on parameter space,

meaning 2SLS can produce logical absurdities such as probabilities greater than one or even negative). Through analogy to 2SLS, the naive analyst may consider changing the second-stage regression to be a logistic model (or perhaps a probit) in lieu of the linear model. This would be a 2SPS. Unfortunately, in general, 2SPS models do not have the nice orthogonality properties of 2SLS and produce biased estimates (Angrist and Pischke 2009; Wooldridge 1997). Other approaches should be considered. These approaches include the parametric approaches of Hirano et al. (2000) and the semiparametric approaches of Abadie (2003), Tan (2006), and Vansteelandt et al. (2011)). Two other widely used approaches (two-stage residual inclusion and a binary probit model) and a relatively new approach (effect ratios) will be considered in detail below.

## Two-Stage Residual Inclusion

Two-stage residual inclusion (2SRI) is a two-stage regression method that is equivalent to 2SPS when the outcome is continuous but differs when the outcome is binary. Consider the non-linear model

$$E(Y|D, X, U) = M(D\beta_D + \mathbf{X}^T\beta_X + \mathbf{U}^T\beta_U) \quad (15)$$

where  $M(\cdot)$  is a known function of the treatment  $D$ , a vector of observed covariates  $\mathbf{X}$ , and a vector of unobserved covariates  $\mathbf{U}$ . The unobserved covariates  $\mathbf{U}$  are correlated with the treatment  $D$  when there is unmeasured confounding.

In a 2SPS model, the actual treatment is replaced by some predicted values, like so

$$E(Y|D, \mathbf{X}, U) = M(\hat{D}\beta_D + \mathbf{X}^T\beta_X + \mathbf{U}^T\beta_U) \quad (16)$$

where  $\hat{D}$  is estimated using the IV. This is how 2SLS is done. If the model,  $M(\cdot)$ , is linear then – speaking loosely – 2SLS makes use of the

additivity of the terms on the right-hand side of the regression to separate the endogeneity of the treatment and allow unbiased estimation of the treatment effect. If  $M(\cdot)$  is nonlinear, though, generally 2SPS will not maintain the separability of the confounding variables through the substitution method.

Another approach here is to use a two-stage residual inclusion (2SRI) model. The idea in a 2SRI is to model the unobserved covariates using the instrument, not the treatment, and thereby remove the endogeneity. The first stage in a 2SRI model is the same in that you model the treatment selection. But the difference is that in the second stage you substitute in the residuals from the first stage, not the predicted treatment. In formula this is to say:

$$E(Y|D, \mathbf{X}, U) = M(D\beta_D + \mathbf{X}^T\beta_X + \mathbf{U}^T\hat{\beta}_U) \quad (17)$$

where  $\mathbf{U}^T\hat{\beta}_U$  is estimated as the difference between the actual treatment value and the predicted treatment value from the first stage (i.e., the residual). The difference between a 2SPS and a 2SRI is what information from the first stage is used in the second stage. 2SPS and 2SRI produce the same estimates for linear models but not for nonlinear models. For an introduction to 2SRI models and how they differ from 2SPS (of which 2SLS is a special case), see Terza et al. (2008). It was shown using simulation studies in Cai et al. (2012) that for the estimation of the causal odds ratio for compliers, the 2SPS and 2SRI models performed similarly; see also Cai et al. (2011) for an analytical comparison. The simulation studies of Cai et al. (2012) also showed that the generalized structural mean model (GSMM) in an IV framework with binary outcomes tended to perform quite well vis-a-vis 2SPS and 2SRI models. See Vansteelandt et al. (2011) for an introduction to GSMM in an IV framework.

## Bivariate Probit Models

The bivariate probit model is a parameterized model that assumes an explicit functional form of the



bivariate distribution of the error terms from the selection model and the error terms from the outcome model (Bhattacharya et al. 2006; Muthen 1979). This model leans on the parametric assumptions of the error terms, leaving the conclusions sensitive to modifications of the assumptions. Additionally, these models suffer from difficulty in maximizing the likelihood functions and trouble with calculating appropriate standard errors (Freedman and Sekhon 2010).

**Matching-Based Estimator: Effect Ratio**

Coming out of a different tradition, a class of estimator has been proposed which is also capable of dealing with binary outcomes in an IV setting. Proposed in Baiocchi et al. (2010), the “effect ratio” in a binary setting can be thought of as a risk difference estimator for the compliers. The effect ratio is predicated on having matched sets. In Baiocchi et al. (2010) matched pairs were constructed using a study design-based approach called near-far matching. Near-far matching will be discussed in the next section.

First, notation will be introduced required to discuss the effect ratio. Assume there are  $I$  matched pairs,  $i = 1, \dots, I$ , with 2 subjects,  $j = 1, 2$ , one treated subject and one control, or  $2I$  subjects in total. If the  $j$ th subject in pair  $i$  receives the treatment, write  $Z_{ij} = 1$ , whereas if this subject receives the control, write  $Z_{ij} = 0$ , so  $1 = Z_{i1} + Z_{i2}$  for  $i = 1, \dots, I$ . The matched pairs were formed by matching for an observed covariate  $\mathbf{x}_{ij}$  but may have failed to control an unobserved covariate  $u_{ij}$ ; that is,  $\mathbf{x}_{ij} = \mathbf{x}_{ik}$  for all  $i, j, k$ , but possibly  $u_{ij} \neq u_{ik}$ .

For any outcome, each subject has two potential responses, one seen when the instrument encourages the subject to take the treatment,  $Z_{ij} = 1$ , the other seen when the instrument randomly assigns the subject to be encouraged to take the control,  $Z_{ij} = 0$ . Here, there are two responses, the potential outcomes  $(Y^{(Z_{ij}=1)}, Y^{(Z_{ij}=0)})$  and the potential treatment selections  $(D^{(Z_{ij}=1)}, D^{(Z_{ij}=0)})$ . Abbreviate these as  $(Y_{ij}^0, Y_{ij}^1)$  and  $(D_{ij}^0, D_{ij}^1)$ .

The effect ratio,  $\lambda$ , is the parameter

$$\lambda = \frac{\sum_{i=0}^I \sum_{j=0}^2 (Y_{ij}^1 - Y_{ij}^0)}{\sum_{i=0}^I \sum_{j=0}^2 (D_{ij}^1 - D_{ij}^0)}, \tag{18}$$

where it is implicitly assumed that  $0 \neq \sum_{i=0}^I \sum_{j=0}^2 (D_{ij}^1 - D_{ij}^0)$ . Here,  $\lambda$  is a parameter of the finite population of  $2I$  individuals, and because  $(Y_{ij}^0, Y_{ij}^1)$  and  $(D_{ij}^0, D_{ij}^1)$  are not jointly observed,  $\lambda$  cannot be calculated from observable data so inference is required.

To test the null hypothesis  $H_0: \lambda = \lambda_0$ , construct the following statistics

$$\begin{aligned} T(\lambda_0) &= \frac{1}{I} \sum_{i=1}^I \left\{ \sum_{j=1}^2 Z_{ij} (Y_{ij} - \lambda_0 D_{ij}) \right. \\ &\quad \left. - \sum_{j=1}^2 (1 - Z_{ij}) (Y_{ij} - \lambda_0 D_{ij}) \right\} \\ &= \frac{1}{I} \sum_{i=1}^I V_i(\lambda_0), \text{ say,} \end{aligned} \tag{19}$$

where, because  $Y_{ij} - \lambda_0 D_{ij} = Y_{ij}^1 - \lambda_0 D_{ij}^1$  if  $Z_{ij} = 1$  and  $Y_{ij} - \lambda_0 D_{ij} = Y_{ij}^0 - \lambda_0 D_{ij}^0$  if  $Z_{ij} = 0$ , write

$$\begin{aligned} V_i(\lambda_0) &= \sum_{j=1}^2 Z_{ij} (Y_{ij}^1 - \lambda_0 D_{ij}^1) \\ &\quad - \sum_{j=1}^2 (1 - Z_{ij}) (Y_{ij}^0 - \lambda_0 D_{ij}^0). \end{aligned} \tag{20}$$

Also, define

$$S^2(\lambda_0) = \frac{1}{I(I-1)} \sum_{j=1}^I \{V_i(\lambda_0) - T(\lambda_0)\}^2.$$

As shown in Baiocchi et al. (2010), under reasonable conditions, the hypothesis  $H_0: \lambda = \lambda_0$  may be tested by comparing the test statistic  $T(\lambda_0) / S(\lambda_0)$  to the standard normal.

## Multinomial, Survival and Distributional Outcomes

### Multinomial Outcome

Multinomial outcomes (i.e., nominal or ordinal outcomes) are common in health services research. For example, Bruce et al. (2004) conducted a randomized trial to improve adherence to prescribed depression treatments among depressed elderly patients in primary care practices; the outcomes of interest included continuous outcomes as well as multinomial outcomes such as the number of depression symptoms, ranging from 0 to 9, and the depression class (major, minor, or no depression). There was noncompliance in this trial, and Ten Have et al. (2004) used random assignment as an IV to estimate the effect of receiving treatment on continuous outcomes. Cheng (2009) considered how to estimate the effect of receiving treatment on the multinomial outcomes using random assignment as an IV.

For ordinal outcomes, the CACE is a function of coding scores and probabilities with respect to the categories:

$$\begin{aligned} \text{CACE} &= E(Y_i^1 - Y_i^0 | C_i = co) \\ &= \sum_j (W_j \times t_j) - \sum_j (W_j \times v_j) \\ &= \sum_j (W_j \times t_j) \\ &\quad - \frac{1}{\pi_c} \left[ \sum_j (W_j \times g_j) - (1 - \pi_c) \sum_j (W_j \times s_j) \right] \end{aligned}$$

where  $W_j$  is the coding score;  $t_j$ ,  $v_j$ , and  $s_j$  are the probabilities for compliers under treatment and control and never takers, respectively; and  $q_j$  is the probability for observed group  $Z_i = 0$ ,  $D_i = 0$  for the  $j^{\text{th}}$  category. For estimating the CACE for ordinal outcomes, the coding score needs to be chosen. Equally spaced scores or linear transformations of them, midranks and ridit scores are among the options. A sensitivity analysis can be performed with different choices of scores to see how the results differ.

In addition to the CACE, Cheng (2009) considered some other functions of outcome distributions

for understanding the causal effect for ordinal outcomes, including the measure of stochastic superiority of treatment over control for compliers –

$$\begin{aligned} \text{SSC} &= P(Y_i^1 > Y_i^0 | C_i = \text{complier}) \\ &\quad + \frac{1}{2} P(Y_i^1 = Y_i^0 | C_i = \text{complier}) \\ &= \sum_{j=k}^{J-1} \sum_{k=1}^{J-j} t_{j+k} v_j + \frac{1}{2} \sum_{j=1}^J t_j v_j \\ &= \sum_{j=k}^{J-1} \sum_{k=1}^{J-j} t_{j+k} v_j \left[ \frac{q_j - (1 - \pi_c) s_j}{\pi_c} \right] \\ &\quad + \frac{1}{2} \sum_{j=1}^J t_j \left[ \frac{q_j - (1 - \pi_c) s_j}{\pi_c} \right] \end{aligned} \tag{21}$$

$\text{SSC} = 0.5$  indicates no causal effect, and  $\text{SSC} > 0.5$  indicates beneficial effect of the treatment for compliers if a higher value of the outcome is a better result. Compared to the CACE, SSC is easy to interpret and avoids the problem of choosing scores  $W_j$ , but without use of weighting scores, it may not describe the strength of the effect well when some specific categories are known to be more important than other categories in measuring the treatment effect.

For nominal outcomes, it is difficult to get a summary measure of the causal effect such as the CACE or SSC for ordinal outcomes. Instead, the treatment effect on the entire outcome distributions of compliers with and without treatment can be evaluated, that is, to compare  $t_j$  to  $v_j$ ,  $j = 1, \dots, J$  and test the equality of  $t_j$  and  $v_j$ ,  $j = 1, \dots, J$ . Cheng (2009) estimated those causal effects with the likelihood method and proposed a bootstrap/double bootstrap version of a likelihood ratio test for the inference when the true values of parameters are on the boundary of the parameter spaces under the null.

### Survival Outcome

Compared to trials with continuous, binary, and multinomial outcomes, randomized trials with survival outcomes often have an issue of

administrative censoring in addition to noncompliance. For those studies, Robins and Tsiatis (1991) considered a structural accelerated failure time model and developed semiparametric estimators for this model. Joffe (2001) provided a good discussion of their approach and comparisons with other survival analysis methods. Loeys and Goetghebeur (2003) and Cuzick et al. (2007) considered a structural proportional hazards model in which the hazard of the potential failure time under treatment for a certain group of subjects is proportional to the hazard of the potential failure time under control for these same subjects. Both the structural accelerated failure time model and the structural proportional hazards model are semiparametric models, where the effect of the treatment on the distribution of failure times is modeled parametrically.

Baker (1998) extended the models and assumptions for discrete-time survival data and derived closed form expressions for estimating the difference in the hazards at a specific time between compliers under treatment and control based on maximum likelihood. Baker (1998)’s estimator is analogous to the standard IV estimator for a survival outcome. Nie et al. (2011) discussed this standard IV approach and parametric maximum likelihood methods for the difference in survival at a specific time between compliers under treatment and control.

Here, the standard IV approach of Baker (1998) will be reviewed. Let  $S_{c1}(V)$ ,  $S_{c0}(V)$ ,  $S_{at}(V)$ , and  $S_{nt}(V)$  be the potential survival functions at time  $V$  of compliers in the treatment and control groups and of always takers and never takers, respectively,  $S_z(V)$  be the survival probabilities at time  $V$  for the group with assignment  $Z = z$ , and  $S_{zd}(V)$  be the survival probabilities at time  $V$  for the group with assignment  $Z = z$  and treatment received  $D = d$ . By Table 1, the following holds

$$S_1(V) = \pi_c S_{c1}(V) + \pi_{at} S_{at}(V) + \pi_{nt} S_{nt}(V),$$

$$S_{11}(V) = \frac{\pi_c}{\pi_c + \pi_{at}} S_{c1}(V) + \frac{\pi_{at}}{\pi_c + \pi_{at}} S_{at}(V)$$

$$S_{10}(V) = S_{nt}(V) S_0(V) = \pi_c S_{c0}(V) + \pi_{at} S_{at}(V) + \pi_{nt} S_{nt}(V),$$

$$S_{00}(V) = \frac{\pi_c}{\pi_c + \pi_{nt}} S_{c0}(V) + \frac{\pi_{nt}}{\pi_c + \pi_{nt}} S_{nt}(V) S_0(V) = S_{at}(V)$$

Similar to the standard IV estimator for CACE, the standard IV estimator for the compliers difference in survival probabilities is

$$\hat{S}_{c1}(V) - \hat{S}_{c0}(V) = \frac{\hat{S}_1(V) - \hat{S}_0(V)}{\hat{E}(D|Z = 1) - \hat{E}(D|Z = 0)},$$

which is the difference of the observed survival probabilities at time  $V$  between compliers under treatment and control divided by the proportion of compliers.  $\hat{S}_z(V)$  is the Kaplan-Meier estimator under assignment  $z$ . In addition to the five IV assumptions discussed in section “[Framework and Notation](#),” an additional assumption is needed to ensure that the estimator based on Kaplan-Meier estimates is consistent:

*Independence Assumption of Failure Times and Censoring Times* The distributions of potential failure times  $T$  and administrative censoring times  $C$  are independent of each other. Type I censoring (i.e., censoring times are the same for all subjects) and random censoring are two special cases.

Although the standard IV estimator is very useful, it may give negative estimates for hazards and be inefficient because it does not make full use of the mixture structure implied by the latent compliance model. When the survival functions follow some parametric distributions, Nie et al. (2011) used the EM algorithm to obtain the MLE on the difference in survival probabilities for compliers. However, the MLEs could be biased when the parametric assumptions are not valid. To address this concern, Nie et al. (2011) developed a nonparametric estimator based on empirical likelihood that makes use of the mixture structure to gain efficiency over the standard IV method while not depending on parametric assumptions to be consistent.

### Effect of Treatment on Distribution of Outcomes

As discussed in previous sections, a large literature on methods of analysis for treatment effects

focuses on estimating the effect of treatment on average outcomes, for example, the CACE (Imbens and Angrist 1994; Angrist et al. 1996). However, in addition to the average effect, knowledge of the causal effect of a treatment on the outcome distribution and its general functions can often provide additional insights into the impact of the treatment and therefore can be of significant interest in many situations (Poulson et al. 2012). For example, in a study of the effect of school subsidized meal programs on children's weight, both low weight and high weight are adverse outcomes; therefore, knowing the effect of the program on the entire distribution of outcomes rather than just average weight is important for understanding the impact of the program. For an individual patient deciding which treatment to take, the patient must weight the effects of the possible treatments on the distribution of outcomes, the costs of the treatments and the potential side effects of the treatments (Hunink et al. 2001). Therefore, making the best decision

requires information on the treatment's effect on the entire distribution of outcomes rather than just the average effect because a patient's utility over outcomes may be nonlinear over the outcome scale (Karni 2009; Pliskin et al. 1980). Hogan and Lee (2004), Saigal et al. (1999), and Sommers et al. (2007) provide examples in HIV care, neonatal care, and cancer care, respectively.

For distributional treatment effects on non-degenerate outcome variables with bounded support, without any parametric assumption, Abadie (2002) used the standard IV approach to estimate the counterfactual cumulative distribution functions (cdf) of the outcome of compliers with and without the treatment and proposed a bootstrap procedure to test distributional hypotheses with the Kolmogorov-Smirnov statistic. However, Abadie (2002) and Imbens and Rubin (1997a) pointed out that the standard IV estimates of the potential cdfs for compliers may not be nondecreasing functions:

$$\begin{aligned} \hat{H}_{c1}(y)^{SIV} &= \frac{\hat{E}\{1(Y_i \leq y)D_i | Z_i = 1\} - \hat{E}\{1(Y_i \leq y)D_i | Z_i = 0\}}{\hat{E}(D_i | Z_i = 1) - \hat{E}(D_i | Z_i = 0)} \hat{H}_{c0}(y)^{SIV} \\ &= \frac{\hat{E}\{1(Y_i \leq y)(1 - D_i) | Z_i = 1\} - \hat{E}\{1(Y_i \leq y)(1 - D_i) | Z_i = 0\}}{\hat{E}\{(1 - D_i) | Z_i = 1\} - \hat{E}\{(1 - D_i) | Z_i = 0\}}, \end{aligned}$$

where  $\hat{H}_{c1}(y)^{SIV}$  and  $\hat{H}_{c0}(y)^{SIV}$  are the standard IV estimators for compliers' cumulative distribution function (cdf) under treatment and control, respectively. Furthermore, as discussed in section "More Efficient Estimation," the standard IV approach does not make full use of the mixture structure (Imbens and Rubin 1997a) implied by the latent compliance class model (see Table 1) and hence could be less efficient. Instead, Imbens and Rubin (1997a) proposed a normal approximation and two multinomial approximations to the outcome distributions. However, the estimator based on a normal approximation could be biased when the outcomes are not normal, and for the approach based on multinomial approximations, a systematic approach for choosing the multinomial approximations is needed.

Cheng et al. (2009a) developed a semi-parametric instrumental variable method based on the empirical likelihood approach. Their approach makes full use of the mixture structure implied by the latent compliance class model without parametric assumptions on the outcome distributions as well as takes into account the nondecreasing property of cdfs and can be easily constructed based on data. Their method can be applied to general outcomes and general functions of outcome distributions. Cheng et al. (2009a) showed that their estimator has good properties and is substantially more efficient than the standard IV estimator.

For the mixture structure implied by the latent compliance model (see Table 1), Cheng et al. (2009a) adopted a density ratio model proposed

by Anderson (1979) to relate the densities of the latent compliance classes by an exponential tilt:

$$\frac{h_j(y)}{h_0(y)} = \exp(\alpha_j + \beta_j y), \quad j = 1, 2, 3 \quad (22)$$

where  $h_0(y)$  is unspecified and  $h_0(y) = P(Y_i = y|Z_i = 0, C_i = co)$ ,  $h_1(y) = P(Y_i = y|C_i = nt)$ ,  $h_2(y) = P(Y_i = y|Z_i = 1, C_i = co)$ ,  $h_3(y) = P(Y_i = y|C_i = at)$  are the outcome density (mass) functions of the latent compliance groups: compliers under control, never takers, compliers under treatment, and always takers, respectively; The densities are modeled nonparametrically except for being related by a parametric “exponential tilt.” The idea is similar to Cox’s proportional hazard models, and many conventional parametric families fall in the exponential tilt model category, including two normals with common variance but different means, two exponential distributions, and two Poissons. The exponential tilt model provides a good fit to the data when many conventional parametric models do not fit the data well.

Let  $f_{zd}(y) = P(Y_i = y|Z_i = z, D_i = d)$  and  $F_{zd}(y) = P(Y_i \leq y|Z_i = z, D_i = d)$  be the probability density (mass) function and cumulative distribution function of the observed group ( $Z_i = z, D_i = d$ ) for continuous (discrete) outcome, respectively, where  $z, d = 0, 1$ . Then, by the IV assumptions and latent compliance class model (see Table 1), the following holds

$$\begin{aligned} f_{11}(y) &= \lambda h_2(y) + (1 - \lambda)h_3(y), \\ f_{10}(y) &= h_1(y), \quad f_{00}(y) = \tau h_0(y) \\ &+ (1 - \tau)h_1(y), \quad f_{01}(y) = h_3(y). \end{aligned} \quad (23)$$

where

$$\begin{aligned} \lambda &= \frac{\phi_c}{\phi_c + \phi_a} = \frac{1 - \phi_a - \phi_n}{1 - \phi_n}, \quad \tau = \frac{\phi_c}{\phi_c + \phi_n} \\ &= \frac{1 - \phi_a - \phi_n}{1 - \phi_a} \end{aligned}$$

The causal effect of actually receiving the treatment on the outcome distribution for compliers can be examined by considering  $h_0(y)$  and  $h_2(y)$ .

Under the density ratio model (22), the log likelihood is

$$\begin{aligned} \ell &= n_{01} \log \phi_a + n_{00} \log(1 - \phi_a) \\ &+ n_{10} \log \phi_n + n_{11} \log(1 - \phi_n) \\ &+ \sum_{i=1}^n [I(Z_i = 0, D_i = 1)(\alpha_3 + \beta_3 y_i) \\ &+ I(Z_i = D_i = 0) \log\{\lambda + (1 - \lambda) \exp(\alpha_1 + \beta_1 y_i)\}] \\ &+ \sum_{i=1}^n [I(Z_i = D_i = 1) \log\{\tau \exp(\alpha_2 + \beta_2 y_i) \\ &+ (1 - \tau) \exp(\alpha_3 + \beta_3 y_i)\}] \\ &+ \sum_{i=1}^n [I(Z_i = 1, D_i = 0)(\alpha_1 + \beta_1 y_i)] + \sum_{i=1}^n \log h_0(y_i) \end{aligned}$$

where  $h_0(\cdot)$  is unspecified, and

$$\begin{aligned} h_0 \in C = \left\{ h_0 | h_0(y_i) \geq 0, \sum_{i=1}^n h_0(y_i) = 1, \sum_{i=1}^n h_0(y_i) \right. \\ \left. \exp(\alpha_j + \beta_j y_i) = 1, j = 1, 2, 3 \right\} \quad (24) \end{aligned}$$

Note that  $h_0(\cdot)$  will put its support on observed data points  $y_1, \dots, y_n$  (Owen 2002) and constraint (24) ensures that the estimators for outcome distributions  $H_0, H_1, H_2$ , and  $H_3$  are cumulative distribution functions. Similar to Qin and Zhang (1997), after maximizing the log likelihood with constraint (24) through Lagrange multipliers, the following holds:

$$\begin{aligned} h_0(y_i) &= \frac{1}{n} \frac{1}{n + \sum_{j=1}^3 \xi_j \{\exp(\alpha_j + \beta_j y_i) - 1\}}, \\ &j = 1, 2, 3 \end{aligned} \quad (25)$$

where  $\xi_j$ 's ( $j = 1, 2, 3$ ) are Lagrange multipliers determined by

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \frac{\exp(\alpha_j + \beta_j y_i) - 1}{1 + \sum_{j=1}^3 \xi_j \{\exp(\alpha_j + \beta_j y_i) - 1\}} \\ = 0, \quad j = 1, 2, 3 \end{aligned} \quad (26)$$

and the limiting values of  $\xi$  are

$$\xi^0 = \begin{Bmatrix} \xi_1^0 \\ \xi_2^0 \\ \xi_3^0 \end{Bmatrix} = \begin{Bmatrix} \delta\phi_n + (1-\delta)(1-\phi_a)(1-\lambda) \\ \tau\delta(1-\phi_n) \\ (1-\delta)\phi_a + \delta(1-\phi_n)(1-\tau) \end{Bmatrix}$$

Then, the maximum semiparametric empirical likelihood estimate of  $\eta = (\phi_a, \phi_n, \alpha_1, \beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3)$  can be obtained by maximizing the profiled log likelihood through the EM algorithm. And then the outcome densities (masses) of compliers under control ( $h_0(y)$ ) and treatment ( $h_2(y)$ ) can be estimated by  $\hat{h}_0(y_i)$ ; see Eq. 25, and  $\hat{h}_0(y_i)\exp(\hat{\alpha}_2 + \hat{\beta}_2 y_i)$ , respectively, and their corresponding cdfs  $H_0(y)$  and  $H_2(y)$  are estimated by  $\hat{H}_0(y) = \sum_i \hat{h}_0(y_i)I(y_i \leq y)$  and  $\hat{H}_2(y) = \sum_i \hat{h}_0(y_i)\exp(\hat{\alpha}_2 + \hat{\beta}_2 y_i)I(y_i \leq y)$ , respectively. To examine the causal effect of actually receiving treatment on the outcome distribution for compliers, the equality of  $h_0(y)$  and  $h_2(y)$  can be tested by testing  $H_0: \alpha_2 = \beta_2 = 0$  by the semiparametric empirical likelihood ratio statistic

$$R = 2 \left\{ \max_{\eta} \ell(\eta) - \max_{\eta_1, \alpha_2 = \beta_2 = 0} \ell(\eta_1, \alpha_2 = \beta_2 = 0) \right\},$$

$$\eta_1 = (\alpha_1, \beta_1, \alpha_3, \beta_3, \phi_a, \phi_n)$$

where  $\alpha_2$  must equal 0 when  $\beta_2$  equals 0 because of constraint (24). Under regularity conditions,  $R$  follows a chi-squared distribution with one degree of freedom asymptotically under the null hypothesis.

In addition to investigating the distributional treatment effect, some function of the outcome distributions,  $g(\eta)$ , where  $g$  is a real-valued function with nonzero first partial derivatives, can also be estimated. For example, under the semiparametric setting in Cheng et al. (2009a), the CACE can be estimated by using

$$\widehat{CACE}^{SEM} = \sum_{i=1}^n y_i \hat{h}_0(y_i) \{ \exp(\hat{\alpha}_2 + \hat{\beta}_2 y_i) - 1 \}.$$

One can also compare the  $\iota$  – quantiles of outcome distributions of compliers with and without treatment (marginal distributions of  $Y^1$  and  $Y^0$ ):

$$\widehat{CQCE}^{SEM} = \hat{H}_2^{-1}(\iota) - \hat{H}_0^{-1}(\iota)$$

When  $\iota = 0.5$ , it is the difference of the medians for the compliers under treatment and control.

The goodness of fit of the density ratio model can be tested by comparing estimated outcome cdfs based on the density ratio model to the empirical distribution function estimates (Qin and Zhang 1997):

$$\Delta_{zd} = \sup_{-\infty < y < \infty} \sqrt{n} |\hat{F}_{zd}(y) - \tilde{F}_{zd}(y)|, \quad z, d = 0, 1. \quad (28)$$

The  $p$ -value of the goodness-of-fit test can be estimated by a bootstrap  $p$ -value

$$\hat{P}_{zd}^B = \hat{P}_{zd}^B(\Delta_{zd}^* \geq \Delta_{zd}^{obs}) \quad (29)$$

where  $\Delta_{zd}^{obs}$  is obtained from the actually observed data and  $\Delta_{zd}^*$  is calculated from  $B$  bootstrap samples generated under the null hypothesis: the density ratio model (22) is true.

---

## Study Design IV and Multiple IVs

### Study Design IV: Near-Far Matching

Study design focuses attention on the data which is to be analyzed. The manner in which the data are structured largely determines the statistical procedures appropriate for analysis. The separation between study design and statistical analysis is quickly illustrated by considering a uniform randomized paired analysis. The process of matching individual units of observation into pairs based on observed, pretreatment covariates, and then randomizing one unit within each pair to treatment and the other to control is study design. The researcher constructs the pairs by carefully controlling the assignments to increase efficiency by decreasing within pair variation (by constructing matched pairs) as well as to minimize unobserved bias (by randomization). These steps increase the validity of the results and go a

long way toward reassuring the audience of the reliability of the reported conclusions. Only the manner in which the data are prepared has thus far been described. This is the design of the study.

Once the experiment is run and the data are recorded, then the results need to be analyzed. Given the study design, most analysts would select a paired *t*-test, perhaps using student's *t*. But that is not the only choice; one could justifiably use a permutational test or, with some additional assumptions, a model-based approach (e.g., regression) to adjust for potential covariate imbalances which routinely occur in finite sample randomizations. This is the statistical inference phase of the study. Statistical inference is distinct from, though predicated on and preceded by, the study design. The more well understood the study design, the more credibility the statistical inference is likely to have. This is true in experimentation and even truer in the observational setting.

In observational settings data is often plentiful, especially compared to the experimental setting. The trouble with observational data is that estimates of treatment effects tend to be plagued by confounding by both observed and unobserved covariates. The goal of study design in the observational setting can be thought of as finding the subset of the data which will produce the best study given the limitations of the data (usually in the sense of internal validity).

In the literature, study design is also sometimes referred to as “preprocessing” (Ho et al. 2007). For those new to study design, perhaps the most unintuitive insight is that the analysis can actually be improved by removing observations from consideration before performing the statistical inference. This is unintuitive because, loosely speaking, it seems like the study with the most observations is the most informative. This is a recognized problem in the observational literature. For example, it has become standard practice to use propensity scores to limit the analysis to only the observation units which have corresponding propensity score values in either the treated or control group, removing from inference the observational units with extreme values close to 1 or 0 (Rosenbaum 2002, 2009).

Analogously for instrumental variables, it is known that if the goal is to have greater power and results which are more robust to small violations of the IV assumptions, then a smaller data set with a stronger instrument is preferable to a larger data set with a weaker instrument (Small and Rosenbaum 2008). The trade-off between bigger but weaker and smaller but stronger was thought to be informative, but not useful once the analyst has committed to using a particular data set. Contrary to this belief, Baiocchi et al. (2010) demonstrated that even within a particular data set, the analyst may use near-far matching to go from a weaker-but-bigger study to a more robust smaller-but-stronger study.

There are two objectives in near-far matching. As in a randomized controlled trial (RCT) with a matched-pair design, one objective in near-far matching is to create matched pairs where the covariates are similar within a pair. Creating pairs with very similar covariate values (i.e., pairs which are near each other in covariate space) is used to improve efficiency. The other objective in near-far matching is to separate observations' instrument values within a matched pair. In the neonatal intensive care example outlined in the introduction, within a matched pair, one wants one mother to be highly encouraged to deliver at a high-level NICU and the other to be highly encouraged to deliver at a low-level NICU. This is similar to the matched-pair design when there is the potential for non-compliance. If the level of encouragement can be varied, then it is preferential to have two mothers who are highly dissimilar (far) in their levels of randomly assigned encouragement because it is then more likely that within the pair, one mother will comply with the encouragement and take the treatment and the other will comply with the lack of encouragement and take the control. As outlined in Baiocchi et al. (2010), algorithms exist which will construct pairs which maximize both of these objectives at the same time.

In most real-world examples, there will be a trade-off between the “near” and the “far” part of the matching. The technical aspects of this trade-off, and how to construct such pairs, are context specific – for guidance see Baiocchi et al. (2010, 2012). The intuition is that as the analyst forces

separation in the instrument values between pairs of patients it becomes more difficult to find patients with quite dissimilar instrument values but very similar covariates. The Baiocchi et al. (2010) paper outlines both theoretical arguments as well as practical reasons for designing studies with greater separation in the instrument.

It should be noted that pair matching is being referred to, but all of these arguments hold for larger block designs. Near-far matching would work with k:1 matching and other more exotic designs. The primary difference would be the optimization algorithm used to construct the sets.

This process is similar to propensity score matching and other matching techniques in general. The goal is to prepare the data, by finding the parts of the data set which lend themselves to causal inference, so as to improve the reliability of the statistical analysis to be performed. Note that, just as with propensity score matching, the analyst may decide to use whichever appropriate statistical method of analysis post-matching. That is, after performing near-far matching, the analyst may then decide to use a 2SRI model if that is appropriate for the given data set. But, the selection of the statistical method must be made with justification, not out of convenience. This is why most analysts will decide to use the effect ratio (discussed in section “Binary Outcomes”) after performing near-far matching as the study design leads naturally into the statistical analysis.

## Multilevel and Continuous IVs

In some settings, the IV has multiple levels or is continuous. For example, in the neonatal intensive care example, the mother’s excess travel time from the nearest high-level NICU compared to the nearest low-level NICU is continuous. Multiple levels of the IV provides us with the opportunity to identify a richer set of causal effects (Imbens 2007). Suppose the IV is continuous and the following extended monotonicity assumption holds  $D_i^z \geq D_i^{z'}$  for all  $z_i \geq z'_i$ , that is, a higher level of the IV always leads to at least as high a level of the treatment. The limit of the treatment

effect for subjects who would take treatment if the IV was equal to  $z$  but not take the treatment if the IV was a little less than  $z$  is  $\lim_{\epsilon \rightarrow 0} E [Y_i^{d=1} - Y_i^{d=0} | D_i^z = 1, D_i^{z-\epsilon} = 0]$ ; Heckman and Vytlacil (1999) refer to this as the marginal treatment effect at  $z$ . Treatment effects of interest can all be expressed as a weighted average of these marginal treatment effects (Heckman and Vytlacil 1999). For example, the treatment effect estimated by dichotomizing the IV as 1 or 0 according to whether the IV is above some cutoff or the treatment effect estimated by two-stage least squares using the continuous IV can be expressed as a weighted average of the marginal treatment effects. The average treatment effect over the whole population can also be expressed as a weighted average of the marginal treatment effects. Identification of the average treatment effect over the whole population requires identification of all the marginal treatment effects. In order for all the marginal treatment effects to be identified using the IV (and thus the average treatment effect identified), it is required that for large values of  $Z$ ,  $P(D = 1 | Z)$  approaches 1 and for small values of  $Z$ ,  $P(D = 1 | Z)$  approaches 0 (Heckman and Vytlacil 1999). Basu et al. (2007) show how to estimate marginal treatment effects and the average treatment effect when this condition is satisfied.

## Multiple IVs

In some settings, there may be multiple IVs available. For example, Malkin et al. (2000) used IV methods to estimate the effect of longer postpartum stays on newborn readmissions. Malkin et al. (2000) used two IVs, (1) hour of birth and (2) method of delivery (vaginal vs. C-section). Hour of birth influences length of stay because it affects whether a newborn will spend an extra night in the hospital; for example, Malkin et al. (2000) found that newborns born in the a.m. have longer lengths of stay than newborns born in the p.m. Method of delivery influences length of stay because mothers need more time to recuperate after a C-section than following a vaginal delivery, and newborns are rarely discharged before



their mothers. Each IV identifies the treatment effect for a different set of compliers. If treatment effects are heterogeneous, the complier average causal effects may differ. For example, newborns who would only stay an extra day if born in the a.m. compared to the p.m. may differ in their risk characteristics compared to newborns who would only stay an extra day if delivered by C-section compared to vaginal delivery, and length of stay may have a different effect on newborns with different risk characteristics.

Two-stage least squares can be used to combine the IVs – in the first stage, regress  $D$  on both  $Z_1$  and  $Z_2$  (as well as  $\mathbf{X}$ ) and then use the predicted  $D$  as usual in the second stage. Under the assumption of homogeneous treatment effects and constant variance, the two-stage least squares estimate is the optimal way to combine the IVs (White 1984). When treatment effects are heterogeneous, two-stage least squares estimates a weighted average of the complier average causal effect for the IVs with stronger IVs getting greater weight (Imbens and Angrist 1994; Angrist and Imbens 1995). When there are two or more distinct IVs, it is useful to report the estimates from the individual IVs in addition to the combined IVs since the IVs may be estimating treatment effects for different types of people.

When there are multiple IVs and treatment effects are homogeneous, the overidentifying restrictions test can be used to test the validity of the IVs (Davidson and MacKinnon 1993; Sargan 1958). The overidentifying restrictions test tests whether the estimates from the different IVs are the same. When treatment effects are homogeneous, if the estimates from two different IVs converge to different limits, this would show that at least one of the IVs is invalid. There are two problems with using the overidentifying restrictions test to test the validity of IVs. First, if treatment effects are heterogeneous, then the complier average causal effects for the two IVs may be different even though both IVs are valid; in this case, the overidentifying restrictions test would falsely indicate that at least one of the IVs is invalid. Second, even if treatment effects are homogeneous, two IVs  $A$  and  $B$  may both be biased but in the same way so that the asymptotic

limit of the estimators based on IV  $A$  and  $B$ , respectively, is the same; in this case, the overidentifying restrictions test would give false assurance that the IVs are valid (Small 2007).

### Multilevel and Continuously Valued Treatments

The treatment under study may take on multiple or continuous values, for example, the dose of a medication. Two-stage least squares can still be applied. Angrist and Imbens (1995) present the following formula that shows that the two-stage least squares estimator converges to a weighted average of the effect of one unit changes in the treatment level. Suppose the treatment can take on levels  $0, 1, \dots, \bar{d}$  and that monotonicity holds in the sense that  $D_i^{z=1} \geq D_i^{z=0}$ . Assume there are no covariates. Then, the two-stage least squares estimator converges to

$$\begin{aligned} & \frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(D_i|Z_i = 1) - E(D_i|Z_i = 0)} \\ &= \sum_{d=1}^{\bar{d}} \omega_d E[Y^d - Y^{d-1} | D^{z=1} \geq d > D^{z=0}], \end{aligned} \tag{30}$$

where  $\omega_d = \frac{P(D^{z=1} \geq d > D^{z=0})}{\sum_{d=1}^{\bar{d}} P(D^{z=1} \geq d > D^{z=0})}$ . The numerator of  $\omega_d$  is the proportion of compliers at point  $d$ , that is, the proportion of individuals driven by the encouraging level of the IV from a treatment intensity less than  $d$  to at least  $d$ . The  $\omega_d$ 's are nonnegative and sum to one. The quantity  $E[Y^d - Y^{d-1} | D^{z=1} \geq d > D^{z=0}]$  in Eq. 30 is the causal effect of a one unit increase in the treatment from  $d - 1$  to  $d$  for compliers at point  $d$ . Equation 30 shows that the two-stage least squares estimator converges to a weighted average of the causal effects of one unit increases in the treatment from  $d - 1$  to  $d$  for compliers at point  $d$ , where the points  $d$  at which there are more compliers get greater weight. The weights  $\omega d$  can be estimated since under monotonicity and the assumption that the IV is independent of the potential treatment received,  $P(D^{z=1} \geq d > D^{z=0}) = P(D^{z=1} \geq d) - P$

$(D^{z=0} \geq d) = P(D \geq d|Z = 1) - P(D \geq d|Z = 0)$ . See Angrist and Imbens (1995) for an extension of these formulas to the setting where there are covariates  $\mathbf{X}$  that are controlled for.

Researchers often times dichotomize multi-level or continuous treatments. However, using IV methods with a dichotomized continuous treatment can lead to an overestimate of the treatment effect. Let  $\beta$  denote the average causal effect (30) that the two-stage least squares estimator for a multilevel treatment converges to Angrist and Imbens (1995) that show that if this treatment is dichotomized as  $B = 1$  if  $D \geq l$ ,  $B = 0$  if  $D < l$  for some  $1 \leq l \leq \bar{d}$ , then the two-stage least squares estimator using the binary treatment  $B$  converges to  $\phi\beta$  where

$$\begin{aligned} \phi &= \frac{E(D|Z = 1) - E(D|Z = 0)}{E(B|Z = 1) - E(B|Z = 0)} \\ &= \frac{\sum_{j=1}^{\bar{d}} P(D^{z=1} \geq j > D^{z=0})}{P(D^{z=1} \geq l > D^{z=0})} \geq 1 \end{aligned}$$

The only situation when  $\phi = 1$  is when the IV has no effect other than to cause people to switch from  $D = l - 1$  to  $D = l$ . Otherwise, when a multilevel treatment is incorrectly parameterized as binary, the resulting estimate tends to be too large relative to the average per-unit effect of the treatment. The problem with dichotomizing a multilevel treatment is that the IV has a direct effect because the encouraging level of the IV can push a person to a higher level of treatment even if  $B$  is 1 under both the non-encouraging and encouraging levels of the IV.

Although dichotomizing a continuous treatment results in a biased IV estimate, the sign of the treatment effect is still consistently estimated.

If the treatment effect for compliers is linear, that is, the causal effect of a one unit increase in the treatment from  $d - 1$  to  $d$  for compliers at point  $d$  is the same for all  $d$ , then the two-stage least squares estimator estimates this linear treatment effect. If the treatment effect is nonlinear, then with a binary IV, it is not possible to estimate anything other than the weighted treatment effect (30). If the IV is continuous, then the IV can be

used to form multiple IVs (e.g.,  $Z$ ,  $Z^2$ ,  $Z^3$ , etc.), and a nonlinear treatment effect can be estimated (Kelejian 1971). For example, suppose  $Y^{D=d} = Y^0 + \beta_1 d + \beta_2 d^2$ . Then,  $\beta_1$  and  $\beta_2$  can be consistently estimated with a continuous IV  $Z$  by using two least squares where  $\hat{D}$  is estimated by regressing  $D$  on  $Z$  and  $Z^2$ ,  $\hat{D}^2$  is estimated by regressing  $D^2$  on  $Z$  and  $Z^2$ , and  $\beta_1$  and  $\beta_2$  are estimated by regressing  $Y$  on  $\hat{D}$  and  $\hat{D}^2$ . Tan (2010) discusses other estimation approaches for estimating nonlinear treatment effects.

A common setting is to have a treatment with three levels that may not be strictly ordered by dose. Cheng and Small (2006) consider the setting of a treatment with three levels – control (0) and two active levels  $A$  and  $B$ , where  $A$  and  $B$  are not ordered by dose and some subjects may prefer  $A$  to  $B$  and some may prefer  $B$  to  $A$ . Subjects are randomly assigned to one of the three arms 0,  $A$  and  $B$ , and then could either take the assigned treatment or not take it and receive the control (for the control arm, all subjects receive the control 0). The effect of treatment  $A$  versus control for subjects who would take treatment  $A$  if offered it (i.e., compliers with treatment  $A$ ) is identified by analyzing only subjects who were either assigned to the control arm or the treatment  $A$  arm. But for this setting, Cheng and Small (2006) showed that the effect of treatment  $A$  for subjects who would take treatment  $A$  if assigned to it but not treatment  $B$  and the effect of treatment  $A$  for subjects who would take treatments  $A$  or  $B$  if assigned to  $A$  or  $B$ , respectively, is not point identified. However, the data provides information that can be used to narrow bounds on these treatment effects. These treatment effects are of interest for individuals making decisions about which treatment to take, for example, for a very compliant subject who knows she would take either treatment  $A$  or  $B$  if offered it, she would like to know whether treatment  $A$  or  $B$  is better among very compliant subjects like herself; the treatment effects are also of interest for clinicians deciding which treatment to offer first and for health policymakers anticipating what would happen were the treatment (s) to be introduced into general practice in a

setting in which compliance patterns are expected to differ from those of the trial (Cheng and Small 2006).

---

### Extended Instrumental Variable Method for When Proposed IV Has a Direct Effect

When a proposed IV  $Z$  is thought to be independent of unmeasured confounders but there is concern that  $Z$  might have a direct effect on the outcome, Joffe et al. (2008) proposed an extended instrumental variables strategy for obtaining a consistent (i.e., asymptotically unbiased) estimate of the causal effect of treatment that requires having a covariate  $W$  for which:

- (eiv-a1). *The covariate  $W$  interacts with  $Z$  in affecting treatment.*
- (eiv-a2). *The direct effect of  $Z$  does not depend on  $W$ .*

For such a setting, Joffe et al. (2008) show that a consistent estimate of the treatment effect can be obtained under the additional assumption that the treatment effect is constant across subjects by using two-stage least squares where  $Z \times W$  is the IV and  $Z$  and  $W$  are included as measured covariates (other covariates can also be included in addition). As an example of this approach, Card (1995) studied the effect of education on earnings and considered having grown up near a 4-year college as an IV, but was concerned that growing up near a college might have a direct effect on earnings, for example, through the presence of a college being associated with higher school quality at nearby elementary and secondary schools. Card considered the covariate  $W =$  whether the person grew up in a low-income household. The interaction between growing up near a 4-year college and being from a low-income household predicts going to college, because college proximity lowers the cost of higher education and this cost lowering has a bigger effect on going to college for children from low-income families. In order for the extended instrumental variable

strategy to produce a consistent estimate, the effect of higher elementary/secondary school quality on earnings would have to be the same for children from low-income and high-income families – this is assumption (eiv-a2).

---

### Software

Software for implementing IV analyses is available in R, SAS, and Stata. Here an IV analysis will be illustrated using the AER package in the freely available software R. Consider estimating the causal effect of military service during the World War II era on men's future earnings using data from the 5% public use 1980 Census. The Census data contain information on a man's race and Census division of birth, but is missing information on variables such as health and criminal behavior, which were important barriers to serving in the war and are important determinants of earnings. Motivated by this concern about unmeasured confounding, Angrist and Krueger (1994) proposed to use time of birth as an IV; see also Small and Rosenbaum (2008) for follow-up analyses. Time of birth is associated with military service because a man only becomes eligible to serve in the military when he turns 18; men who turned 18 after World War II was over are substantially less likely to have served in the military. Here, consider the binary IV,  $Z = 1$  if a man was born between 1925 and 1927 (most men born in these time periods turned 18 during World War II) and  $Z = 0$  if a man was born in 1928 (so turned 18 after World War II was over). The data set used in the analysis `military_earnings.csv` is available at [www-stat.wharton.upenn.edu/dsmall/military-earnings.csv](http://www-stat.wharton.upenn.edu/dsmall/military-earnings.csv), and the data is described in the file [www-stat.wharton.upenn.edu/dsmall/military-earnings-readme.txt](http://www-stat.wharton.upenn.edu/dsmall/military-earnings-readme.txt).

```
library(AER)
dataset=read.csv("military-earnings.
csv",header=TRUE) attach(dataset);
# earnings = earnings in 1980
# veteran = 1 if World War II veteran,
0 if not
```

```

# yrquarter = year/quarter of birth,
e.g., born in 1927 first quarter =
1927, born in 1927 second qua
# born in 1927 fourth quarter =
1927.75
# racecat = 1, white; 2, black;
3, other
# Make race into a categorical variable
# racecat=as.factor(racecat);
# division = Census division of
birthplace,
# Census Division of birthplace
# 1= New England, 2= Middle Atlantic,
3 = East North Central,
# 4= West North Central, 5 = South
Atlantic, 6 = East South Central
# 7 = Mountain, 8 = Pacific, 9 =
American Territories
# Make division into a categorical
variable division=as.factor(division);
# IV is 1 if born in 1925-1927, 0 if
born in 1928 z=(yrquarter<=1927.75)
# Strength of the IV
> mean(veteran[z==1]) [1] 0.7363794
> mean(veteran[z==0]). [1] 0.3169782
# It is estimated that 0.736-
0.317=0.419 of the men are compliers
and the IV is moderately strong
# First stage of the two stage least
squares regression
# Find partial F test statistic for IV
fsreg=lm(veteran~racecat+division+z)
reg.without.iv=lm(veteran~racecat
+division) anova(fsreg,reg.without.iv)
Analysis of Variance Table
Model 1: veteran ~ racecat + division
Model 2: veteran ~ racecat + division +
z
Res.Df RSS Df Sum of Sq F Pr(>F)
<inline figure>
# The partial F statistic is 21,747,
much greater than 10, and thus there is
no concern about the IV
# being too weak for two stage least
squares to be reliable
# Two stage least squares regression
using z as the IV and controlling for
race and Census division
# of birth

```

```

tslsreg=ivreg(earnings~veteran
+racecat+division,~z+racecat+division)
summary(tslsreg)
Call:
ivreg(formula = earnings ~ veteran +
racecat + division | z + racecat +
division)
Residuals:
<inline figure>
Coefficients:
<inline figure>
Signif. codes: 0 *** 0.001 ** 0.01 *
0.05 . 0.1 1
Residual standard error: 12750 on
127073 degrees of freedom Multiple
R-Squared: 0.02788, Adjusted R-squared:
0.02779 Wald test: 408.2 on 11 and
127073 DF, p-value: < 2.2e-16
# It is estimated that military service
decreases a man's earnings by $834 with a
# standard error of $197. There is
strong evidence that military service
# decreases earnings (p-value <
0.0001).

```

**Acknowledgments** Jing Cheng and Dylan Small were supported by grant RC4MH092722 from the National Institute of Mental Health. The authors thank Scott Lorch for the use of the data from the NICU study.

---

## References

- Abadie A. Bootstrap tests for distributional treatment effects in instrumental variable models. *J Am Stat Assoc.* 2002;97:284–92.
- Abadie A. Semiparametric instrumental variable estimation of treatment response models. *J Econ.* 2003;113:231–63.
- Aidoo M, Terlouw D, Kolczak M, McElroy P, ter Kuile F, Kariuki S, Nahlen B, Lal A, Udhayakumar V. Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet.* 2002; 359:1311–2.
- Anderson J. Multivariate logistic compounds. *Biometrika.* 1979;66:17–26.
- Angrist J. Estimation of limited dependent variable models with dummy endogenous regressors. *J Bus Econ Stat.* 2001;19:2–28.
- Angrist J, Imbens G. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J Am Stat Assoc.* 1995;90:430–42.

- Angrist J, Krueger A. Does compulsory school attendance affect schooling and earnings? *Q J Econ.* 1991;106:979–1014.
- Angrist J, Krueger A. The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *J Am Stat Assoc.* 1992;87:328–36.
- Angrist J, Krueger A. Why do World War II veterans earn more than nonveterans? *J Labor Econ.* 1994;12:74–97.
- Angrist J, Pischke J-S. *Mostly harmless econometrics: an empiricist's companion.* Princeton/Oxford: Princeton University Press; 2009.
- Angrist J, Imbens G, Rubin D. Identification of causal effects using instrumental variables. *J Am Stat Assoc.* 1996;91:444–55.
- Baiocchi M, Small D, Lorch S, Rosenbaum P. Building a stronger instrument in an observational study of perinatal care for premature infants. *J Am Stat Assoc.* 2010;105:1285–96.
- Baiocchi M, Small D, Yang L, Polsky D, Groeneveld P. Near/far matching: a study design approach to instrumental variables. *Health Serv Outcome Res Methodol.* 2012;12:237–53.
- Baker S. Analysis of survival data from a randomized trial with all-or-none compliance: estimating the cost-effectiveness of a cancer screening program. *J Am Stat Assoc.* 1998;93:929–34.
- Balke A, Pearl J. Bounds on treatment effects for studies with imperfect compliance. *J Am Stat Assoc.* 1997;92:1171–6.
- Basu A, Heckman J, Navarro-Lozano S, Urzua S. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. *Health Econ.* 2007;16:1133–57.
- Bhattacharya J, Goldman D, McCaffrey D. Estimating probit models with self-selected treatments. *Stat Med.* 2006;25:389–413.
- Bhattacharya J, Shaikh A, Vytlacil E. Treatment effect bounds under monotonicity assumptions: an application to Swan-Ganz catheterization. *Am Econ Rev.* 2008;98:351–6.
- Bound JD, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak. *J Am Stat Assoc.* 1995;90:443–50.
- Brookhart M, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *Int J Biostat.* 2007;3:14.
- Brookhart M, Wang P, Solomon D, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology.* 2006;17:268–75.
- Brookhart M, Rassen J, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf.* 2010;19:537–54.
- Brooks J, Chrischilles E, Scott S, Chen-Hardee S. Was breast conserving surgery underutilized for early stage breast cancer? Instrumental variables evidence for stage II patients from Iowa. *Health Serv Res.* 2004;38:1385–402.
- Bruce M, Ten Have T, Reynolds C III, Katz I, Schulberg H, Mulsant B, Brown G, McAvay G, Pearson J, Alexopoulos G. Reducing suicidal ideation and depressive symptoms in depressed older primary care patients: a randomized trial. *J Am Med Assoc.* 2004;291:1081–91.
- Cai B, Small D, Ten Have T. Two-stage instrumental variable methods for estimating the causal odds ratio: analysis of bias. *Stat Med.* 2011;30:1809–24.
- Cai B, Hennessy S, Flory JH, Sha D, Ten Have TR, Small DS. Simulation study of instrumental variable approaches with an application to a study of the anti-diabetic effect of bezafibrate. *Pharmacoepidemiol Drug Saf.* 2012;21:114–20.
- Card D. *Using geographic variation in college proximity to estimate the return to schooling.* Toronto: University of Toronto Press; 1995. p. 201–22.
- Cheng J. Estimation and inference for the causal effect of receiving treatment on a multinomial outcome. *Biometrics.* 2009;65:96–103.
- Cheng J, Small D. Bounds on causal effects in three-arm trials with noncompliance. *J R Stat Soc Ser B.* 2006;68:815–36.
- Cheng J, Qin J, Zhang B. Semiparametric estimation and inference for distributional and general treatment effects. *J R Stat Soc Ser B Stat Methodol.* 2009a;71:881–904.
- Cheng J, Small D, Tan Z, Ten Have T. Efficient nonparametric estimation of causal effects in randomized trials with noncompliance. *Biometrika.* 2009b;96:19–36.
- Clarke P, Windmeijer F. Instrumental variable estimators for binary outcomes. *J Am Stat Assoc.* 2012;107:1638–52.
- Cole J, Norman H, Weatherby L, Walker A. Drug copayment and adherence in chronic heart failure: effect on costs and outcomes. *Pharmacotherapy.* 2006;26:1157–64.
- Cox D. *Planning of experiments.* New York: Wiley; 1958.
- Cuzick J, Sasieni P, Myles J, Tyler J. Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. *J R Stat Soc Ser B Methodol.* 2007;69:565–88.
- Davidson R, MacKinnon J. *Estimation and inference in econometrics.* New York: Oxford University Press; 1993.
- Demissie K, Rhoads G, Ananth C, Alexander G, Kramer M, Kogan M, Joseph K. Trends in preterm birth and neonatal mortality among blacks and whites in the United States from 1989 to 1997. *Am J Epidemiol.* 2001;154:307–15.
- Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res.* 2007;16:309–30.

- Durbin J. Errors in variables. *Rev Inst Int Stat.* 1954; 22:23–32.
- Fisher R. Design of experiments. Edinburgh: Oliver and Boyd; 1949.
- Freedman D. Statistical models: theory and practice. Cambridge: Cambridge University Press; 2009.
- Freedman D, Sekhon J. Endogeneity in probit response models. *Polit Anal.* 2010;18:138–50.
- Goedde H, Agarwal D, Fritze G, Meier-Tackmann D, Singh S, Beckmann G, Bhatia K, Chen L, Fang B, Lisker R. Distribution of ADH2 and ALDH2 genotypes in different populations. *Hum Genet.* 1992; 88:344–6.
- Goyal N, Zubizarreta J, Small D, Lorch S. Length of stay and readmission among late preterm infants: an instrumental variable approach. *Hosp Pediatr.* In press.
- Heckman J, Robb R. Alternative methods for evaluating the impacts of interventions: an overview. *J Econ.* 1985;30:239–67.
- Heckman J, Vytlacil E. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proc Natl Acad Sci.* 1999;96:4730–4.
- Hernán M, Robins J. Instruments for causal inference: an epidemiologist's dream? *Epidemiology.* 2006;17:360.
- Hernán M, Robins J. Causal inference; 2013.
- Hirano K, Imbens G, Rubin D, Zhou X. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics.* 2000;1:69–88.
- Ho V, Hamilton B, Roos L. Multiple approaches to assessing the effects of delays for hip fracture patients in the United States and Canada. *Health Serv Res.* 2000;34:1499–518.
- Ho D, Imai K, King G, Stuart E. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal.* 2007;15: 199–236.
- Hogan J, Lee J. Marginal structural quantile models for longitudinal observational studies with time-varying treatment. *Stat Sin.* 2004;14:927–44.
- Holland P. Causal inference, path analysis, and recursive structural equations models. *Sociol Methodol.* 1988;18:449–84.
- Hudgens M, Halloran M. Towards causal inference with interference. *J Am Stat Assoc.* 2008;103:832–42.
- Hunink M, Glasziou P, Siegel J, Weeks J, Pliskin J, Elstein A, Weinstein M. Making in health and medicine: integrating evidence and values. Cambridge: Cambridge University Press; 2001.
- Imbens G. Nonadditive models with endogenous regressors. New York: Cambridge University Press; 2007.
- Imbens G, Angrist J. Identification and estimation of local average treatment effects. *Econometrica.* 1994;62: 467–75.
- Imbens G, Rosenbaum P. Robust, accurate confidence intervals with weak instruments: quarter of birth and education. *J R Stat Soc Ser A.* 2005;168:109–26.
- Imbens G, Rubin D. Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann Stat.* 1997a;25:305–27.
- Imbens G, Rubin D. Estimating outcome distributions for compliers in instrumental variables models. *Rev Econ Stud.* 1997b;64:555–74.
- Inoue A, Solon G. Two-sample instrumental variables estimators. *Rev Econ Stat.* 2010;92:557–61.
- Joffe M. Administrative and artificial censoring in censored regression models. *Stat Med.* 2001;20:2287–304.
- Joffe M. Principal stratification and attribution prohibition: good ideas taken too far. *Int J Biostat.* 2011;7(1):1–22.
- Joffe M, Small D, Brunelli S, Ten Have T, Feldman H. Extended instrumental variables estimation for overall effects. *Int J Biostat.* 2008;4.
- Johnston S. Combining ecological and individual variables to reduce confounding by indication: case study – subarachnoid hemorrhage treatment. *J Clin Epidemiol.* 2000;53:1236–41.
- Kang H, Kreuels B, Adjei O, May J, Small D. The causal effect of malaria on stunting: a Mendelian randomization and matching approach, Working Paper.
- Karni E. A theory of medical decision making under uncertainty. *J Risk Uncertain.* 2009;39:1–16.
- Kaushal N. Do food stamps cause obesity? Evidence from immigrant experience. *J Health Econ.* 2007;26:968–91.
- Kelejian H. Two-stage least squares and econometric systems linear in parameters but nonlinear in the endogenous variables. *J Am Stat Assoc.* 1971;66:373–4.
- Kitcheman J, Adams C, Prevaiz A, Kader I, Mohandas D, Brookes G. Does an encouraging letter encourage attendance at psychiatric outpatient clinics? The Leeds PROMPTS randomized study. *Psychol Med.* 2008;38:717–23.
- Korn E, Baumrind S. Clinician preferences and the estimation of causal treatment differences. *Stat Sci.* 1998;13:209–35.
- Kramer M, Rooks Y, Pearson H. Growth and development in children with sickle-cell trait. *N Engl J Med.* 1978;299:686–9.
- Lawlor D, Harbord R, Sterne J, Timpson N, Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med.* 2008;27:1133–63.
- Little R, Yau L. Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psychol Methods.* 1998;3: 147–59.
- Loeys T, Goetghebeur E. A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. *Biometrics.* 2003;59:100–5.
- Lorch S, Baiocchi M, Ahlberg C, Small D. The differential impact of delivery hospital on the outcomes of premature infants. *Pediatrics.* 2012a.
- Lorch S, Kroelinger C, Ahlberg C, Barfield W. Factors that mediate racial/ethnic disparities in us fetal death rates. *Am J Public Health.* 2012b;102:1902–10.
- Malkin J, Broder M, Keeler E. Do longer postpartum stays reduce newborn readmissions? Analysis using instrumental variables. *Health Serv Res.* 2000;35: 1071–91.

- McClellan M, McNeil B, Newhouse J. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA*. 1994;272:859.
- Moreira M. A conditional likelihood ratio test for structural models. *Econometrica*. 1990;71:463–80.
- Muthen B. A structural probit model with latent variables. *J Am Stat Assoc*. 1979;74:807–11.
- Newman T, Vittinghoff E, McCulloch C. Efficacy of phototherapy for newborns with hyperbilirubinemia: a cautionary example of an instrumental variable analysis. *Med Decis Mak*. 2012;32:83–92.
- Neyman J. On the application of probability theory to agricultural experiments. *Stat Sci*. 1990;5:463–80.
- Nie H, Cheng J, Small D. Inference for the effect of treatment on survival probability in randomized trials with noncompliance and administrative censoring. *Biometrics*. 2011;67:1397–405.
- O'Malley A, Frank R, Normand S. Estimating cost-offsets of new medications: use of new antipsychotics and mental health costs for schizophrenia. *Stat Med*. 2011;30:1971–88.
- Okui R, Small D, Tan Z, Robins J. Doubly robust instrumental variables regression. *Stat Sin*. 2012;22:173–205.
- Owen A. *Empirical likelihood*. Boca Raton: Chapman & Hall/CRC; 2002.
- Pearl J. *Causality*. Cambridge: Cambridge University Press; 2009.
- Permutt T, Hebel J. Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics*. 1989;45:619–22.
- Phibbs C, Mark D, Luft H, Peltzman-Rennie D, Garnick D, Lichtenberg E, McPhee S. Choice of hospital for delivery: a comparison of high-risk and low-risk women. *Health Serv Res*. 1993;28:201.
- Pliskin J, Shepard D, Weinstein M. Utility functions for life years and health status. *Oper Res*. 1980;28:206–24.
- Poulsen R, Gadbury G, Allison D. Treatment heterogeneity and individual qualitative interaction. *Am Stat*. 2012;66:16–24.
- Qin J, Zhang B. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*. 1997;84:609–18.
- Rehan N. Growth status of children with and without sickle cell trait. *Clin Pediatr*. 1981;20:705–9.
- Robins J, Greenland S. A comment on Angrist, Imbens and Rubin: Identification of causal effects using instrumental variables. *J Am Stat Assoc*. 1996;91:456–8.
- Robins J, Tsiatis A. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun Stat Theory Methods*. 1991;20:2609–31.
- Rosenbaum P. *Observational studies*. New York: Springer; 2002.
- Rosenbaum P. *Design of observational studies*. New York: Springer; 2009.
- Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41–55.
- Rubin D. Estimating causal effects of treatments in randomized and non-randomized studies. *J Educ Psychol*. 1974;66:688–701.
- Rubin D. Formal modes of statistical inference for causal effects. *J Stat Plan Inference*. 1990;25:279–92.
- Saigal S, Stoskopf B, Feeny D, Furlong W, Burrows E, Rosenbaum P, Hoult L. Differences in preferences for neonatal outcomes among health care professionals, parents, and adolescents. *J Am Med Assoc*. 1999;281:1991–7.
- Sargan J. The estimation of economic relationships using instrumental variables. *Econometrica*. 1958;26:393–415.
- Sexton M, Hebel J. A clinical trial of change in maternal smoking and its effect on birth weight. *J Am Med Assoc*. 1984;251:911–5.
- Sham P. *Statistics in human genetics*. London: Arnold; 1998.
- Shea J. Instrument relevance in multivariate linear models: a simple measure. *Rev Econ Stat*. 1997;79:348–52.
- Shetty K, Vogt W, Bhattacharya J. Hormone replacement therapy and cardiovascular health in the United States. *Med Care*. 2009;47:600–6.
- Siddique Z. Partially identified treatment effects under imperfect compliance: the case of domestic violence. IZA Discussion Paper No. 4565. 2009.
- Small D. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J Am Stat Assoc*. 2007;102:1049–58.
- Small D, Rosenbaum P. War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *J Am Stat Assoc*. 2008;103:924–33.
- Sobel M. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *J Am Stat Assoc*. 2006;101:1398–407.
- Sommers BD, Beard CJ, Dahl D, D'Amico AV, Kaplan IP, Richie JP, Zeckhauser RJ. Decision analysis using individual patient preferences to determine optimal treatment for localized prostate cancer. *Cancer*. 2007;110:2210–7.
- Stock J, Wright J, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. *J Bus Econ Stat*. 2002;20:518–29.
- Tan Z. Regression and weighting methods for causal inference using instrumental variables. *J Am Stat Assoc*. 2006;101:1607–18.
- Tan Z. Marginal and nested structural models using instrumental variables. *J Am Stat Assoc*. 2010;105:157–69.
- Ten Have T, Elliott M, Joffe M, Zanutto E, Datto C. Causal models for randomized physician encouragement trials in treating primary care depression. *J Am Stat Assoc*. 2004;99:16–25.
- Terza J, Basu A, Rathouz P. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Health Econ*. 2008;27:527–43.
- Vansteelandt S, Bowden J, Babinezhad M, Goetghebuer E. On instrumental variables estimation of causal odds ratios. *Stat Sci*. 2011;26:403–22.

- Voight B, Peloso G, Orho-Melander M, Frikke-Schmidt R, Barbalic M, Jensen M, Hindy G, Hólm H, Ding E, Johnson T, et al. Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study. *Lancet*. 2012;380:572–80.
- Vytlacil E. Independence, monotonicity, and latent index models: an equivalence result. *Econometrica*. 2002; 70:331–41.
- Wehby G, Jugessur A, Moreno L, Murray J, Wilcox A, Lie R. Genetic instrumental variable studies of the impacts of risk behaviors: an application to maternal smoking and orofacial clefts. *Health Serv Outcome Res Methodol*. 2011;11:54–78.
- White H. Asymptotic theory for econometricians. 1984.
- Wooldridge J. On two stage least squares estimation of the average treatment effect in a random coefficient model. *Econ Lett*. 1997;56:129–33.
- Zelen M. A new design for randomized clinical trials. *N Engl J Med*. 1979;300:1242–5.





# Introduction to Causal Inference Approaches

# 22

Elizabeth A. Stuart and Sarah Naeger

## Contents

<b>Introduction</b> .....	524
Defining Causal Effects .....	524
Two Concepts: SUTVA and Assignment Mechanism .....	526
Careful Design .....	526
<b>Strategies for Estimating Causal Effects</b> .....	527
Randomized Experiments .....	527
Natural Experiments: Instrumental Variables .....	528
Regression Discontinuity .....	529
Difference-in-Difference and Interrupted Time Series Designs .....	530
Propensity Scores and Other Matching Methods .....	531
<b>Conclusions</b> .....	533
<b>References</b> .....	533

## Abstract

Many questions in health services research require causal estimates of the effects of policies or programs on a health outcome. Although randomized experiments are seen as the gold standard for estimating causal effects, randomization is often unfeasible and/or impractical or will not answer the question of interest. In those cases, rigorous

nonexperimental study designs can be used, as highlighted in this chapter. The chapter first takes care to carefully define the causal effects of interest and stresses the importance of careful study design. Overviews of four common nonexperimental study designs are then provided: instrumental variables, regression discontinuity, interrupted time series (and the related approach of difference in differences), and propensity score matching. An emphasis is on applications of these methods in health services research and the assumptions underlying each approach. The chapter concludes with open topics and suggestions for the conduct of studies aiming to estimate causal effects in health services research.

---

E. A. Stuart (✉)  
Department of Mental Health, Johns Hopkins Bloomberg  
School of Public Health, Baltimore, MD, USA  
e-mail: [estuart@jhsph.edu](mailto:estuart@jhsph.edu)

S. Naeger  
Behavioral Health Research and Policy, IBM Watson  
Health, Bethesda, MD, USA  
e-mail: [snaeger@jhsph.edu](mailto:snaeger@jhsph.edu)

## Introduction

Many questions in health services research require causal estimates of the effects of policies or programs on a health outcome, for example, the effects of expanding access to public health insurance on health (Finkelstein et al. 2004) or the effects of public reporting on quality of care in nursing homes (Werner et al. 2009). Either due to practical or ethical concerns, many of these questions cannot be answered with randomized experiments and require sophisticated nonexperimental methods instead. This is particularly true given recent interest in comparative effectiveness research and patient-centered outcomes research, which are both interested in examining the effects of interventions in real-world settings, among a broad set of patients, including those who may normally be excluded from randomized trials (Berger et al. 2009; Mullins et al. 2012; Oliver et al. 2009; Rosenberg 2009).

Like many other fields in the social sciences, health services researchers are continually faced with the challenge of making causal inferences from nonexperimental data. As stated by Escarce and Flood (2011) in an introduction to a special section of *Health Services Research* on causality, “Explicit in both definitions [of health services research, by AcademyHealth and the Agency for Healthcare Research and Quality] is the notion that health services researchers should strive to identify and estimate the causal effects on outcomes of interest of alternative organizational structures, management approaches, financing systems, provider practices, and personal choices regarding lifestyle and behavior. Without a focus on causal effects, it would be impossible to identify the most effective ways to achieve the outcomes we seek through clinical, management, or policy interventions” (2011, p. 394). At stake is the need to determine if programs, treatments, and prevention efforts are having a measurable impact on people and the populations being served in face of declining resources and funds.

Health services research is an inherently interdisciplinary field; researchers come to the field with varying levels of familiarity with various methods for estimating causal effects. Many

health researchers, in particular, come from academic traditions that emphasize one method over another (Dowd 2011). This chapter aims to provide an overview of methods for estimating causal effects, providing a brief introduction to the methods commonly used in health services research, including propensity scores, instrumental variables, and interrupted time series. For more information on the history behind some of these methods (and their use in health services research), see Dowd (2011) and O’Malley (2011), and for more description of the role of structural equation models for assessing causality more broadly, see, for example, Pearl (2011).

## Defining Causal Effects

To help clarify the concepts and goals of causal inference methods, consider a motivating example where interest is in examining whether access to public health insurance (such as Medicare) improves health outcomes. For simplicity, assume health outcomes are measured by self-reported health status at a particular point in time. For each individual, there are two “potential outcomes:” their self-reported health status if they are in the Medicare program (the “treatment” condition), denoted  $Y(1)$ , and their self-reported health status if they are not in Medicare (the “control” condition), denoted  $Y(0)$  (these could be indexed by  $i$  to denote individual  $i$ ; for simplicity we leave that out). The causal effect of Medicare enrollment is then a comparison of these two potential outcomes. When the outcome of interest is continuous, the comparison between the sets of potential outcomes is often a difference in means, for binary outcomes an odds ratio can be used.

The comparison of potential outcomes allows the definition of the causal effects of interest; they are an essential component to the causal inference framework. There are two other key components to this framework. First, the *treatment* or *exposure* of interest: a condition that could theoretically be given to or withheld from an individual (or a community). The treatment condition should be defined in relation to a control or reference condition. In some cases, the control condition

may be no treatment, and in others, it may be an established treatment. For example, if the research question of interest is on the impact of electronic health records on behavioral health screening rates in pediatric clinics, the “treatment” condition could be using electronic health records, and the “control” condition could be using paper health records (Hacker et al. 2012). Often the first step in clearly stating causal questions (and thus providing causal answers) is in clearly stating what is the intervention of interest and what is the appropriate comparison condition. In CER, for example, a researcher may have to decide whether to compare a particular treatment to another (e.g., one drug vs. another) or one drug versus “usual practice” (which may be a mix of treatments). Although concepts can be defined and some methods are available for cases where there are more than two treatment conditions of interest, for simplicity this chapter focuses on binary treatment (treatment vs. control or treatment 1 vs. treatment 2) comparisons.

The next key concept is the *units*: the entities that the treatment could be given to or withheld from at a particular point in time. Units can be individuals, medical clinics, or communities, but the units should correspond to the level of the treatment being evaluated. In the study of electronic health records mentioned above, the units would be pediatric clinics. In a study of a new therapy for diabetes, the units would be individual patients.

The treatment, units, and potential outcomes form the framework for causal inference. The “fundamental problem of causal inference,” however, is that only one of the potential outcomes for each individual or unit can be observed (Holland 1986). For individuals/units in the treatment condition,  $Y(1)$  is observed, and for individuals in the control group,  $Y(0)$  is observed; at a given point in time, each individual is either in Medicare or not. For the individuals in Medicare,  $Y(1)$  is observed and interest is in predicting what  $Y(0)$  would have been if they had not been in Medicare. Similarly, for individuals not in Medicare,  $Y(0)$  is observed, and interest is in predicting what  $Y(1)$  would have been had they been in Medicare. Causal inference can thus be thought of as a missing data problem,

where interest is in predicting these missing potential outcomes.

The causal effect of interest is the difference in potential outcomes ( $Y(1)$  and  $Y(0)$ ) for the *same* individual. The statistical problem of causal inference relates to how we can best predict those missing potential outcomes to make estimates. An important distinction is that these effects are defined in relation to potential outcomes and become the *estimand* of interest (i.e., the quantity we are interested in estimating), independent of what study design we might use to learn about them (e.g., randomized vs. nonrandomized). The estimand is the comparison of the potential outcomes that defines the causal effect of interest.

In prelude to concepts that will be discussed further below, there may be different estimands of interest, and different methods may estimate different estimands, as discussed further below. Two common estimands are the “average treatment effect on the treated” (ATT) and the “average treatment effect” (ATE). The ATE is the effect of some treatment if everyone in the population receives the treatment versus no one receiving the treatment. The ATT, in contrast, is the effect for the treatment group – the difference in average outcomes if everyone in the treatment group is treated and the average outcome if everyone in the treatment group actually receives the control (this is the “counterfactual” condition). Which estimand is of more interest will depend on the substantive question. For example, when investigating the effects of potentially harmful “treatments” (such as adolescent drug use), the ATT may be more relevant since that treatment would never be imposed on the full population; instead, interest is on what the effects are for those people who are actually drug users. In contrast, the ATE is a useful estimate when it is plausible that treatment could be disseminated to the entire population, for example, fluoride in the public water system. Note that in a randomized experiment, these quantities are equal in expectation, and so this distinction does not arise. Other methods, such as regression discontinuity and instrumental variables, estimate other “local average treatment effects,” which are effects for a particular subgroup of individuals and are discussed further

below. Imai et al. (2008) also differentiate “sample” versus “population” effects, a distinction not further discussed in this chapter.

As mentioned earlier, in most causal frameworks, the treatments of interest are thought of as (at least hypothetical) “interventions” that one can imagine giving or withholding. As stated by Holland (1986, p. 959), “No causation without manipulation.” In part this is to ensure that the estimand is clear and that everyone has the same understanding of what “treatment” versus “control” means. However, some of the methods discussed in this chapter have also been used to examine noncausal questions, such as to investigate racial disparities, using the framework of “balanced comparisons,” where we want to compare two groups that are as similar as possible on a set of observed characteristics. Zaslavsky et al. (2012) discuss these ideas in more detail. In this chapter, the focus is on the “effects of causes” rather than the “causes of effects,” as delineated by Holland (1986). In this way, interest is in questions regarding what are the effects of particular policies, interventions, or “treatments,” rather than broader (and perhaps less specific) questions about causal mechanisms or causal models more generally.

## Two Concepts: SUTVA and Assignment Mechanism

Since causal effects are estimated at the group level but potential outcomes are defined as individual level phenomena, several assumptions are required (Little and Rubin 2000). The first is the stable unit treatment value assumption (SUTVA). SUTVA has two components: first, that there is only one “version” of each treatment: that the “treatment” is well defined in that there are not two different types of treatments within the “treatment” condition, and second, that there is no interference in that the treatment assignment of one unit does not affect the potential outcomes of any other units. An implication of this assumption is that each unit has a unique potential outcome under the treatment and control conditions; i.e., if person  $i$  is treated,  $Y_i(1)$  is observed, and if person

$i$  is in the comparison group,  $Y_i(0)$  is observed (Holland 1986), and these quantities each take on a single value, regardless of other conditions, such as the treatment assignments of other individuals. This is known as “consistency” in the epidemiology literature (Cole and Frangakis 2009).

The assignment mechanism is the process by which individuals are assigned to receive treatment or not. In randomized experiments, the assignment mechanism is the randomization process; knowing the assignment mechanism frees the researcher from making any further assumptions about the distribution of the data. This is because, in a randomized experiment, treatment assignment is independent of the individual’s potential outcome. In observational studies, the researcher must infer the mechanism or process by which individuals end up in the treatment or comparison group. In the example above, the researcher would need to model the process through which some individuals receive Medicaid coverage and some do not. This relates back to the problem of missing data, in that the process that creates the missing potential outcomes must be accounted for when estimating causal effects (Greenland 2005; Little and Rubin 2000).

## Careful Design

One theme of this chapter is the importance of careful design. Randomized experiments have a particularly useful design; individuals are assigned to receive the treatment or control condition randomly. (The benefits of randomization will be discussed further below). When the assignment mechanism is known, it is possible to obtain unbiased estimates of treatment effects with no assumptions (for now assuming away any non-compliance or missing data).

In contrast, any nonexperimental study must rely on some (mostly untestable) assumptions. Those assumptions are discussed briefly below for each method and in more detail in the accompanying chapters. For that reason nonexperimental studies require smart choices, “choice as an alternative to control” in the words of Paul Rosenbaum (1999, 2005a) and thoughtful designs to isolate the effects

of the treatments of interest. In other words, when you can't randomize, make smart choices to yield robust causal inferences.

Many of these choices will involve selecting an appropriate control group, as stressed in Rosenbaum (2010) and Cook et al. (2008). The key feature of a randomized experiment is that it produces comparable or balanced treatment and control groups that lead to unbiased and consistent estimates of causal effects. Therefore, in terms of internal validity, when experimental designs are not available creating comparable treatment and control groups in observational studies is more important than creating samples that represent a population. A study by Lehman et al. (1987) on the long-term effects of the sudden and unexpected loss of a spouse or a child provides an example. Individuals who had either lost a spouse or child in a car accident in the 4–7 years prior to the study made up the treatment group. In order to isolate the bereavement effects the authors created a control group by identifying 7581 individuals through driver's license renewals and then matched one control subject to each treatment group member on gender, age, family income in 1976 (i.e., the time period before the crash), education level, and the number and ages of children (Lehman et al. 1987). By carefully creating balanced treatment and control groups, the authors were able to demonstrate that psychological distress was significantly greater in the treatment subjects (Lehman et al. 1987; Rosenbaum 2005a). In an example from road safety, Rosenbaum (2010) describes a study that was looking at the association of road features with accidents; the "treatment" conditions were accident sites, and the comparison conditions were sites exactly one mile prior to the accident at the same time as the accident, with the idea that the car in the accident passed by that site (with no problem) just before the accident, thus controlling for factors such as weather and characteristics of the drivers. Because of the need to rely on untestable assumptions, sensitivity analyses are particularly crucial in nonexperimental studies – assessing the robustness of results to other (plausible) assumptions and considering other possible designs.

This chapter aims to give researchers some tools to start thinking about those possible designs, outlining the basics of study designs with a focus on nonexperimental studies. Readers who are interested in learning more about the careful design of nonexperimental studies should refer to the discussion of threats to validity in Shadish et al. (2002) discussion of the importance of careful design and methods of design sensitivity in Rosenbaum (1999, 2010) and discussion of the role of design versus analysis in Rubin (2007).

---

## Strategies for Estimating Causal Effects

This section provides an overview of common study designs that aim to estimate causal effects. These descriptions are not meant to be fully detailed but rather to provide a broad understanding of the approach, when it can be used, and what its underlying assumptions are. Examples of how each design has been used in health services research are provided.

## Randomized Experiments

First formalized by Fisher (1926), randomized experiments are considered the gold standard of causal inference, since, as mentioned above, (when "clean") they yield unbiased estimates of treatment effects (at least for the sample at hand) with no additional assumptions. In contrast, all of the nonexperimental methods discussed below rely on at least some assumptions. Intuitively, randomization to treatment or control groups means that the groups are equivalent on everything at baseline, except which treatment they receive. This means that any difference in outcomes between groups can be attributed to the treatment and not to any preexisting differences. Mathematically it can be shown that the average potential outcomes observed in each group (treatment or control) provides an unbiased estimate of the average potential outcome

under that condition for the population (Neyman 1923, 1934).

The three key properties of randomized experiments that ensure estimates of causal effects are unbiased are as follows. First, the treatment assignment is “unconfounded” which means the randomization process is independent of the potential outcomes. Second, each individual or unit in the experiment has a positive probability of receiving each treatment condition (i.e., each person could potentially be in either the treatment or control group). And finally, the study is designed without any knowledge of the potential outcomes.

Examples of randomized experiments in health services research include the Oregon Medicaid Coverage experiment (Baicker and Finkelstein 2011). Researchers used a lottery to randomly allocate low-income adults between 19 and 64 years old to either receive Medicaid or be assigned to a waiting list for Medicaid. Although not originally implemented for this purpose, the lottery process allowed researchers to estimate the causal effects of Medicaid enrollment compared to being uninsured. Preliminary results for the study indicated that Medicaid coverage increases health-care use (Baicker and Finkelstein 2011).

However, as has been widely discussed (Gluud 2006; Marcus et al. 2012; Rothwell 2005), randomized trials do have their own complications. These include noncompliance, where people do not take their assigned treatments (Frangakis and Rubin 2002; Marasinghe and Amarasinghe 2007; Peduzzi et al. 1993), missing outcome data (Frangakis et al. 2007), worries that the people who enroll in a trial may be different from those of broader interest (Marcus et al. 2012; Zimmerman et al. 2005), and ethical concerns about randomization (Crawford et al. 2011; De Melo-Martín et al. 2011; Hughes 2009). Because of these concerns, nonexperimental studies are sometimes used to estimate the causal effects of “treatments,” interventions, or exposures of interest. We will see that many of these designs attempt to replicate key features of experiments.

## Natural Experiments: Instrumental Variables

In some cases researchers do not have power over the treatments individuals (or providers or communities) do or do not receive but can identify some naturally occurring randomness in who receives which treatment. These methods rely on finding an “instrument” that is (or can be thought of as) randomly assigned, affects the treatment individuals receive, but does not affect their outcomes directly. Instrumental variable designs are sometimes referred to as “encouragement designs” as the instrument can be thought of as something that encourages individuals to take the treatment of interest (or not). Examples of instrumental variables (IVs) in HSR include Bao et al. (2006), who, in examining the effect of providers giving smoking cessation advice, used whether or not the provider provided diet/nutrition or physical activity advice as an instrument. Linden and Adams (2006) use zip code as an instrument for participation in disease management programs, since not all geographic areas are covered by such programs. Geography is commonly used as an instrument, as it takes advantage of the fact that many medical treatments are more accessible in some geographic areas than others (e.g., McClellan et al. 1994).

IV methods essentially work by fitting two models: first, a model of treatment received as a function of the instrument and covariates and, second, a model of outcome as a function of treatment received and the covariates. The “exclusion restriction” (described further below) means that the instrument is “excluded” (not in) the second-stage model. Because these two equations are related (and the error terms therefore correlated), the models are generally fit using two-stage least squares models (Angrist and Imbens 1995, 1996).

There are two primary assumptions on which IV methods rely (in addition to the SUTVA assumption described above). The first is known as “monotonicity” and basically implies that there are no “defiers:” no people who go against the instrument in terms of what treatment they receive. In other words, no one who would take

the treatment if not “encouraged to” by the instrument but who would not take the treatment when “encouraged” to do so by the instrument. The second set of assumptions are what are known as the “exclusion restrictions.” These say that there is no effect of the instrument on individuals whose behavior is not changed by the instrument. In other words, there is no effect of the instrument on outcomes for people who would either always take the treatment (whether encouraged to or not by the instrument) or for people who would never take the treatment (whether encouraged to or not). This is sometimes stated as that there is “no direct effect” of the instrument on the outcomes; the only way the instrument can change outcomes is by changing the treatment that individuals receive. This assumption is often questionable.

To illustrate these two assumptions, consider treatment assignment and actual treatment status. In a randomized experiment, these two conditions are typically one and the same and are manipulated by the researcher. In the context of an IV design, the instrument influences (encourages) an individual’s treatment assignment, but other factors, such as individual-level covariates, influence compliance with the assignment (i.e., treatment status). The monotonicity assumption means that there is a positive correlation between treatment assignment and status. As an example, in the Long et al. (2005) study of the impact of Medicaid on improving access to care, treatment status was defined as being privately insured, having Medicaid coverage, or being uninsured. The four instrumental variables (i.e., the treatment assignment variables) included accessibility of private insurance, availability of public coverage, and family and community attitudes toward public assistance; under the monotonicity assumption, the influence of these variables can only increase the likelihood that an individual is privately insured or has Medicaid coverage. The exclusion restrictions require that accessibility of private insurance, availability of public coverage, and attitudes toward public assistance only influence insurance coverage and do not have any effect on health-care utilization directly (Long et al. 2005).

One important point about IV methods is that they technically estimate what is known as the

“local average treatment effect,” also known as the “complier average causal effect:” the effect of the treatment for the “compliers,” those individuals whose behavior is affected by the instrument and who will take the treatment when “told” to do so (when encouraged by the instrument) but not when not encouraged. In the example above, compliers would be individuals who seek out health insurance coverage when private or public options are available and there are positive attitudes toward public assistance, but who do not seek out insurance coverage when these are not operating (Long et al. 2005). The complier average causal effect is also known as a “marginal treatment effect” in the economics literature (Carneiro et al. 2011).

Another consideration when using IV methods is what is known as the “strength” of the instrument: how correlated the treatment assignment (instrument) is with the actual treatment status (the treatment received). A strong instrument is highly correlated with the actual treatment received. A weak instrument, in contrast, is only weakly associated with the actual treatment received (i.e., it is a poor predictor of treatment status). Weak instruments lead to reduced power and biased IV estimates (Bound et al. 1995).

## Regression Discontinuity

Introduced by Thistlethwaite and Campbell (1960), regression discontinuity (RD) is a particularly strong nonexperimental design that can be used when the treatment of interest is assigned on the basis of some “assignment variable” and cutoff. For example, individuals with cholesterol levels above 200 may be put into a care management program, whereas those with lower cholesterol are not given access to the program. The idea is to compare individuals just below and just above the cutoff, who should be otherwise similar but with one group receiving the treatment of interest and the other not. The analysis examines whether there is a “discontinuity” in the outcome variable at the cutoff, which would indicate an effect of the treatment. RD is similar to randomized experiments in that the assignment

mechanism is known, and that is what allows us to obtain reliable treatment effect estimates.

Examples of RD designs in HSR include studies of disease management programs (Linden and Adams 2006), which may be a particularly good setting for RD since eligibility for the program is often determined by clinical measures to ensure that the program is provided to those most in need. RD designs may also be appropriate when resources permit serving only a portion of the population and those most in need are served first, in which case there may be a discontinuity at the point at which resources are gone. This sort of idea was used by Ludwig and Miller (2006) in estimating the effects of Head Start, who used a discontinuity in grant writing support for original Head Start grants, with that support given to the 300 poorest counties in the country.

This section highlights a few assumptions and requirements of the RD method, as described by Trochim (1984). First, for the most basic form of RD analyses, the cutoff must be followed. (In fact, more advanced “fuzzy” RD designs can be used if there is some “noncompliance,” where some individuals who were eligible didn’t receive the treatment and some individuals who were not eligible did receive it; see Imbens and Lemieux (2008)). Second, accurate modeling of the relationship between the assignment variable and the outcome is crucial, for example, allowing for a nonlinear relationship or other flexible models. Ludwig and Miller (2006) consider a variety of functional forms in order to assess sensitivity to the model. Third, the sample size around the cutoff must be large enough to fit those models reliably and with sufficient precision. Goldberger (2008) indicates that sample sizes 2.75 times larger than would be required for adequate power in an RCT are needed for RD designs.

Threats to the validity of RD designs include cases where the assignment variable is manipulated because of the treatment assignment process, for example, clinicians manipulating the assignment variable so that patients they want to have participate in the program are seen as eligible.

Similar to the idea of the “local average treatment effect” in instrumental variables analyses, a limitation of the RD design is that it formally

estimates the effect only for those just around the cutoff. This arguably, however, is the group for whom the effect is most relevant as presumably these are the people who may or may not receive the intervention (i.e., those with very high or very low scores may not be reasonable candidates for the intervention under investigation). The design is not appropriate for estimating the effect of the treatment for individuals with assignment variables nowhere near the cutoff.

Sensitivity analyses are important in RD designs. Important sensitivity analysis options include “zero checks” where the analysis is repeated using fake cutoffs, to confirm that no “effect” is seen there, as well as assessing sensitivity to the model specification, as mentioned above. It is also important to note that RD designs only work when the treatment was in fact given out on the basis of the cutoff variable; they cannot be used in a “post hoc” way if that was not in fact how the treatment was administered.

For more information on RD designs, see Imbens and Lemieux (2008). Wong et al. (2012) provides discussion of extensions for studies with multiple assignment variables or cutoff points. The appendix of Linden et al. (2006) provides a relatively easy to read description of the actual models run to estimate effects in RD designs.

## **Difference-in-Difference and Interrupted Time Series Designs**

A common approach for estimating the effects of discrete policy changes is interrupted time series (ITS) analyses (or a simplified version, difference in differences). These methods rely on sophisticated before-after analyses to compare observed trends in the presence of an intervention with the time trends that would have been predicted in the absence of the intervention. Briefly, at its most basic level, the treatment effect is estimated by modeling the “outcome” of interest in the pre-period, extrapolating that model fit to the post period, and estimating the effect as the difference between the expected values (from that model fit) and the observed values. Interest may be in determining whether the intervention leads



to a jump at the time of implementation (an “interruption”) or also possibly a change in the slope of the time series trend. The simpler model, difference in differences, essentially collapses the “pre”- and “post”-time periods, comparing the change in the outcome from pre-intervention to post-intervention between the intervention group and a comparison group (see O’Malley et al. 2006, for an example).

ITS designs abound in HSR. Campbell et al. (2009) use an ITS design to evaluate the effect of pay for performance on the quality of care in primary care practices. They collected data from 42 primary care practices at two time points prior to the policy implementation and at two time points post policy implementation. Data on patient care, patient perception of access to care, and continuity of case were used to determine if care for patients with asthma, diabetes, or coronary heart disease improved after the pay-for performance plan was implemented (Campbell et al. 2009). As another example, Andersson et al. (2006) use interrupted time series to investigate the effects of changes in the pharmaceutical reimbursement schedule in Sweden on costs and volumes of pharmaceuticals.

ITS methods are most useful when (1) there is an abrupt policy change (e.g., a new law) and (2) there is sufficient pre-change data with which to estimate trends reliably. And while not required, a comparison group that did not experience the policy change can be very useful in terms of providing accurate results. In particular, comparative interrupted time series designs are particularly strong since they provide information on trends in the post-period in comparison units (e.g., states) that did not experience the policy change. Without such a comparison group, the results are more reliant on the time series models themselves; this can be misleading, for example, when there are strong time trends even in the absence of the intervention (e.g., increasing test scores in education research). Linden and Adams (2010) provide an example of combining ITS methods with propensity score weighting (discussed more below) to create a particularly good comparison group for the ITS analysis. Their study estimates the effect of California’s Proposition 99, which in 1988

raised the cigarette excise tax by 25 cents per pack in order to fund anti-smoking initiatives across the state. Similarly, O’Malley et al. (2006) discuss the careful choice of comparison groups in the context of a difference-in-difference analysis of interventions aimed to encourage the use of generic drugs.

An important consideration in ITS models is serial correlation and accounting for the correlation of measures across time. Since the error terms in the regression models will likely be correlated, it is important to test for autocorrelation using a test such as Durbin’s test (Durbin 1970) and appropriately model that autocorrelation, for example, using AR-1 models (Mills 1990). See Wagenaar et al. (2009) for an example.

## Propensity Scores and Other Matching Methods

The final nonexperimental method discussed is that of propensity score methods, which broadly are used to equate two groups and ensure that the treatment effect is being estimated among treated and comparison subjects who are otherwise similar. In this respect, propensity score methods aim to replicate two key features of a randomized experiment: (1) create groups that are similar on background characteristics (or at least the observed ones) and (2) the outcome is not used in setting up the “design” of the study. The propensity score itself is defined as the probability of receiving the treatment and is estimated by modeling treatment status as a function of baseline characteristics. Because of the properties of the propensity score (Rosenbaum and Rubin 1983), they are particularly useful for creating groups that look similar with respect to a large set of characteristics; researchers can then match, subclassify, or weight using just the propensity score itself, rather than having to deal with each variable individually. See Stuart (2010) for more details.

Propensity score methods involve two stages: (1) fitting a propensity score model and (2) using those propensity scores to create balanced samples. Common propensity score estimation methods include logistic regression as well as

nonparametric methods such as boosted CART or random forests (Lee et al. 2010). Common methods of using propensity scores include matching, subclassification, and weighting (Stuart 2010). Matching aims to find one (or more) matched comparison subjects for each treated subject; most matching methods estimate the average treatment effect on the treated. Subclassification groups subjects into small sets with similar propensity score values (e.g., by the deciles of the propensity score distribution). Weighting uses ideas similar to survey weighting, where individuals are weighted by functions of the propensity score. The most common weighting approach, known as inverse probability of treatment weighting (IPTW), weights the treatment and comparison group up to the combined sample and estimates the average treatment effect.

Examples of propensity score methods in HSR include Werner et al. (2009), which used propensity score matching combined with ITS to estimate the effect of public reporting of nursing home quality measures on quality of care. In that work, nursing home residents before the policy change were matched to residents after the change, to ensure a comparable case-mix over time.

The key assumption underlying propensity score methods is that of unconfounded treatment assignment, also known as “no hidden bias” or “strong ignorability” (Rosenbaum and Rubin 1983). This assumes that there are no unobserved differences between the treatment and comparison groups, given the observed covariates. It is crucial to think carefully about the validity of this assumption in any given study and whether the data at hand are sufficient in terms of the variables observed and available for the propensity score model. Propensity score methods generally work best when there is a large set of variables on which to match (demographics are generally not sufficient) and in particular when baseline measures of the outcomes are available (Steiner et al. 2010). For example, when assessing self-reported health as an outcome, it is important to have a pre-intervention baseline measure of self-reported health status (or of variables highly correlated with such a measure).

The validity of the unconfoundedness assumption will also likely depend on the setting, in that, for example, the assumption may be more believable when the treatment assignment is made by an external party on the basis of observed characteristics (e.g., a physician, using medical records that researchers have access to, or a teacher selecting students on the basis of test scores), as compared to studies where individuals self-select into treatments, in ways that may be related to unobserved factors such as motivation or an individual’s own assessment of how effective they think the treatment will be for them. This may be a particular concern in many HSR studies that rely on publicly available data that was not originally designed to answer the question of current interest and where the variables that predict treatment assignment may not be observed (e.g., clinical measures may not be available in a claims file). In this case, one strategy is to follow the strategy recommended by Rosenbaum (2010), which is to deal as well as possible with the observed characteristics (“overt bias”) using methods such as propensity scores, and follow that with an analysis of sensitivity to unobserved confounding (“hidden bias”). While not yet fully disseminated, there are a number of sensitivity analyses that can be done to assess sensitivity to an unobserved confounder; these analyses ask the question “How strongly related to treatment assignment and the outcome would an unobserved confounder have to be to make my observed treatment effect go away?” See Rosenbaum (2005b), Schneeweiss (2006), and Liu et al. (2013) for more background on these methods, including links to software to implement these approaches. A second challenge with the sorts of large datasets often used in HSR is that they are often cross-sectional, without repeated measures of individuals. The challenge in this setting is to identify which variables can be safely considered “pretreatment” and therefore matched on, versus those that may have been affected by the treatment and thus should not be matched on (known generally as “posttreatment bias”; Imai et al. 2010).

## Conclusions

Health services research involves answering many important questions, many of which are causal, aiming to understand what are the effects of particular policies or programs. This chapter has aimed to provide an overview of the methods available to answer such causal questions, with a goal of providing an overview of a variety of methods so that researchers know the breadth of methods available. There is no one single method that will be best for answering every possible study; the method needs to be chosen in the context of any given research question. For example, regression discontinuity and interrupted time series designs can work very well when the data arise from such settings but cannot be used when the data does not.

In many cases, researchers may be left selecting between instrumental variables and propensity score approaches. Again, the optimal choice will depend on the particular question: whether a plausible instrument exists and whether the assumption of unconfounded treatment assignment is believable. In brief: which method's assumptions are more likely satisfied? How much is known about the process that determined who was treated and who was not and what are the characteristics associated with that choice observed? This decision may also relate to who was making the treatment decisions: when an individual is self-selecting the treatment, there may be more concern about the plausibility of unconfounded treatment assignment. In contrast, if another decision maker (e.g., a physician) is making the decision based on variables that are largely observed, unconfounded treatment assignment may be more reasonable.

At a minimum, analyses of sensitivity to those assumptions should be done, such as the sensitivity analyses discussed above for propensity score methods, as well as methods that help assess the validity of the exclusion restrictions in IV (Greenland 2000). In some cases, both analyses may be plausible, and doing the analysis both ways may help provide a sense for the robustness of the results.

There are also many important directions for further research in the field of causal inference relevant for HSR. These include modifications of existing methods to handle very large datasets such as electronic health records or medical claims (e.g., the high-dimensional propensity score approach of Schneeweiss et al. (2009)). A second challenge in the coming years is to identify methods to better detect treatment effect heterogeneity. Some progress has been made recently, but this will be an important area for further work, especially as there is increasing interest in determining “what works for whom” and under what settings treatments are effective.

Methods for causal inference are an important, and expanding, set of tools for health services researchers. Answering causal questions well will ultimately help us better understand how to improve health and health care for people across the globe.

---

## References

- Andersson K, Petzold MG, Sonesson C, Lonnroth K, Carlsten A. Do policy changes in the pharmaceutical reimbursement schedule affect drug expenditures? Interrupted time series analysis of cost, volume, and cost per volume trends in Sweden 1986–2002. *Health Policy*. 2006;79:231–43.
- Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J Am Stat Assoc*. 1995;90(430): 431–42. <https://doi.org/10.1080/01621459.1995.10476535>.
- Angrist JD, Imbens GW. Identification of causal effects using instrumental variables. *J Am Stat Assoc*. 1996;91:444–55.
- Baicker K, Finkelstein A. The effects of Medicaid coverage – learning from the Oregon experiment. *N Engl J Med*. 2011;365(8):683–5.
- Bao Y, Duan N, Fox SA. Is some provider advice on smoking cessation better than no advice? An instrumental variable analysis of the 2001 National Health Interview Survey. *Health Serv Res*. 2006;41(6):2114–35.
- Berger ML, Mamdani M, Atkins D, Johnson ML. Good research practices for comparative effectiveness research: defining, reporting and interpreting non-randomized studies of treatment effects using secondary data sources: the ISPOR good research practices for retrospective database analysis task force report – part I. *Value Health*. 2009;12(8):1044–52.
- Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between

- the instruments and the endogenous explanatory variable is weak. *J Am Stat Assoc.* 1995;90(430): 443–50.
- Campbell SM, Reeves D, Kontopantelis E, Sibbald B, Roland M. Effects of pay for performance on the quality of primary care in England. *N Engl J Med.* 2009; 361(4):368–78. <https://doi.org/10.1056/NEJMsa0807651>.
- Carneiro P, Heckman JJ, Vytlacil EJ. Estimating marginal returns to education. *Am Econ Rev.* 2011;101(6): 2754–81.
- Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology.* 2009;20(1):3–5.
- Cook TD, Shadish WR, Wong VC. Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. *J Policy Anal Manage.* 2008;27(4): 724–50. <https://doi.org/10.1002/pam.20375>.
- Crawford MJ, Thana L, Methuen C, Ghosh P, Stanley SV, Ross J, Gordon F, et al. Impact of screening for risk of suicide: randomized controlled trial. *Br J Psychiatry.* 2011;198(5):379–84.
- De Melo-Martín I, Sondhi D, Crystal RG. When ethics constrains clinical research: trial design of control arms in “greater than minimal risk” pediatric trials. *Hum Gene Ther.* 2011;22(9):1121–7.
- Dowd BE. Separated at birth: statisticians, social scientists, and causality in health services research. *Health Serv Res.* 2011;46(2):397–420.
- Durbin J. Testing for serial correlation in least-squares regression when some of the Regressors are lagged dependent variables. *Econometrica.* 1970;38(3): 410–21.
- Escarce JJ, Flood AB. Introduction to special section: causality in health services research. *Health Serv Res.* 2011;46(2):394–6. <https://doi.org/10.1111/j.1475-6773.2011.01255.x>.
- Finkelstein EA, Fiebelkorn IC, Wang G. State-level estimates of annual medical expenditures attributable to obesity\*. *Obes Res.* 2004;12(1):18–24. <https://doi.org/10.1038/oby.2004.4>.
- Fisher R. The arrangement of field experiments. *Journal of Ministry of Agriculture.* 1926;33:500–13.
- Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics.* 2002;58(1):21–9.
- Frangakis CE, Rubin DB, An MW, MacKenzie E. Principal stratification designs to estimate input data missing due to death. *Biometrics.* 2007;63(3): 641–9.
- Glued LL. Bias in clinical intervention research. *Am J Epidemiol.* 2006;163(6):493–501. <https://doi.org/10.1093/aje/kwj069>.
- Goldberger A. Selection bias in evaluating treatment effects: some formal illustrations. In: *Modelling and evaluating treatment effects in econometrics, Advances in econometrics.* Bingley: Emerald Group Publishing Limited; 2008. p. 1–31.
- Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol.* 2000;29(4):722–9.
- Greenland S. Epidemiologic measures and policy formulation: lessons from potential outcomes. *Emerging Themes in Epidemiology.* 2005;2(1):5.
- Hacker K, Penfold R, Zhang F, Soumerai SB. Impact of electronic health record transition on behavioral health screening in a large pediatric practice. *Psychiatr Serv.* 2012;63(3):256–61.
- Holland PW. Statistics and causal inference. *J Am Stat Assoc.* 1986;81(396):945–60.
- Hughes JR. Ethical concerns about non-active conditions in smoking cessation trials and methods to decrease such concerns. *Drug Alcohol Depend.* 2009;100(3):187–93.
- Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. *Stat Sci.* 2010;25(1):51–71.
- Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *J R Stat Soc Ser A Stat Soc.* 2008;171(2): 481–502.
- Imbens GW, Lemieux T. Regression discontinuity designs: a guide to practice. *J Econ.* 2008;142(2):615–35.
- Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med.* 2010; 29(3):337–46.
- Lehman DR, Wortman CB, Williams AF. Long-term effects of losing a spouse or child in a motor vehicle crash. *J Pers Soc Psychol.* 1987;52(1):218–31.
- Linden A, Adams JL. Evaluating disease management programme effectiveness: an introduction to instrumental variables. *J Eval Clin Pract.* 2006;12(2):148–54. <https://doi.org/10.1111/j.1365-2753.2006.00615.x>.
- Linden A, Adams JL, Roberts N. Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design. *J Eval Clin Pract.* 2006;12(2):124–31.
- Linden A, Adams JL. Using propensity score-based weighting in the evaluation of health management programme effectiveness. *J Eval Clin Pract.* 2010;16(1):175–9.
- Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health.* 2000;21:121–45. <https://doi.org/10.1146/annurev.publhealth.21.1.121>.
- Liu W, Kuramoto SK, Stuart EA. An introduction to sensitivity analysis for unobserved confounding in non-experimental prevention research. *Prev Sci.* 2013;14(6):570–80. PMID:3800481.
- Long SK, Coughlin T, King J. How well does medicaid work in improving access to care? *Health Serv Res.* 2005;40(1):36–58. <https://doi.org/10.1111/j.1475-6773.2005.00341.x>.
- Ludwig J, Miller DL. Does head start improve children’s life chances? Evidence from a regression discontinuity design. Institute for the Study of Labor (IZA). 2006. Retrieved from <http://ideas.repec.org/p/iza/izadps/dp2111.html>

- Marasinghe JP, Amarasinghe AAW. Noncompliance in randomized controlled trials [4]. *CMAJ*. 2007; 176(12):1735.
- Marcus SM, Stuart EA, Wang P, Shadish WR, Steiner PM. Estimating the causal effect of randomization versus treatment preference in a doubly randomized preference trial. *Psychol Methods*. 2012;17(2):244–54.
- McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA*. 1994;272:859–66.
- Mills TC. *Time series techniques for economists*. Cambridge: Cambridge University Press; 1990.
- Mullins CD, Abdulhalim AM, Lavallee DC. Continuous patient engagement in comparative effectiveness research. *JAMA J Am Med Assoc*. 2012;307(15):1587–8.
- Neyman J. On the application of probability theory to agricultural experiments. *Essay on principles*. *Stat Sci*. 1923;5(4):465–80.
- Neyman J. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J R Stat Soc*. 1934;97:558–606.
- Oliver S, Armes DG, Gyte G. Public involvement in setting a national research agenda: a mixed methods evaluation. *Patient Patient-Cent Outcomes Res*. 2009;2(3):179–90.
- O’Malley AJ. Commentary on Bryan Dowd’s paper “Separated at birth: statisticians, social scientists, and causality in health services research”. *Health Serv Res*. 2011;46(2):430–6.
- O’Malley AJ, Frank RG, Kaddis A, Rothenberg BM, McNeil BJ. Impact of alternative interventions on changes in generic dispensing rates. *Health Serv Res*. 2006;41(5):1876–94.
- Pearl J. Statistics and causality: Separated to reunite – commentary on Bryan Dowd’s “Separated at birth”. *Health Serv Res*. 2011;46(2):421–9.
- Peduzzi P, Wittes J, Detre K, Holford T. Analysis as-randomized and the problem of non-adherence: an example from the veterans affairs randomized trial of coronary artery bypass surgery. *Stat Med*. 1993;12(13): 1185–95. <https://doi.org/10.1002/sim.4780121302>.
- Rosenbaum PR. Choice as an alternative to control in observational studies. *Stat Sci*. 1999;14(3):259–304.
- Rosenbaum PR. Observational study. In: Everitt B, Howell D, editors. *Encyclopedia of statistics in behavioral science*. Chichester: Wiley; 2005a.
- Rosenbaum PR. Sensitivity analysis in observational studies. In: Everitt BS, Howell DC, editors. *Encyclopedia of statistics in behavioral science*, vol. 4. Chichester: Wiley; 2005b. p. 1809–14.
- Rosenbaum PR. *Design of observational studies*, Springer series in statistics. New York: Springer; 2010.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
- Rosenberg L. Comparative effectiveness research: making it work for those we serve. *J Behav Health Serv Res*. 2009;36(3):283–4.
- Rothwell PM. External validity of randomised controlled trials? To whom do the results of this trial apply?? *Lancet*. 2005;365(9453):82–93.
- Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007; 26(1):20–36.
- Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf*. 2006;15(5):291–303.
- Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20(4):512–22.
- Shadish WR, Cook TD, Campbell DT. *Experimental and quasi-experimental designs for generalized causal inference*. 2nd ed. Belmont: Wadsworth Publishing; 2002.
- Steiner PM, Cook TD, Shadish WR, Clark MH. The importance of covariate selection in controlling for selection bias in observational studies. *Psychol Methods*. 2010;15(3):250–67.
- Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1–21.
- Thistlethwaite DL, Campbell DT. Regression-discontinuity analysis: an alternative to the ex post facto experiment. *J Educ Psychol*. 1960;51(6):309–17.
- Trochim W. *Research design for program evaluation; the regression-discontinuity design*. Beverly Hills: Sage; 1984.
- Wagenaar AC, Maldonado-Molina MM, Wagenaar BH. Effects of alcohol tax increases on alcohol-related disease mortality in Alaska: time-series analysis from 1976 to 2004. *Am J Public Health*. 2009; 99(8):1464–70.
- Werner RM, Konetzka RT, Stuart EA, Norton EC, Polsky D, Park J. Impact of public reporting on quality of Postacute care. *Health Serv Res*. 2009;44(4): 1169–87. <https://doi.org/10.1111/j.1475-6773.2009.00967.x>.
- Wong VC, Steiner PM, Cook TD. Analyzing regression-discontinuity designs with multiple assignment variables: a comparative study of four estimation methods. *J Educ Behav Stat*. 2012; <https://doi.org/10.3102/1076998611432172>.
- Zaslavsky AM, Ayanian JZ, Zaborski LB. The validity of race and ethnicity in enrollment data for medicare beneficiaries. *Health Serv Res*. 2012;47(3 Part 2): 1300–21.
- Zimmerman M, Chelminski I, Posternak MA. Generalizability of antidepressant efficacy trials: differences between depressed psychiatric outpatients who would or would not qualify for an efficacy trial. *Am J Psychiatr*. 2005;162(7):1370–2.



# Measurement of Patient-Reported Outcomes of Health Services

# 23

Joseph C. Cappelleri and Andrew G. Bushmakin

## Contents

<b>Introduction</b> .....	538
<b>Research Basis and Goals</b> .....	539
Background and Rationale .....	539
Research Objectives .....	540
<b>Selection of Subjects</b> .....	541
<b>Longitudinal Designs</b> .....	542
Event- or Condition-Driven Designs .....	542
Time-Driven Designs .....	543
Timing of the Initial PRO Assessment .....	543
Timing of Follow-Up PRO Assessments .....	544
Frequency of Evaluations .....	544
<b>Selection and Evaluation of the Measurement Instrument</b> .....	545
Step 1: Formulating Study Objectives .....	545
Step 2: Developing or Selecting an Instrument .....	546
Step 3: Developing Data Collection Strategies .....	549
Step 4: Analyzing Data .....	551
Step 5: Reporting Data .....	554
Interpreting Study Findings .....	555
<b>References</b> .....	556

## Abstract

A patient-reported outcome (PRO) is any report on the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else. The measurement of PROs should

address key protocol elements that include the rationale for the specific aspect of PRO being measured, explicit research objectives and endpoints, strategies for minimizing the exclusion of subjects from the study, rationale for timing of assessments and off-study rules, rationale for instruction selection, details for administration of PRO assessments to minimize bias and missing data, sample size estimation, and analytic plan. Another key element involves the measurement properties of a PRO. These protocol

J. C. Cappelleri (✉) · A. G. Bushmakin  
Global Product Development, Pfizer Inc, Groton, CT, USA  
e-mail: [joseph.c.cappelleri@pfizer.com](mailto:joseph.c.cappelleri@pfizer.com);  
[andrew.g.bushmakin@pfizer.com](mailto:andrew.g.bushmakin@pfizer.com)

elements are central to this chapter as they relate to the design and measurement of PROs. These elements are discussed and framed within the five characteristics that tend to be associated with PROs: missing and incomplete data, psychometric validation, interpretation, multiple testing, and longitudinal analysis. Special consideration is given for developing a PRO measurement strategy in a regulatory context where the intent is to have a labeling claim on a PRO.

---

## Introduction

A patient-reported outcome (PRO) is any report on the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else (Food and Drug Administration 2009). Patient-reported outcome is an umbrella term that includes a whole host of subjective outcomes such as pain, fatigue, depression, aspects of well-being (e.g., physical, functional, psychological), treatment satisfaction, health-related quality of life, and physical symptoms such as nausea and vomiting. Patient-reported outcomes are often relevant in studying a variety of conditions – including pain, erectile dysfunction, fatigue, migraine, mental functioning, physical functioning, and depression – that cannot be assessed adequately without a patient's evaluation and whose key questions require patient's input on the impact of a disease or a treatment. After all, who knows better than the patient herself/himself? To be useful to patients and other decision-makers (e.g., physicians, regulatory agencies, reimbursement authorities), who are stakeholders in medical care, PRO must undergo a validation process to confirm that it is reliably measuring what it is intended to measure.

In general the same clinical trial design principles that apply to directly assessable clinical endpoint measures, like blood pressure, also apply to PROs. Although not necessarily unique to PROs, at least five characteristics tend to be associated with PROs (Fairclough 2004). One characteristic is, by definition, PROs require the subject's (patient's) active participation, resulting in the

potential for missing data from not only missed assessments on an entire PRO but also non-response of some items on a PRO used in a study. A second characteristic is that being subjective and not a so-called “hard” endpoint like death, PROs require their measurement properties to be assessed, leading to additional steps of validation (reliability and validity) prior to their analysis on treatment effect. A third characteristic, related to the second one, is that the interpretation of PROs may require methods that can enrich and enhance their interpretation. A fourth characteristic is that most PROs are multidimensional and hence produce multiple scores on various aspects of what is being measured, engendering multiple comparisons and testing of outcomes that need to be methodologically and statistically addressed. The fifth characteristic is that the outcomes are generally repeated over time, calling for methods that effectively handle longitudinal data in the context of the research question.

Identifying which components of a PRO are relevant to measuring the impact of a disease and its treatment is essential to good study design and subsequent scientific scrutiny. Successful measurement of PROs begins with the development of a protocol to provide a recipe for the conduct of the study. The protocol provides not only key elements of the study design but also provides the scientific rationale and planned analysis for the study, which are inextricably linked to study design.

Because the validation of PROs is an ongoing process, multiple protocols with each having its specific purpose may often be necessary. A protocol for a study, be it a clinical trial or a method study, should contain several essential elements. A clinical trial protocol should describe the following: the rationale for the specific aspect of PRO being measured, explicit research objectives and endpoints, strategies for minimizing the exclusion of subjects from the study, rationale for timing of assessments and off-study rules, rationale for instruction selection, details for administration of PRO assessments to minimize bias and missing data, sample size estimation, and analytic plan. A method study protocol involves by definition methodological considerations, such as which measurement properties of a PRO will be

tested, and these considerations will define the design of the study. For example, if an objective is to obtain test-retest reliability data, data should be collected at least on two occasions. Contrary to a clinical trial design, which includes a pre-selected diseased population at baseline, method studies may not involve any treatment and may include a variety of subjects from healthy to severely ill for whom a PRO is designed to assess.

The aforementioned protocol elements are central to this chapter as they relate to the design and measurement of PROs. These elements are discussed and framed within the five characteristics that tend to be associated with PROs: missing data, validation, interpretation, multiple testing, and handling of longitudinal data.

Specifically, Section “[Research Basis and Goals](#)” covers the research basis surrounding PROs, with focus on the background and rationale and also on research objectives. Section “[Selection of Subjects](#)” centers on selection of subjects. Section “[Longitudinal Designs](#)” focuses on longitudinal designs. It discusses event- or condition-driven designs, time-driven designs, timing of the initial PRO assessments, timing of the follow-up PRO assessments, and frequency of evaluation. Section “[Selection and Evaluation of the Measurement Instrument](#)” concentrates on the selection and evaluation of the measurement instrument: formulating study objectives, developing or selecting an instrument (its relevance, psychometric properties, and feasibility), developing data collection strategies, analyzing data (multiple testing, missing data), reporting data, and interpreting study findings. Moreover, in this chapter, special consideration is given for developing a PRO measurement strategy in a regulatory context where the intent is to have a labeling claim on a PRO.

---

## Research Basis and Goals

### Background and Rationale

Providing sufficient background and rationale to justify the resources required for an investigation of PROs will contribute to the success of the investigation. The background to why PROs are

of relevance in assessing outcomes of interest in relation to the disease, as well as the characteristics of the patient population under consideration, needs to be described and linked to previous research and to the planned treatment. The reason for using the PRO component in relation to the research question needs to be lucid, and the PROs need to be clearly defined in the study (Wiklund 2004). A rationale should be given for not only why a PRO is being studied but also which specific aspect of a PRO is central and especially worthwhile.

Inherent in PROs is its ability to assist in providing a better understanding of disease and treatment outcomes from the patient’s perspective, and PROs do so by translating clinical improvement, stability, or deterioration into patient-centered outcomes. As such, PROs represent a unique indicator of the impact of disease and its treatment by enabling physicians and other health-care professionals to rely significantly on patient reports in evaluating disease activity and symptoms.

In the management and monitoring of certain chronic conditions – such as arthritis, neuropathic pain, irritable bowel syndrome, sexual dysfunction, and chronic obstructive pulmonary disease – PROs have become the central outcomes of choice. In other chronic diseases, such as cancer and cardiovascular disease, increased attention has been paid to PROs in order to highlight the humanistic side of the disease and its treatment. In oncology studies, for instance, the impact of treatment on survival and tumor shrinkage is often accompanied by and weighted against the impact of the treatment on aspects of a patient’s health-related quality of life, for example, the impact of chemotherapy on toxicity and adverse events. Since 1985, the FDA has recommended patient-centered evaluations in clinical trials relating to cancer research (Johnson and Temple 1985).

The rationale for measuring PROs needs to be made explicit in the planning and documenting of clinical trials in order to put forward labeling or promotional claims on PROs. From an industry and regulatory perspective, a well-defined and reliable PRO instrument in suitably designed investigations can be used to support a claim provided that the medical product labeling of the



claim is consistent with the suitably documented measurement capability of the instrument (Food and Drug Administration 2009). ("Instrument," as defined here, refers to a questionnaire plus all the information and documentation that supports its use, including the method of administration, instructions for administration, the scoring algorithm, analysis, and interpretation.) The Food and Drug Administration, the regulator and approver of medicines in the United States, has produced a guidance document for use in medical product development to support labeling claims (Food and Drug Administration 2009).

Data from a PRO instrument can be used to highlight any distinctive treatment advantages and disadvantages of a drug, which are not possible to be measured in other ways. Conversely, without PRO data, a drug's profile may be incomplete and as such does not represent the full base of potential benefits or harms patients would experience when using the medicine under investigation.

## Research Objectives

The most critical component of a study is its research objectives and goals. The implementation of study is successful only when its goals and research objectives are well defined with sufficient detail to guide its design, conduct, and analysis. The development of a clear and explicit a priori objective is vital for subsequent trial design and study conduct, especially if a sponsor wishes to seek a label claim or promote benefits of an intervention.

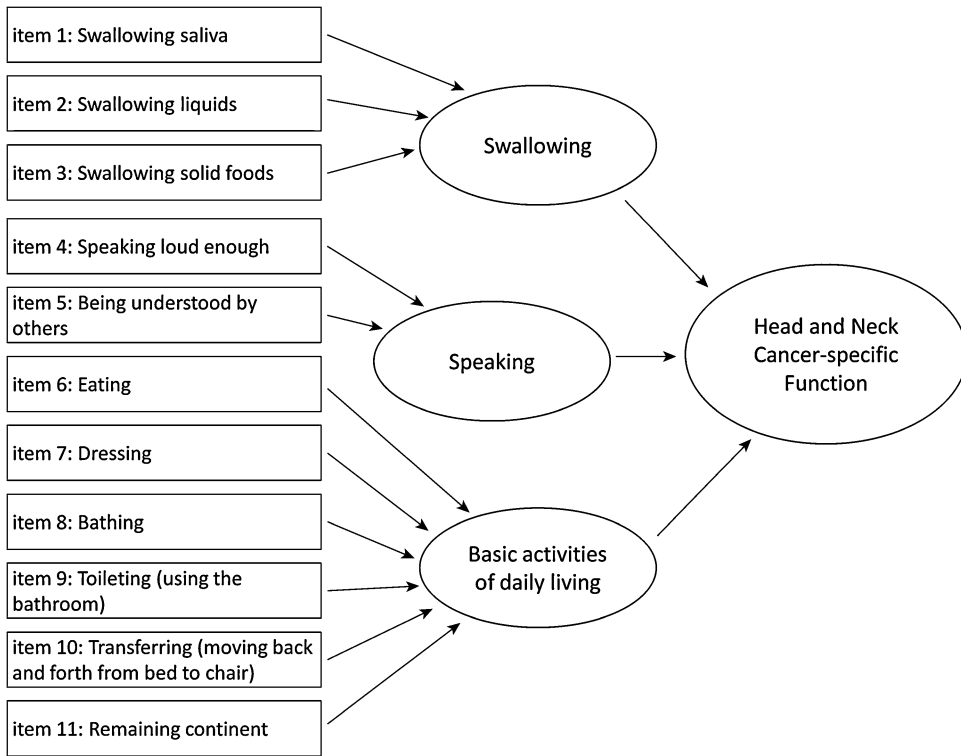
Stated objectives should breathe concrete and specificity, not vagueness and ambiguity. For example, the objective "To compare PROs between regimen A and regimen B" fails to provide specific information about the patient population of interest, time of assessment, and which aspects of a patient's condition will be assessed and compared. The stated objectives should refer to what is being measured and not the measurement instrument. Clear specifications of the details can help to better design study protocols and are vital to the ultimate success of a clinical

study. Here is an example of a specific and concrete objective: "Moodlift 20 mg, taken once daily, will lead to improvement in symptoms of depression, psychological function, and social function by 2 weeks among adult men with major depressive disorder" (Luo and Cappelleri 2008; Rothman et al. 2007). Based on the aspects of a patient's condition under investigation, relevant PRO instruments and relevant domains of those instruments should be identified.

In addition to identifying the relevant domains of a PRO, the population of interest, and the time frame of interest, objectives should have clear hypotheses as to whether the intent is to obtain a label claim or not, demonstrate superiority or non-inferiority, and seek confirmatory or exploratory evidence. Different endpoints may serve different purposes; for example, one PRO endpoint may be sought for a label claim and require confirmatory evidence, whereas another PRO endpoint may be considered exploratory with no intention of a label claim.

In seeking a PRO label claim being sought in the United States, sponsors of medicines are advised to place their research objectives and goals in terms of a conceptual framework, which may be useful in developing and refining the goals for PRO measurement. Guided by an appropriate conceptual model, which identifies and describes the PRO concepts and hypotheses that underlie a PRO-based product labeling claim, a conceptual framework explicitly defines or depicts the relationships between the items in a PRO instrument and the concept measured (Food and Drug Administration 2009; Rothman et al. 2007; Snyder et al. 2007). The concept is the specific measurement goal, that is, the attribute or characteristic measured by a PRO instrument.

If the desired overall claim, for instance, is "product X reduces problems with swallowing and speaking to others and improves daily activities for individuals with head and neck cancer," the diagram in Fig. 1 depicts a plausible conceptual framework of a PRO instrument where a set of items is associated with a specific domain, such as "swallowing," "speaking," and "basic activities of daily living"; moreover, the domains represent related but separate concepts (Patrick et al. 2007).



**Fig. 1** Diagram of the conceptual framework of a patient-reported outcome instrument

An instrument may create a single score, thereby measuring a single concept, or, as in Fig. 1, may be developed with multiple domain scores each represented by a concept, possibly within a more general concept of measurement, represented by the “head and neck cancer-specific function” domain. The conceptual model of a PRO instrument will evolve and be confirmed over the course of measurement development as a sponsor gathers empiric evidence to support item grouping and scores (Food and Drug Administration 2009).

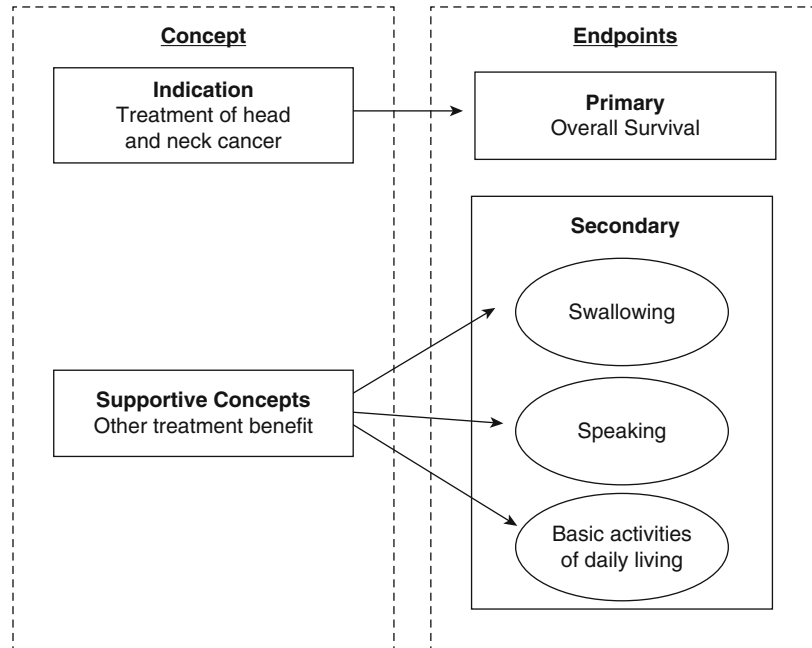
Related to the conceptual framework, an endpoint model should be described and depicted if a label claim is to be sought in the United States. It represents a diagram of the hierarchy of relationship among endpoints, both PRO and non-PRO, that corresponds to the clinical trial’s objectives, design, and data analysis plan (Food and Drug Administration 2009). Figure 2 depicts a hypothetical endpoint model for a head and neck cancer example (Patrick et al. 2007). Primary endpoints here include overall survival which, if

statistically significant and medically important, would be sufficient for a claim. If this survival endpoint showed a statistically significant and clinically meaningful treatment benefit, the domains of the PRO instrument – “swallowing,” “speaking,” and “basic activities of daily living” – are subsequently listed in order of importance as complementary endpoints that may result in a claim.

## Selection of Subjects

It is strongly recommended that protocol eligibility, whenever possible, be restricted to patients willing and able to participate in the PRO assessment. This challenging recommendation is motivated by following two fundamental rationales (Gotay et al. 1992). The first is practical. Study implementation is easier and more efficient when all patients require the same assessments (PROs as well as non-PROs). The second is scientific.

**Fig. 2** Hypothetical endpoint model for head and neck cancer example



Credibility and interpretation of the overall results, and their overall conclusions, are enhanced when all subjects are available for all endpoints.

Assessments on PROs should not be seen as optional by physician or patients. Optional assessments would jeopardize the ability of study results to be generalized to study population and, with randomization compromised, would likely lead to selection bias. The goal here is to avoid differential assessments on different patients because otherwise results can be biased and likely seriously biased. All measurements over time should be sought for all patients, not just some, in order to maintain validity and extension of inferences.

Physical, cognitive, or language barriers may make the evaluation of PROs impossible in practice for specific groups of patients (Gotay et al. 1992). In this case, alternative strategies for collecting PRO data should be considered. Such strategies include translation and culturally fine-tuning of PRO instruments, assistance for patients with visual or auditory impairments, and proxy assessments for patients with cognitive deficits. However, given the need to include patients who are elderly or in minority populations, a preferred

course of action is to make eligible patients a top priority and to have them complete all of their assessments in the same manner.

## Longitudinal Designs

Patient-reported outcomes are often incorporated into a study by administering questionnaires at multiple time points with the goal of characterizing the outcome over time (Fairclough 2004, 2005, 2010). Such longitudinal data arise in most PRO investigations because interest centers on how a disease or intervention affects an individual's functioning and well-being over time. The number and timing of PRO assessments is influenced by the study objectives, such as when meaningful change is expected, and practical considerations, such as patient burden. Key considerations in the design of a longitudinal study follow.

## Event- or Condition-Driven Designs

When the objective of a study is to compare a PRO in subjects who experience the same type

of condition during a given phase of treatment, assessments can be planned at times when clinically relevant events are expected to occur or at times that correspond to a distinct, meaningful phase of the intervention or disease. Such assessment is more common for a design with a relatively short duration. Many variations exist. Among them, for example, is when differences in PRO values are expected during only the early period of therapy. A breast cancer trial of adjuvant therapy in which a 16-week dose-intensive regimen was compared with a more traditional 24-week regimen is an illustration of such an event-driven design (Fetting et al. 1998). Three assessments were planned – prior to (baseline assessment), during, and after therapy – where each phase of the disease or its treatment was considered distinct with respect to the PRO of interest.

In event-driven designs where each assessment is conceptually identified with a landmark event, repeated measures models for longitudinal data (with time taken as a categorical covariate) are an appropriate choice. Note that assessments for all subjects should be taken at the same points in time (e.g., week 6, week 10, and week 24), where points in time need not be equally spaced. Repeated measures models may also be useful in some studies with only a few assessments.

### Time-Driven Designs

When the scientific questions involve a more extended period, or when the phases of the disease or its treatment are not distinct, the longitudinal designs are based on or driven by time (Fairclough 2005, 2010). These designs are appropriate for chronic conditions where therapies are given over elongated periods, such as diabetes and arthritis.

In time-driven designs, the duration of therapy may be indeterminate at study onset, with therapy intended to be given to a patient until it is not efficacious or produces unacceptable toxicity. For instance, patients with advanced renal cell carcinoma were randomized to receive either repeated 6-week cycles of sunitinib (experimental) or

interferon alpha (control) (Cella et al. 2008). Doses were adjusted in response to symptoms of toxicity. Treatment in both groups was continued until the occurrence of death, unacceptable adverse events, or withdrawal of consent. Patients were asked to complete the PRO questionnaires before any clinical activities during visits to the study clinics at screening, on days 1 and 28 of each 42-day treatment cycle, and at the end of treatment or study withdrawal.

Time-driven designs are associated with mixed-effect models for studies where time is often conceptualized and taken as a continuous variable. Mixed-effect models are useful when the timing of assessment differs widely among individuals, studies have a large number of PRO assessments, or changes over time are to be modeled with a smaller number of parameters than that required for a repeated measures model (with time as a categorical covariate).

### Timing of the Initial PRO Assessment

The initial assessment is the first and one of most important assessments in a study. This initial assessment, usually referred as a baseline assessment, plays crucial role in estimation of changes on PRO outcomes. If the baseline assessment is not present, all other data for this subject could be useless in the modeling of differences between treatments. It is also critical that the initial assessment occurs prior to randomization in randomized trials. Because the measurement of a PRO is generally based on self-evaluation, an initial assessment that follows randomization runs the risk that a subject's responses are influenced by knowledge of treatment assignment (Brooks et al. 1998). This risk becomes especially evident when one of the interventions is a new, promising therapy.

Sometimes multiple assessments, assessed before randomization, from daily patient diaries are collected and averaged to arrive at an overall baseline score. Such averaging increases the reliability (precision) of measurement relative to a single assessment. In two randomized, double-blind, placebo-controlled trials of pregabalin for fibromyalgia, a patient's baseline score on self-

reported sleep quality was computed as the average rating over the 7 days prior to taking study medication (Russell et al. 2009). In this daily diary assessment, patients completed the rating in the morning upon awakening and reported the quality of their sleep over the past 24 h on an 11-point numeric rating scale ranging from 0 (“best possible sleep”) to 10 (“worst possible sleep”).

### Timing of Follow-Up PRO Assessments

As with the timing of the initial PRO assessment, the timing of follow-up assessments should receive careful consideration (Fairclough 2010). A tenet of appropriate timing for follow-up assessments is that they should be made consistently across the treatment arms. It is important not to choose a particular time that will bias the results against one treatment or another. Measuring immediately after an untoward event such as toxicity will emphasize that experience at the expense of de-emphasizing the potential benefits of treatment and disease symptoms. When follow-up assessments on PROs are to be collected, they are usually positioned at all or some of the visits that other clinical assessments or lab measurements are collected.

A major factor when deciding on the timing of the PRO assessment, both initially and subsequently, is the recall period of the PRO questionnaire. Because individuals have better recall for major events and more recent experiences, the period of accurate recall for measuring certain areas (e.g., erectile dysfunction, physical well-being) is between 1 and 4 weeks, whereas the period of recall for the frequency and severity of symptoms (e.g., pain, fatigue) is accurate over shorter periods such as at the time of patient completion of the PRO or the past 24 h. That said, it should be noted that recall period established by the developers of the PRO instrument should be used. It is unadvisable to change a recall period of a PRO instrument to fit a particular study design, but rather a PRO instrument should be selected (or maybe even newly developed) to fit the study design. If a recall period for a PRO instrument was changed, some aspects, such as

test-retest, should be reevaluated. To be considered statistically independent observations, the timing of one assessment should not have a recall period that overlaps with the timing of another assessment on the same instrument; assessments should be based on distinct recall periods.

### Frequency of Evaluations

The frequency of the assessments depends on the natural history of the disease, the likelihood of meaningful changes during the study period, the recall period of a PRO (if the PRO is based on recall over the previous month, assessments should not be made weekly or daily), and how discontinuation of therapy relates to the research objective. All of these considerations should be balanced with practical considerations such as the burden placed on individuals who complete questionnaires and the timing of therapeutic and diagnostic interventions. Hence the assessments on PROs should be frequent enough to capture meaningful change over a sufficient duration but not frequent enough to cause excessive burden on participants.

In long-term studies with mortality as the primary endpoint, as in chronic heart failure trials, it is often useful to have more frequent assessments at the end of the study to enable detection of deterioration. If, on the other hand, rapid change is expected during the early part of a study, as is typically the case for renal cell carcinoma studies, more frequent assessments earlier on may be needed.

Assessments should not be more frequent than the period of recall defined for the PRO instrument. Instruments on satisfaction, functioning, and well-being are often based on the last 7 days or 4 weeks. Symptoms assessment scales often use the last 24 h or ask about the severity right now. Shorter periods of recall are generally more appropriate when the severity of symptoms are being evaluated, with more rapid changes in symptoms requiring a shorter recall duration, while the same or longer periods may be required to assess the impact of those symptoms on activities of daily living. Such was the case in a

non-small lung cancer trial where the severity of multiple symptoms and the impact of those symptoms on daily functioning from chemoradiation were evaluated during the last 24 h before the start of this intervention and weekly for 12 weeks during and after it (Wang et al. 2006).

In many cases what is of real interest is not the integrated effect over a short period (e.g., 2-week period), but the effect at regular intervals (e.g., 2, 4, and 6 weeks), similar to how measurements might be made every 2 weeks in a blood pressure trial (Food and Drug Administration 2009). For regulatory claims on a PRO, the recall period with the shortest time frame consistent with the instrument's purpose or intended use (e.g., when feasible, a recall period referenced to the patient's current or recent state) is preferable to a recall period that is based on a longer period, a comparison of a patient's current state with an earlier period, and a self-reported average over time (Food and Drug Administration 2009).

Patients who drop out of a study prematurely are generally more likely to have a less favorable score on a PRO because of side effects or no effect of treatment. A treatment arm with a high rate of dropout is likely to give an artificially more favorable outcome because only the healthiest of the patients remain on treatment, leading to selection bias and overly optimistic estimates of treatment effect. It is therefore desirable to have a PRO assessment in conjunction with premature withdrawal from the study. If the research objective extends to off-therapy assessments, then they can be made by continuing the PRO assessments after discontinuation. The off-therapy assessments can always be excluded if deemed uninformative or irrelevant to the research question. Including the off-therapy assessments after discontinuation allows them to be available should they be determined to be of interest.

---

## Selection and Evaluation of the Measurement Instrument

The PRO measurement strategy should be operationalized according to what study questions are to be answered. It is necessary to understand

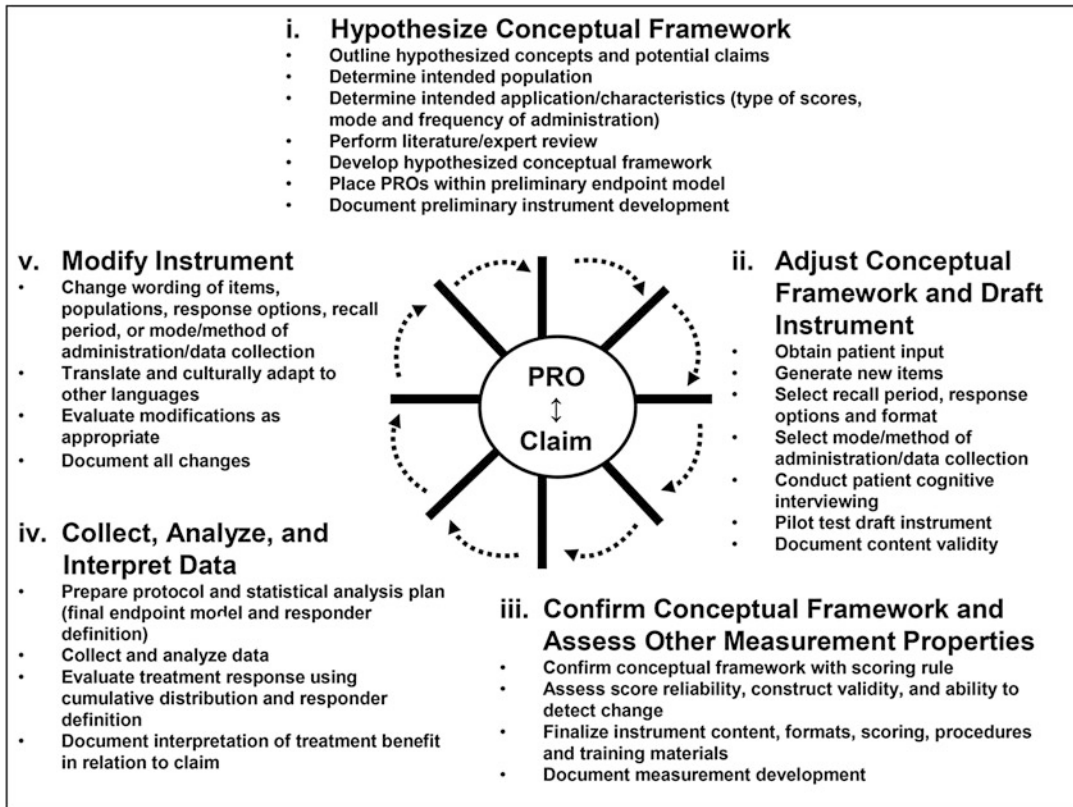
and justify the relevance of the selected PROs to the target disease, patient population, and study setting. Information on relevance can be obtained from the medical literature, previous studies, and direct input from patients and other stakeholders like families and health-care professionals. What is also needed is an understanding of the epidemiology and burden of disease from the patient's perspective and the postulated and empirical relationships between treatment, PROs, and other clinical outcomes.

The FDA guidance on PROs for a label claim in clinical trials recommends a wheel and spoke diagram as a way to organize the development process and provide the path by which the PRO can lead to a claim (Food and Drug Administration 2009; Patrick et al. 2007). The diagram is reproduced in Fig. 3. The five major steps highlighted in the diagram, which summarizes the iterative process used in developing a PRO instrument for use in clinical trials, apply regardless of whether sponsors use an existing instrument, modify an existing instrument, or develop a new instrument. This diagram encapsulates why the standards and preparations required for a PRO label claim are much more involved than when a label claim is not sought.

In what follows a series of key steps on good research practices that centers around the common theme of selecting and evaluating a PRO measurement instrument, be it for a regulatory claim or not (Luo and Cappelleri 2008).

### Step 1: Formulating Study Objectives

The evaluation of PROs begins with the formulation of study objectives (Fig. 4). If a sponsor wishes to seek a label claim or promote benefits of a drug, the development of a clear and explicit a priori objective is critical for subsequent trial design and study conduct. Stated objectives should breathe concrete and specificity, not vagueness and ambiguity, as stated in the section "Research Basis and Goals."



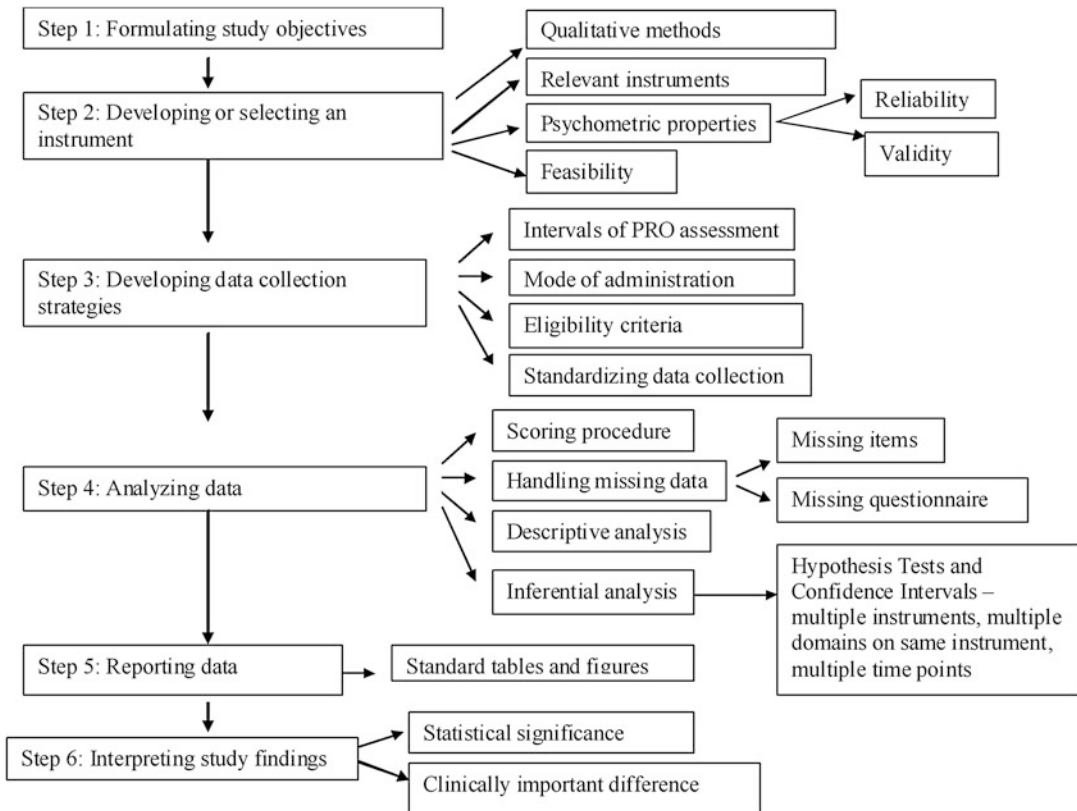
**Fig. 3** Development of a patient-reported outcome instrument for a label claim in a FDA application: an iterative process (Source: Food and Drug Administration 2009)

## Step 2: Developing or Selecting an Instrument

Instrument development can be an expensive and a time-consuming process. It usually involves a number of considerations: qualitative methods (item generation, cognitive debriefing, expert panels, qualitative interviews, focus groups), data collection from a sample in the target population of interest, item reduction, instrument validation, translation, and cultural adaptation. The importance of establishing content validity through qualitative methods – ascertaining that the measured concepts cover what patients consider the important outcomes of the condition and its therapy – cannot be overemphasized (Fig. 4). It is essential that the patients’ perspectives be taken in account when developing PROs, as the

whole point of the endeavor is to measure patients’ experiences; measurement properties have little value if the relevant concepts important to patients are not measured. Adequacy for the development process of a PRO, especially for a label claim, is contingent on the qualitative interview strategy, description of qualitative interviews and focus groups, transcripts, coding procedure, and justification for each step in the development of an instrument (Patrick et al. 2007)

The whole procedure of instrument development and validation can easily require at least 1 year. Therefore, the use of a previously validated instrument is typically preferable to the development of a new instrument that requires validation. For researchers who are not familiar with various instruments, updated information on currently available instruments can be accessed from databases such as



**Fig. 4** Key steps for selecting and evaluating patient-reported outcomes (Source: Reprinted with permission from Luo and Cappelleri 2008)

the Patient-Reported Outcome and Quality of Life Instruments Database (<http://www.proqolid.org>) and the On-Line Guide to Quality-of-Life Assessment (<http://www.OLGA-Qol.com>).

With many instruments currently available, the choice of the most appropriate instruments becomes vital to the success of a study in which PROs are included as a key endpoint. In what follows, several issues that need to be taken into consideration for instrument selection are highlighted (Fig. 4).

### Relevance of the Selected Instrument

As part of aligning a selective instrument with study objectives, the instrument should reflect the concrete, unambiguous questions being asked that are relevant to the targeted disease and study population. The instrument should also be able to measure intended benefits and harms of a treatment.

### Psychometric Properties of an Instrument

The selection of an instrument must also consider the instrument's measurement properties. Is the instrument measuring what it intended to measure – is it valid? Does it give accurate measurements – is it reliable? The selected instrument must be psychometrically sound. Measurement characteristics including reliability and validity are fundamental aspects for judging the quality and merits of an instrument (Fayers and Machin 2007; Streiner and Norman 2008).

Reliability measures to what extent an instrument yields reproducible and consistent results. Evidence on two types of reliability is usually required. One is internal consistency reliability, and another is test-retest reliability. The internal consistency reliability assesses to what extent the items of a domain or subscale are correlated – to



what extent the items move in tandem to measure different aspects of the same concept. The assessment of internal consistency reliability is usually carried out using Cronbach's alpha coefficient. Test-retest reliability measures to what degree an instrument gives similar scores when it is repeatedly administered to the same patient under a stable condition. It is often based on an intraclass correlation coefficient. For Cronbach's alpha and intraclass correlation coefficient, a minimum value of 0.7 is considered acceptable (Fayers and Machin 2007).

Assessing reliability is not sufficient for the validation of an instrument. An instrument may be reliable (accurate and precise in measuring the something) yet not measure what it is supposed to measure and hence not be valid. There are at least three major types of validity: content validity, construct validity, and criterion validity. Criterion validity is not assessed when there no criteria or "gold standard" measure, as is often the case for most of the diseases.

Content validity concerns the extent to which the constituent items reflect the intended concept. The assessment of content validity usually involves critical examination on whether the items are comprehensive enough and clearly cover, without ambiguity, the concept of interest. Content validity is often evaluated by consulting with patients having the disease of interest, physicians, and specialists to ensure that the included items are clear, comprehensive, and acceptable.

Construct validity is another fundamental characteristic of a measurement instrument and assesses to what extent an instrument measures the construct or concept it is supposed to measure. The assessment of construct validity often begins with postulating a relationship between the concept (construct) of interest and other related or unrelated measures or characteristics. Data are then collected, and the assessment is conducted. If the results confirm the postulated relationship, evidence exists to support construct validity.

Different methods can be used to establish construct validity. For example, construct validity can be assessed by comparing instrument scores among different groups of patients that are clinically distinct and anticipated to score differently

(discriminant validity). Construct validity can also be assessed by correlating instrument scores with other measures that are theoretically related (convergent validity) or unrelated (divergent validity) to the underlying concept measured by the instrument.

In addition to corrected item-to-total correlations (correlations between an item and the sum of the other items on the same domain), items in multi-item scales are often evaluated and confirmed by factor analysis. A "factor" is a latent variable, that is, an unobserved or hidden variable; the term "factor" may be defined and interchanged with the terms "domain," "construct," or "concept." A latent variable is a hypothetical construct that is not directly observed but whose existence is inferred from the way it influences the observed or manifest variables. Examples of a latent variable include depression and anxiety.

The statistical technique that can govern and quantify those interrelationships is factor analysis. Factor analysis is a multivariate statistical method concerned with detecting and analyzing patterns based on the correlations among quantitative variables. For PRO assessment, it attempts to identify groups of items such that there are strong correlations among all items within the same domain and weaker correlations among items in different domains. The purposes of factor analysis are mainly for the structural development and validation of scales.

Exploratory and confirmatory factor analyses are two major approaches to factor analysis (Brown 2006; Cappelleri and Gerber 2010; Fayers and Machin 2007). In factor analysis, the underlying structure of a set of measured items is summarized by a smaller set of latent (unobserved) factors that manifest themselves via the measured items. An objective is to identify the number and the nature of the factors that are responsible for covariation in the data and to determine the domain structure of a questionnaire (which items represent which domains), which is what exploratory factor analysis addresses. The domain structure may be unidimensional or multidimensional with several factors or domains (sometimes also called subscales). A further objective may be to confirm an existing domain structure in a separate,

independent group of individuals from the same population, which is what confirmatory factor analysis addresses.

It is difficult to fully and completely prove construct validity. Instead, researchers rely on accumulating amounts of evidence to demonstrate that an instrument is valid in measuring the concept of interest.

Responsiveness, which can also be viewed as another type of validity, is the ability of an instrument to detect small but important changes within a group over time. Responsiveness is one of the most essential characteristics of an instrument; a nonresponsive instrument has little use to discern true drug effects. Two of the most commonly used measures of responsiveness are the standardized response mean and the effect size. The standardized response mean is the ratio of the mean change to the standard deviation of that change. The effect size is the ratio of the mean change to the standard deviation of the initial measurement. The effect size measure is commonly considered more appropriate than the standardized response mean because the effect size uses natural variability stemming from patients' baseline values, which are not influenced by the effects of treatment, in order to help quantify what magnitude of change would be important. Measures of responsiveness like the effect size, being dimensionless, can be used to compare the responsiveness of a new instrument with that of existing ones.

Related to responsiveness is sensitivity: the ability to detect known differences between treatment groups over time or at a specific time. Its standardized measures of effect correspond to those for responsiveness except that the mean change is between groups instead of within group.

With the exception of content validity, which is based on qualitative methods, measurement properties are grounded in quantitative analysis usually involving correlations, means and regression methods, as well as theoretical expectations. Table 1 summarizes key measurement properties of a PRO.

### Feasibility

The final consideration on instrument selection is feasibility. Issues related to feasibility include

language availability, time required to complete the instrument, patient ability to complete the questionnaire, the rate of refusal, and percentage of missing items. All of these issues, each an important element itself, should be thought out when selecting an instrument.

### Step 3: Developing Data Collection Strategies

After determining which instrument will be used in an evaluation on PROs, a carefully planned data collection strategy should be built into study design and research protocol to ensure high quality of data (Fig. 4). Although this is true of any serious study design and research, the fact that PROs are based on a patient's self-report makes it even more important to develop a judicious strategy in order to prevent or minimize bias or missing data. An important consideration when developing the data collection strategies is the time intervals that PROs are assessed, as discussed in the section "[Longitudinal Designs](#)."

Time intervals of assessment should be based on disease progression, treatment response, drug side effects, duration of the study, and number of questionnaires. At a minimum, assessments of PROs should be performed at baseline and at the end of study. But intermediate follow-up measurements may be required to more fully capture changes within group and between groups over time. Therefore, a reasonable number of assessments to capture this trajectory should be planned in a clinical trial.

Assessments of PROs are usually performed at the same time as clinical visits and are best completed before professional encounters with non-PRO measures, which may influence a patient's response on PROs. The mode of administration on PROs can be obtained by paper and pencil, computer administration, electronic devices, or in-person or phone interviews. The same PRO should use the same mode of administration throughout the study.

Standardized data collection procedures need to be established to ensure that the data are collected consistently among different patients and

**Table 1** Measurement properties for PRO instruments

Measurement property	Type	What is assessed?	FDA review considerations
Reliability	Test-retest or intra-interviewer reliability (for interviewer-administered PROs only)	Stability of scores over time when no change is expected in the concept of interest	Intraclass correlation coefficient Time period of assessment
	Internal consistency	Extent to which items comprising a scale measure the same concept Intercorrelation of items that contribute to a score Internal consistency	Cronbach’s alpha for summary scores Item-total correlations
	Inter-interviewer reliability (for interviewer-administered PROs only)	Agreement among responses when the PRO is administered by two or more different interviewers	Interclass correlation coefficient
Validity	Content validity	Evidence that the instrument measures the concept of interest including evidence from qualitative studies that the items and domains of an instrument are appropriate and comprehensive relative to its intended measurement concept, population, and use. Testing other measurement properties will not replace or rectify problems with content validity	Derivation of all items Qualitative interview schedule Interview or focus group transcripts Items derived from the transcripts Composition of patients used to develop content Cognitive interview transcripts to evaluate patient understanding
	Construct validity	Evidence that relationships among items, domains, and concepts conform to a priori hypotheses concerning logical relationships that should exist with measures of related concepts or scores produced in similar or diverse patient groups	Strength of correlation testing a priori hypotheses (discriminant and convergent validity) Degree to which the PRO instrument can distinguish among groups hypothesized a priori to be different (known groups validity)
Ability to detect change		Evidence that a PRO instrument can identify differences in scores over time in individuals or groups (similar to those in the clinical trials) who have changed with respect to the measurement concept	Within person change over time Effect size statistic

Source: Food and Drug Administration 2009

investigators and across various study sites. Before the start of the trial, data collection personnel and study monitors should be carefully and uniformly trained. A detailed guideline on the assessment of PROs should be prepared and serve as a reference book for study monitors and data collection personnel in order to handle issues arising from the assessment.

Missing data can occur at the item level for at least one but not all items on the questionnaire or at the questionnaire level for all of its items. The reasons for missing data should be recorded at the time of occurrence and later considered to lend insight into the potential patterns for why data are missing. Because data quality is directly linked to the validity of study findings, researchers should

have a thorough understanding about the data collection process along with potential issues and biases inherent in this process. Such knowledge can help facilitate the development of appropriate data analysis plans to understand and minimize potential bias.

If missing data do occur for some but not all items on the questionnaire, the non-missing data may still be used for analysis based on some prespecified criteria, usually recommended by the developers of the questionnaire. For example, the EORTC QLQ-C30 (European Organization for Research and Treatment of Cancer Quality of Life Questionnaire – Cancer-30) consists of five functional scales [physical, role, cognitive, emotional, and social], three symptom scales (fatigue, pain, nausea and vomiting), a global health status scale, and six single-item scales (Fayers et al. 2001). The EORTC QLQ-C30 Scoring Manual has specified that under certain conditions, missing values will be imputed for multi-item scales. Specifically, if at least half of the items from the scale have been answered, the missing items are assumed to have values equal to the average of those items which are present for the respondent. For example, the physical function subscale consists of 5 items, and this scale can be estimated whenever at least 3 of its 5 constituent items are present. More is said about missing data in the section “[Missing Data.](#)”

Sample size estimation is an indispensable part of a data collection strategy and depends on the study objective. In principle, there are no major differences in planning studies for a comparison between treatment groups using PROs compared with using non-PRO clinical measures such as blood pressure levels. As such, sample size estimation for PROs will require specification of the significance level, statistical power, anticipated difference or effect size, expected dropout rate, and type of data and method of analysis (Fayers and Machin 2007). As already stressed, it is important and necessary to clearly state and limit the major PROs of interest in the study protocol. Doing so is especially relevant for sample size purposes.

Sample size estimation for PROs becomes specialized for psychometric techniques like factor

analysis and reliability where the objective is on an instrument’s measurement properties, rather than a comparison between treatment groups. Factor analysis is a large-sample procedure, and a valid factor analysis typically involves hundreds of subjects. Sample size estimation for factor analysis depends on several elements such as the distribution of items and correlations between them. One suggested rule of thumb is to recommend a sample size of at least ten times the number of items for an exploratory factor analysis (Fayers and Machin 2007) and at least ten times the number of parameters (measurement-error variances, covariances among domains, factor loadings) for confirmatory factor analysis (Brown 2006). Sample size estimation for test-retest reliability can be based on Fisher’s *Z* transformation for an intraclass correlation using a confidence interval approach (Streiner and Norman 2008).

Although repeated measures and mixed-effect models are often used in the analysis of PRO measurements over time, sample size estimation is most commonly based on calculating the expected difference in the group means at a single time point rather than over time. This calculation amounts to sample size estimation for a univariate analysis and in most cases provides a conservative (larger than necessary) estimate of the sample size. Procedures are also available for the estimation of sample size in a longitudinal analysis with a repeated measures model or mixed-effect model (Fairclough 2010; Fitzmaurice et al. 2011).

#### **Step 4: Analyzing Data**

The next step in the evaluation of PROs is to develop prespecified, comprehensive, and detailed plan on data analysis (Fig. 4). For a clinical trial, the statistical analysis plan (SAP) on PROs is best integrated with other study endpoints as part of an overall analytic strategy. Gains in efficiency arise when PROs are integrated and unified with other endpoints in the SAP.

The SAP part on PROs should be clear and concise, and yet complete and comprehensive, about the stated objective. In addition to the data

analysis on PROs, the SAP should also include a brief description on how the instruments are selected, how domains belonging to an instrument are scored, and how missing items of an instrument are handled. The development of data analysis plan should be based on study objectives and may vary among different phases of clinical trials.

For example, for a phase II trial intended to explore the potential impact of a specific drug treatment on PROs, the analysis plan can focus on a comprehensive descriptive analysis and, if suitable, an inferential analysis. Basic statistics such as instrument compliance rate, the observed mean of domain scores (along with confidence intervals such as a 95% confidence interval), and the observed mean change from baseline (and its 95% confidence interval) to each follow-up time should be included within each group. Additionally, if a trial has multiple arms, a comparison of the domain scores between arms is typically worthwhile to include by analyzing (and then reporting) the between-group difference in changes from baseline to each follow-up time, along with the corresponding difference in mean changes and its 95% confidence interval.

For a phase III trial, especially one intended for a label claim based on a PRO outcome, inferential statistics (hypothesis testing and confidence intervals) should be the focus of the analysis plan, along with a detailed descriptive summary. Regardless of phase of the study, data on PROs should be treated just like any other study points and adopt the same analytical rigors.

As discussed in the section “[Longitudinal Designs](#),” event-driven designs are generally associated with repeated measures longitudinal model, where time is a categorical covariate. Restricted maximum likelihood estimation of repeated measures models can account for incomplete data and time-varying covariates. Time-driven designs are associated with mixed-effect longitudinal models via growth curve models, where time is taken as a continuous covariate. It is generally enough for these models to include polynomial or piecewise linear models and typically allow one to three random effects (intercept; intercept and slopes; intercept, slope, and additional variation over time). Both repeated

measures models and mixed-effect models incorporate all available data and assume that data are missing at random.

Inferential testing of data on PROs should consider the analytical issues specific to the evaluation of PROs in a clinical trial. For example, many instruments have multiple domains, and each instrument may be measured a number of times. Multiple comparisons then become an important issue that deserves special consideration. Missing data usually occur in PRO studies. How to handle the missing data also requires special considerations. More detail on these two issues follow.

### Multiple Testing

It has been well recognized that the multiple comparisons of drug treatments can result in false significant results. Because data on a particular PRO is usually measured over a number of time points, and because the same study may comprise multiple PROs (or multiple subscales within the same PRO instrument), it becomes important to describe in the SAP how to deal with this multiplicity issue, especially if the evaluation in the clinical trial is intended for label claims based on PRO outcomes. Several methods can be applied to address the multiple testing (Fairclough 2010).

One of the methods is to use summary measures or summary statistics. For many instruments, a single score can be constructed by aggregating data across different domains on the same questionnaire. Such a summary score can be used as the primary endpoint for hypothesis testing and, consequently, prevents the concern of repeated testing on multiple domains of the same instrument.

Summary measures can also be constructed on a particular subscale or domain of an instrument to summarize the repeated observations over time on an individual and then across individuals in the same treatment group. Examples include, for each treatment group, the average of within-subject posttreatment values, area under growth curve, and time to reach a peak or prespecified value. The use of these summary measures begins with the construction of a summary measure for each individual, follows with the analysis of a summary measure across individuals for a within

group, and then continues with a corresponding between-group comparison. For instance, it is possible to construct summary statistics on the repeated measures within a group of individuals by taking the average rate of change over time for a treatment group and then comparing these summary statistics between groups.

A potential problem with the use of the summary score is that significant changes in some specific domains may be masked and what is really measured may become clouded or convoluted, resulting in low confidence about the effect of treatment as measured by the summary score. A drawback of summary measures across time is that they do not fully capture the weighted and correlated nature of repeated observations on PROs over time.

Another way to minimize the problem of multiplicity is to restrict the number of key domains and time points, no more than a few. These key domains at specific time points should be prespecified in the SAP as primary endpoints for statistical inference. Other domains at other time points may be regarded as secondary endpoints. While this recommendation provides a straightforward way to handle the multiplicity issue, a major challenge is how to select the most appropriate domains and time points. One way to address this challenge is to rely on substantive knowledge, well-grounded theory, and research objectives in tandem with the nature of the disease and the intended effects of the interventions.

Often several multiple endpoints, both PRO and non-PRO endpoints, would be of clinical interest. One suitable method is to test them using a gatekeeping strategy whereby secondary endpoints are analyzed and tested inferentially in a prespecified sequential order only after success on a primary endpoint (Food and Drug Administration 2009). More generally, the key endpoints are ranked from most important to least important from the list of endpoints considered most relevant. This process can be done using a sequential method by testing additional endpoints in a defined sequence each at the usual alpha at the 0.05 level of statistical significance. The analyses cease when a failure occurs. It is important that the clinical trial protocol specifies all relevant primary

and secondary endpoints and their order for inferential analysis and testing.

The problem of multiplicity can also be addressed in several other ways including through p-value adjustment. Three types of p-value adjustment are commonly considered: (1) Bonferroni, (2) Bonferroni-Holm (step-down) procedure, and (3) Hochberg's (step-up) method. Of the three methods, the Bonferroni procedure is the most conservative. In contrast, the Holm's procedure and Hochberg's method may be more accurate and preferable.

### Missing Data

Missing data on PROs can have at least two major repercussions. At a minimum, the missing data will result in wider confidence intervals and reduced statistical power for detecting a treatment effect. The larger, more troublesome issue is the likelihood that missing data are closely linked to patients' health and treatment, leading possibly to a biased estimation of treatment effects. Given these potential impacts, the SAP should clearly describe how to handle missing data, especially if the evaluation on PROs is intended for label claims or promotional use.

Missing data on PROs can occur as missing items or missing questionnaires. Missing items involve the lack of responses for some specific items; missing questionnaire involves patients who may fail to complete and return the whole questionnaire. Many instruments include well-documented procedures by their developers on how to handle missing items. Such recommendations by developers are typically the preferred way to address missing items.

Missing questionnaires are a more complex situation than missing items. Missing questionnaires can happen as a result of dropout from the study or randomly failing to fill out an entire questionnaire. In any of these situations, it is important to first analyze the rates (proportions) and reasons for missing data. Such information will help to gauge the severity of the nonresponse problem and the underlying mechanisms for missing data.

There are at least four approaches to address the missing data problem (Fairclough 2010). One

approach is to remove patients with missing or incomplete forms from the analysis and only analyze complete cases. While simple, this method is usually not recommended because it can break down initial randomization and reduce sample size and, in doing so, may produce bias results if the missing data are not missing completely at random. (Missing completely at random occurs when the missingness is unrelated to PRO value as when, e.g., a patient moves out of town or a staff member forgets to administer the questionnaire.)

A second approach is to impute the missing data. Different methods can be used for the imputation. The simplest way is to substitute the mean scores of patients with observed data for those with missing data (mean imputation). Unless the missing data are missing completely at random (MCAR), this means imputation method may result in bias estimates and should be used cautiously. Another commonly used method is last observation carried forward, which replaces a patient's missing value with his last completed observation. In the event that data on PROs may not remain stable over time, last observation carried forward may also be suspect and result in a bias representation (Mallinckrodt et al. 2008). Analogous to last observation carried forward approach is the baseline observation carried forward approach, when all missing values for a subject are replaced by his or her baseline observation. Relative to the method based on last observation carried forward, this method can produce more conservative results for treatment differences.

Some more sophisticated techniques have been developed including regression imputation, hot deck imputation, and cold deck imputation. All of these techniques, like the simple mean imputation and last observation carried forward, belong to a single imputation category in which a single value is imputed for a specific missing point. A major limitation with single imputation methods is that estimated errors are generally too small, as the imputed values are treated as actual data when in fact they are not. However, this obstacle can be overcome by multiple imputations whereby several values are imputed instead of just one.

Multiple imputation method, which improves the accuracy of standard error, assumes that the missing data are missing at random (MAR), where the missingness depends only on the observed data such as the most recently observed PRO value.

A third approach to address the problem of missing data is through the application of a likelihood-based approach using repeated measures models or mixed-effect models (Fairclough 2010; Fitzmaurice et al. 2011; Mallinckrodt et al. 2008). In this approach, every subject would contribute his or her available (observed) measurements. Repeated measures models and mixed-effect models employ a likelihood-based approach that is considered attractive because it can provide valid estimate of treatment effects if missing data are MCAR or MAR, where the missing data are said to be ignorable.

The fourth approach is especially relevant when missing data are not MAR and hence depend on the (unknown) missing value, when missing data are said to be non-ignorable. In this case, selection models or pattern-mixture models, which do not assume that data are neither MCAR nor MAR, should be considered as secondary models in sensitivity analyses. For the analysis of longitudinal data, it is generally preferred to consider, depending on the circumstances, a repeated measures model or mixed-effect model as the main model and multiple imputation or pattern-mixture models (or both) as secondary models.

The National Research Council has produced an authoritative account on the prevention and handling of missing data in clinical trials (National Research Council 2010), which can be relevant to prevention and handling of missing PRO data.

## Step 5: Reporting Data

The reporting of data on PROs is a critical component to their evaluation (Fig. 4). Data on PROs should be presented clearly, concisely, and sufficiently to foster clarity, transparency, and comprehension. While a table is a useful way to summarize study results, graphical presentations

is especially appealing in simplifying and depicting the longitudinal and multidimensional nature of data on PROs (Fayers and Machin 2007). Whether a table or graph is used, it is imperative to present information as comprehensively and practically as possible. For example, data on the number of subjects completing the PRO evaluation at each treatment assessment should be reported, as should the metrics of variability embodied as in confidence intervals or standard errors of estimates.

## Interpreting Study Findings

The data analysis may show a statistically significant difference on scores of PROs between treatment groups at a specific time or a significant change within or between groups over time. In addition to statistical significance, a natural ensuing question is whether the treatment difference or change is clinically meaningful (Fig. 4). It has been well recognized that statistical significance may not imply clinical significance. For example, a small difference on PRO scores between two treatment groups may be statistically significant given a large sample size, but clinical relevance may be scant or difficult to interpret in a meaningful manner. Understanding the degree of difference on scores of PROs that is considered to be clinically meaningful can enhance the application and interpretation on PROs.

A number of methods have been proposed for establishing meaningful change in PROs. These methods can be grouped into two broad categories: anchor-based and distribution-based approaches (Fayers and Machin 2007; Food and Drug Administration 2009; Revicki et al. 2008).

Anchor-based methods are those in which differences at a given time or changes over time in PROs are linked – or anchored – to differences or changes in an external clinical measure (e.g., patients' global rating of change and clinical rating of disease severity) or to a yardstick value or even to part of the PRO measure under consideration. When used as an external clinical measure, an anchor should bear an appreciable correlation to the PRO and have clinical understanding and

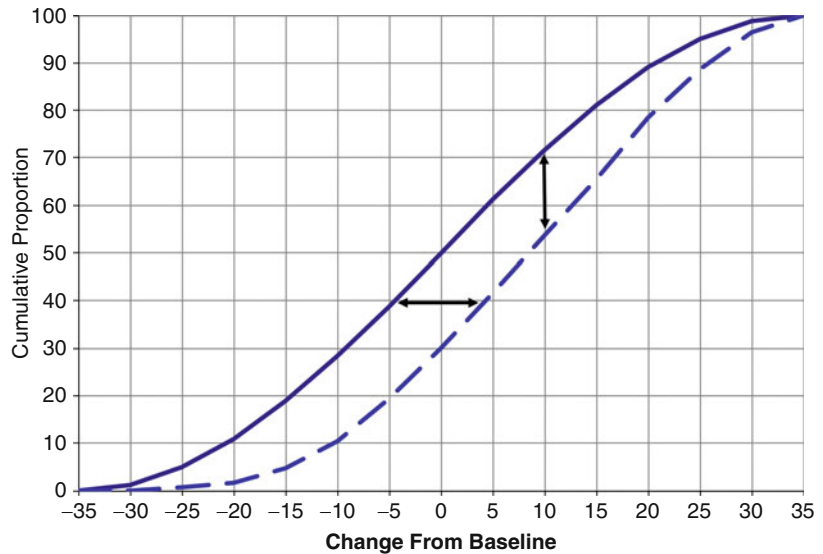
import. Anchor-based approaches include percentages based on thresholds, the percentage of patients above and below some specified value; criterion-group interpretation, the comparison of scores from the particular group of interest to a group or external variable worthy of comparison; content-based interpretation, a representative item internal to the multi-item PRO itself; and clinical important difference, a difference on a PRO that is deemed clinically relevant. For example, a criterion-group interpretation would involve a comparison of PRO scores in the population of interest with norm-based PRO scores from a general population or with external variables such as utilization of health-care services and ability to work.

Distribution-based approaches use the statistical characteristics of the sample (e.g., mean and standard deviation) or instrument (e.g., reliability) to suggest a clinically meaningful change. Distribution-based approaches include effect size, probability of relative benefit, and responder analysis and cumulative proportions. A widely used distribution-based method is the effect size, discussed earlier in the section “[Psychometric Properties of an Instrument](#).” The approach based on the probability of relative benefit, which is based on riddit analysis using the Wilcoxon rank-sum test statistic, gives the probability of a randomly selected individual on one treatment arm having a more favorable score than a randomly selected individual on the other treatment arm (Alcion et al. 2006). Because a distribution-based approach like effect size and probability of relative benefit is derived purely from a statistical distribution, and not from patient input, it does not provide an estimation of *clinical* significance per se.

According to the FDA final guidance on PROs for a label claim, it is recommended to display individual responses using a priori responder definition: the threshold value on an individual PRO change score that is to be interpreted as a treatment benefit (Food and Drug Administration 2009). The proportion of subjects meeting the responder definition can then be reported for each treatment group and compared between groups. As stated in the FDA guidance, it is



**Fig. 5** Illustrative cumulative distribution functions of two treatment groups where more negative change scores are better (solid line = experimental group, dashed line = control group)



usually useful to display individual responses, often using an a priori responder definition (i.e., the individual PRO score change over a pre-determined time period that should be interpreted as a treatment benefit). The responder definition is determined empirically and may vary by target population or other clinical trial design characteristics. The empiric evidence for any responder definition is derived using anchor-based methods, which explore the association between the targeted concept of the PRO instrument and the concept measured by the anchor (or anchors). To be useful, the anchors chosen should be easier to interpret than the PRO measure itself.

A cumulative distribution function can display a continuous plot of the change from baseline on the horizontal axis and the cumulative percent of patients experiencing up to that change on the vertical axis. Consider a situation where lower change or more negative scores are better or more favorable (Fig. 5). In Fig. 4, 70% of the subjects in the experimental group had scores of 10 or less (i.e., 10 or better) compared with 55% of the subjects in the control group. The consistent horizontal separation between the distribution functions suggests that the treatment was beneficial relative to control over the entire range of changes.

Responder analysis and cumulative distribution functions are best suited as descriptive displays and as an adjunct to – as a complement and

supplement to – the main analysis based on the full original scale of measurement using established statistical methods (e.g., repeated measures models or mixed-effect models when the data are longitudinal).

## References

- Alcion L, Petersen JL, Temple S, Arndt S. Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Stat Med.* 2006;25:591–602.
- Brooks MM, Jenkins LS, Schron EB, Steinberg JS, Cross JA, Paeth DS. Quality of life at baseline: is assessment after randomization valid? *Med Care.* 1998;26:1515–9.
- Brown TA. *Confirmatory factor analysis for applied research.* New York: The Guilford Press; 2006.
- Cappelleri JC, Gerber RA. *Exploratory factor analysis.* In: Chow S-C, editor. *Encyclopedia of biopharmaceutical statistics.* 3rd ed., revised and expanded. New York: Informa Healthcare; 2010. p. 480–5.
- Cella D, Li JZ, Cappelleri JC, Bushmakin A, Charbonneau C, Kim ST, Chen I, Michaelson MD, Motzer RJ. Quality of life in patients with metastatic renal cell carcinoma treated with sunitinib versus interferon-alfa: Results from a phase III randomized trial. *J Clin Oncol.* 2008;26:3763–9.
- Fairclough DL. Patient reported outcomes as endpoints in medical research. *Stat Methods Med Res.* 2004;13:115–38.
- Fairclough DL. *Analysing longitudinal studies of QoL.* In: Fayers P, Hayes R, editors. *Assessing quality of life in clinical trials.* Oxford: Oxford University Press; 2005. p. 149–65.

- Fairclough DL. Design and analysis of quality of life studies in clinical trials. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2010.
- Fayers FM, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes. 2nd ed. Chichester: Wiley; 2007.
- Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, Bottomley A. On behalf of the EORTC quality of life group. In: EORTC QLQ-C30 scoring manual. 3rd ed. Brussels: EORTC; 2001.
- Fetting JJ, Gray R, Fairclough DL, Smith TJ, Margolin KA, Citron ML, Grove-Conrad M, Cella D, Pandya K, Robert N, Henderson C, Osborne K, Abeloff MD. A 16-week multidrug regimen versus cyclophosphamide, doxorubicin and 5-fluorouracil as adjuvant therapy for node-positive, receptor negative breast cancer: an intergroup study. *J Clin Oncol*. 1998;16:2382–91.
- Fitzmaurice GH, Laird NM, Ware JH. Applied longitudinal analysis. 2nd ed. Hoboken: Wiley; 2011.
- Food and Drug Administration. Guidance for industry on patient-reported outcome measures: Use in medical product development to support labeling claims. *Fed Regist*. 2009;74(235):65132–3.
- Gotay CC, Korn EL, McCabe MS, Moore TD, Cheson BD. Building quality of life assessment into cancer treatment studies. *Oncology*. 1992;6:25–8.
- Johnson JR, Temple R. Food and drug administration requirements for approval of new anticancer drugs. *Cancer Treat Rep*. 1985;69:1155–9.
- Luo X, Capperli JC. A practical guide on interpreting and evaluating patient-reported outcomes in clinical trials. *Clin Res Regul Aff*. 2008;25:197–211.
- Mallinckrodt CH, Lane PW, Schnell D, Peng Y, Mancuso JP. Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Inf J*. 2008;42:303–19.
- National Research Council. The prevention and treatment of missing data in clinical trials. Washington, DC: The National Academies Press; 2010.
- Patrick DL, Burke LB, Powers JH, Scott JA, Rock EP, Dawisha S, O'Neill R, Kennedy DL. Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value Health*. 2007;10:S125–37.
- Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*. 2008;61:102–9.
- Rothman ML, Beltran P, Cappelleri JC, Lipscomb J, Teschendorf B, Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. Patient-reported outcomes: conceptual issues. *Value Health*. 2007;10:S66–75.
- Russell IJ, Crofford LJ, Leon T, Cappelleri JC, Bushmakina AG, Whalen E, Barrett JA, Sadosky A. The effects of pregabalin on sleep disturbance symptoms among individuals with fibromyalgia syndrome. *Sleep Med*. 2009;10:604–10.
- Snyder CF, Watson ME, Jackson JD, Cella D, Halyard MY, Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. Patient-reported outcomes instruction selection: designing a measurement strategy. *Value Health*. 2007;10:S76–85.
- Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th ed. - New York: Oxford University Press; 2008.
- Wang XS, Fairclough DL, Liao Z, Komaki R, Chang JY, Mobley GM, Cleeland CS. Longitudinal study of the relationship between chemoradiation therapy for non-small-cell lung cancer and patient symptoms. *J Clin Oncol*. 2006;24:4485–91.
- Wiklund I. Assessment of patient-reported outcomes in clinical trials: the example of health-related quality of life. *Fundam Clin Pharmacol*. 2004;18:351–63.



Carolyn M. Rutter

## Contents

<b>Introduction</b> .....	560
<b>Development of a Microsimulation Model</b> .....	560
Step 1: Define the Decision Problem .....	560
Step 2: Specify the Model Structure .....	563
Step 3: Identify Data Sources .....	568
Step 4: Select Model Parameters .....	569
<b>Implementation</b> .....	570
Example: Comparison of Two Tests to Screen for Colorectal Cancer .....	570
Sensitivity Analysis .....	571
Exploration and Description of Model Uncertainty .....	571
Model Validation .....	572
<b>In Conclusion</b> .....	574
<b>References</b> .....	574

## Abstract

Microsimulation models are a tool for informing health policy decisions. Models provide a structure for combining a wide range of evidence that represents the current understanding of both disease and interventions to prevent or treat disease. In the health policy context, *microsimulation* refers to simulation of an entire population by simulating life histories for *individuals* within the population. The basic structure of a microsimulation model includes a description of health states

that describe key events in a disease process. Individuals occupy these health states, and the model includes rules describing how individuals transition between states. Models are developed by specifying states and transition rules that result in predictions that reproduce observed or expected results. Model parameters are selected to achieve good prediction through a process of model calibration. Once calibrated, models are used to predict population-level outcomes under different policy scenarios. Model predictions are increasingly being used to provide information to guide health policy decisions. This increased use brings with it the need both for better understanding of microsimulation models by policy researchers and continued

C. M. Rutter (✉)  
RAND Corporation, Santa Monica, CA, USA  
e-mail: [crutter@rand.org](mailto:crutter@rand.org)

improvement in methods for developing and applying microsimulation models. This chapter reviews the process of developing and applying a microsimulation model, drawing from guidelines for best practices for simulation outlined by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and The Society for Medical Decision Making (SDM) (Caro et al. 2012).

---

## Introduction

Microsimulation models for health policy are a type of decision analytic model that describe disease processes by simulating key events that occur as disease develops. Their purpose is to help decision makers identify trade-offs associated with different policy decisions. For example, the National Cancer Institute has advanced the use of models for the cancer outcomes through the Cancer Intervention and Survival Modeling Network (CISNET) (2014). CISNET models have been used to inform policy recommendations regarding use of newer colorectal cancer screening tests (fecal immunochemical tests, stool-based DNA, and computed tomography colonography) and to assist in development of guidelines for breast and colorectal cancer screening. As early as the 1980s, models were used by the American Cancer Society to aid in guideline development for cervical cancer screening and by the US Congress's Office of Technology Assessment for evaluation of cervical and breast cancer screening policy (Eddy 1987; Muller et al. 1990). Models have also been used to inform policy and clinical practice related to medications, radiology, vaccination, and HIV screening (Mandelblatt et al. 2012). Examples of policy-relevant findings from models include overdiagnosis of prostate cancer among PSA-detected cases (Etzioni et al. 2002); identification of efficient cervical cancer screening policies (van der Akker-van Marle et al. 2002); and the impact of modifiable risk factors, screening, and treatment on colorectal cancer (CRC) mortality rates (Vogelaar et al. 2006).

Models can help decision makers choose among competing courses of action by structuring and combining a wide range of evidence, including information about disease process and clinical and economic outcomes, and then predicting patient outcomes based on this evidence. Microsimulation models are used to predict outcomes under different policy scenarios and are especially useful for outcomes that cannot readily be studied via direct observation for ethical or practical reasons. Model predictions may extend cross-sectional results to longitudinal predictions, extend results to different patient populations, or make direct comparisons not made in available randomized trials. For example, randomized trials demonstrate that both fecal occult blood testing (FOBT) (Hardcastle et al. 1996; Kronborg et al. 1996; Towler et al. 1998) and flexible sigmoidoscopy (Atkin et al. 2010) reduce CRC mortality. There is no direct evidence that either optical colonoscopy or CT colonography reduces mortality, though several studies have estimated their sensitivity and specificity for detecting colorectal adenomas (the primary precursor of colorectal cancer) (Hixson et al. 1990; Johnson et al. 2008; Rex et al. 1997). Microsimulation models for colorectal cancer have been used to combine available information about the natural history of disease and screening tests to compare the effectiveness and cost-effectiveness of all four of these screening modalities (Knudsen et al. 2010; Lansdorp-Vogelaar et al. 2010).

---

## Development of a Microsimulation Model

Table 1 shows the steps in developing a microsimulation model.

### Step 1: Define the Decision Problem

The first job of the modeler is to **define the decision problem**, that is the modeling objectives. It is important to be clear about the objectives, because these will drive model structure and complexity. Modeling is a collaborative process. Consulting with experts knowledgeable about the targeted

**Table 1** Steps in developing a microsimulation model

<b>Step 1: Define the decision problem</b>
What interventions will be modeled?
What events are of interest?
What is the target population and what subgroups are of interest?
<b>Step 2: Conceptualize the model structure</b>
Will models describe events in discrete or continuous time?
What disease states and characteristics will the model describe?
When (and how) do individuals transition between states?
<b>Step 3: Identify and select data sources</b>
Which data will inform the model?
How will each data source inform the model – as an input, calibration target, or validation target?
<b>Step 4: Select model parameters</b>
Which parameters are “inputs” and which parameters will be calibrated?
Which goodness of fit measure will be used to guide calibration?
Which calibration method will be used for parameter selection?

disease from the outset will help to ensure development of a useful model that addresses important policy questions. Clinicians and epidemiologists who are familiar with the disease process can help inform the model structure to ensure face validity of the model and can provide insight into key questions that cannot readily be addressed through direct observation. Policy makers and other end users can help to determine necessary model output and provide additional insight into policy questions.

Three key questions, described below, need to be considered when defining the decision problem.

### **What Interventions Will Be Modeled?**

Interventions can include primary prevention of disease, screening for purposes of early detection, methods for diagnosing disease, and treatment after diagnosis. The action of the intervention will determine key health states that need to be included in the model structure. For example, models for screening need to describe disease states that occur before clinical (symptomatic) presentation, because screening

affects the disease process through detection of preclinical states.

### **What Events Are of Interest?**

Events that are outcomes, such as cases of and deaths from the disease of interest, need to be described by the model. All-cause death is another event that is almost always modeled, because it enables calculation of life-years gained (or lost) that result from intervening on the disease process. The events that are modeled are closely related to the interventions of interest. Models for prevention and screening need to describe preclinical (asymptomatic) disease processes. In contrast, models that focus on treatment focus on detected disease need to describe remission and recurrence.

### **What Is the Target Population and What Subgroups Are of Interest?**

Those eligible for intervention often define the target population, with the earliest age of intervention defining the beginning of the age range, which may extend through the entire simulated life span. For example, models for cervical cancer screening focus on women who are 18 years and older, while models of for breast cancer screening generally focus on women who are 40 years and older. Models examining treatment focus on patients diagnosed with disease. Specific subgroups may be defined by risk factors, such as race/ethnicity and family history or disease severity.

Some models are developed for very specific decision problems, while others are developed to address multiple problems. General purpose models tend to describe disease processes in greater detail, enabling modeling of the action of a wide range of possible interventions and capture of a wide range of possible outcomes. Therefore, models that are used for multiple decision problems tend to be more complex than more focused models. It can be difficult to choose the level of detail that will be described by the model. The modeler must strike a balance between simplicity, which eases communication of model assumptions, and complexity, which may increase face validity.

### Example: CRC Screening

Screening is an effective tool for reducing colorectal cancer incidence and mortality. Screening can detect colorectal cancer at an early stage, when there are better chances of survival (Hardcastle et al. 1996; Imperiale 2013; Kronborg et al. 1996; Towler et al. 1998) and can also detect adenomas, the predominant precursor lesion in colorectal cancer, leading to disease prevention through their removal. Professional societies, including the American Cancer Society, the US Multi-Society Task Force on CRC, and the American College of Radiology, recommend a variety of options for CRC screening, including annual fecal occult blood testing, flexible sigmoidoscopy every 5 years, and colonoscopy every 10 years (Levin et al. 2008; Rex et al. 2009; U. S. Preventive Services Task Force 2008). These tests differ in terms of costs, screening intervals, and invasiveness. A key question faced by patients, providers, and policy makers is how best to screen for colorectal cancer, that is, which test or sequence of tests is most effective for preventing death from colorectal cancer.

In spite of a great deal of accumulated evidence demonstrating the effectiveness of individual colorectal tests, it is difficult to directly compare the effectiveness of different screening regimens. Colorectal cancer is a rare event, and so estimation of the effectiveness of screening to reduce cancer incidence requires large samples sizes, and estimation of the effectiveness of screening to reduce colorectal cancer mortality requires long-term follow-up of this large sample. Direct comparison of multiple screening regimens requires even larger samples. For even short-term outcomes, it is not feasible to directly compare the wide range of potential screening regimens, which include combinations of different tests given at various screening intervals. Models allow researchers to combine available evidence to evaluate a specific decision problem: the effect of different screening regimens on (lifetime) colorectal cancer mortality. The decision problem is further refined by addressing our three questions.

### What Interventions Will Be Modeled?

To simplify this example, consider the impact of two screening interventions: colonoscopy every 10 years and annual fecal immunochemical test (FIT). For both interventions, screening begins at age 50 and continues up to and including age 75. Individuals with a positive FIT result are assumed undergo colonoscopy, with a return to annual FIT screening in 10 years if no adenomas or cancers are detected at colonoscopy. Any adenomas detected at colonoscopy are assumed to be completely removed. Consistent with clinical practice, both screening interventions refer individuals to adenoma surveillance based on findings at colonoscopy: individuals with one or two small (<10 mm) adenomas detected have their next colonoscopy in 5 years; individuals with three or more adenomas or any large ( $\geq 10$  mm) adenomas have their next colonoscopy in 3 years. These analyses simulate patients who are fully adherent to all test. However, models could be developed to examine the effect of differential adherence across screening regimens. For example, models could simulate individuals with different overall rates of adherence for each test type or different rates of patient dropout from the two screening regimens over time.

As part of specifying the intervention, the sensitivity and specificity need to be defined for each test and for detection of both precursor lesions and cancer. FIT was assumed to have 0.95 specificity, so that it results in a positive test 5% of the time when no disease is present, including precursor lesions. FIT was assumed to have sensitivity, the probability of detecting disease when it is present, that depends on adenoma size: 0.05 for adenomas 5 mm and smaller, 0.10 for adenomas larger than 5 mm and less than 10 mm, 0.22 for adenomas 10 mm and larger, and 0.70 for preclinical cancers of any size. Colonoscopy is an endoscopic tests that visually examines the entire large intestine (colon and rectum). Most but not all colonoscopies are complete, and lesions may be missed because they are beyond the reach of the endoscope. Colonoscopy was assumed to be complete to the cecum for 98% of exams. Tissue that is biopsied during colonoscopy is sent to pathology for definitive diagnosis, so colonoscopy has

perfect specificity. The sensitivity of colonoscopy was assumed to depend on the size of the lesion; the probability of missing a lesion that is  $s$  mm in diameter is given by  $P(\text{miss}|\text{size}=s \text{ and } \text{size} < 20) = 0.34 - 0.0349s + 0.0009s^2$ , with perfect sensitivity for adenomas 20 mm and larger. The associated miss rates for lesions that are 1 mm, 5 mm, 10 mm, and 15 mm in size are 31%, 19%, 8%, and 2%, respectively.

### What Events Are of Interest?

For this question, the key outcome event is colorectal cancer death. However, other-cause death also needs to be modeled to enable estimation of life-years saved and accurate description of the screened population. In addition, models need to describe the preclinical disease processes because screening can reduce mortality by its effect on two preclinical process: (1) by detecting cancer at an earlier stage, before it has become clinically detected (through presentation with symptoms), and (2) by preventing disease through detection and removal of precancerous lesions (adenomas). It will be important to describe adenoma size in this model because both the probability of screen-detecting an adenoma and the probability that an adenoma transitions to cancer increases with increasing adenoma size.

### What Is the Target Population and What Subgroups Are of Interest?

The decision problem in this example focuses on average risk individuals, who begin screening at age 50. Individuals at high risk for colorectal cancer, because of family history of colorectal cancer or diagnosis with genetic conditions, often begin screening at earlier ages.

## Step 2: Specify the Model Structure

Once the decision problem is defined, the modeler must **specify the model structure** (Roberts et al. 2012). The structure of the model is driven by the decision problem in combination with an understanding of the disease process, which may be rooted in empirical data representing the cumulative scientific knowledge. In this way, data may

indirectly inform the model; however, data availability should not necessarily determine a model's structure. The structure of the model must be sufficient to address the decision problem, and this may require description of processes that cannot be directly observed (such as tumor growth). If the model structure is not supported by data, this limited understanding of the underlying disease process should be noted. Processes that are not well supported by data can be explored through sensitivity analysis.

When specifying a microsimulation model, the modeler must choose *whether to model time as discrete or continuous*, the *distinct health states* that the model will describe, and *rules for transitioning between states*.

### Will Models Describe Events in Discrete or Continuous Time?

The decision to model time as continuous or discrete is closely tied to the **type of model** used for simulations. Different types of health policy models are described below, including some models that are not used for microsimulation.

**Decision trees** are a relatively simple models that are used to describe outcomes for groups of individuals (Petitti 2000). At each branching point, the tree specifies the probability of each subsequent outcome, for example, whether an individual has disease and, among people who have disease, whether a test is positive or negative. Using a decision tree, alternative courses of action are compared by calculating the expected value of the outcome resulting from each pathway (i.e., multiplying the value assigned to each potential outcome by the probability that each occurs). Because they do not explicitly incorporate time, decision trees are useful for simple decision problems with short time horizons, such as the short-term effects of diagnostic assessment, but they are not well suited to modeling of repeated events, such as a regimen of screening.

**State transition models** are more complex than decision trees and are useful for describing events over longer time frames than decision trees. State transition models incorporate time by updating state membership at discrete time intervals or *cycles*. Because only a single transition can

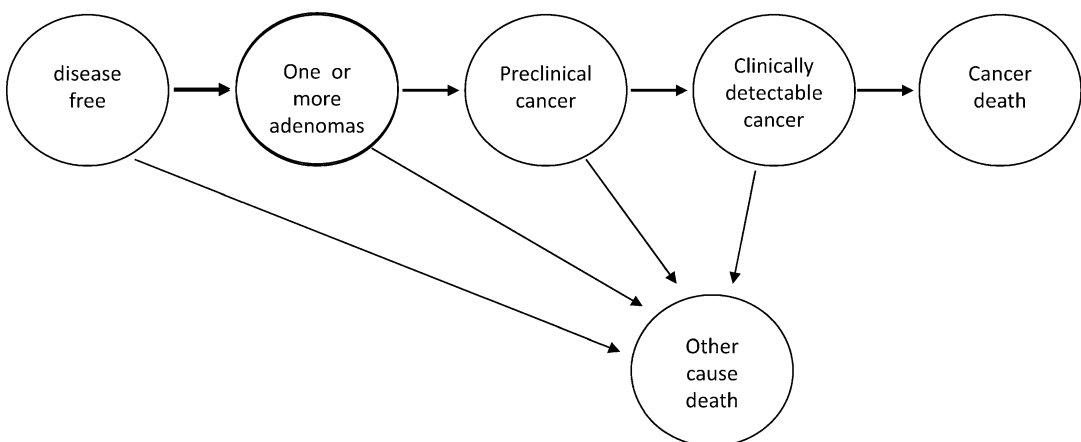
occur in each cycle, cycle length should be selected understanding that only one event can occur within a cycle. For example, if there are disease-free, preclinical disease, and clinical disease states, and individuals are required to pass through the preclinical state, then in one cycle individuals could transition from disease-free to preclinical disease, or from preclinical disease to clinical disease, but not from disease-free to clinical disease. Cycle length does not need to be uniform – it can depend on the state. However, shortening the cycle length for a given timeframe increases the total number of simulated transitions, increasing computational time.

State transition models that describe the transition of groups of individuals are called *Markov process models* (Beck and Pauker 1983; Siebert et al. 2012). Markov models assume that the probability of the transition from one state to the next depends only on the current state and is independent of prior history (i.e., how members got to the state). Because of this, Markov models are commonly described as “memory-less.” For example, when using a Markov model for screening, the probability of transition to the next screening test depends only on the outcome of the current test rather than the entire simulated screening history.

Markov process models assume that individuals who occupy the same state are homogeneous, that is, they are governed by the same rules for transitioning into the next health state. The

number of states must be increased when there is interest in patient subgroups with different transition probabilities that reflect differences in disease characteristics. The number of states can also increase when modelers relax the Markov assumption by carrying past health state information forward. Because of this, the number of disease states needed to adequately describe a disease process can quickly increase, a problem known as “state explosion.” As the number of states increases, Markov process models become intractable.

State transition models that describe the transition of individuals are a type of *microsimulation model*. Simulated individuals can be assigned characteristics (such as age, sex, or race), and the model can allow transitions to depend on these characteristics. By explicitly allowing individuals in the same state to be governed by different transition rules, microsimulation models are able to limit the total number of states. For example, consider the colorectal cancer model shown in Fig. 1, which includes six states: (1) alive and disease-free; (2) alive with one or more adenomas, but no cancer; (3) alive with preclinical cancer; (4) alive with detected cancer; (5) dead from colorectal cancer; and (6) dead from other causes. Suppose now that the model needs to allow all transitions to depend on sex. Using a Markov process model, this would require expansion to an 10 state model (assuming death states are the



**Fig. 1** Bubble graph showing the states and allowed transitions between states for the colorectal cancer model



same for men and women). In contrast, a state transition microsimulation model could describe this process using the same six states, by allowing transitions to depend on the sex of the simulated individuals, and some transitions could be modeled as identical for both men and women.

**Discrete event simulation (DES) models** are another type of microsimulation model that describe the movement individuals through distinct disease states in *continuous time* (Karnon et al. 2012). Discrete event simulation models are useful when modelers can better characterize transitions with time-to-event models than with transition probabilities over fixed periods. For example, when modeling disease incidence, a state transition model would specify incidence probabilities that are tied to the model's cycle length (e.g., annual incidence probabilities), while DES models could use time-to-event (survival) models to simulate the age at disease incidence.

**Models for infectious diseases** are more complicated because they describe transmission of disease between individuals, and therefore individuals are not independently simulated (Pitman et al. 2012). Two broad types of models are used to simulate infectious disease at the population-level: dynamic transition models and agent-based models. Dynamic transition models for infectious disease model groups of individuals and describe transitions using differential equations (Brauer and Castillo-Chavez 2013). These are also known as compartmental models, and they describe the transitions of individuals between compartments (or states) in continuous time. Agent-based models are an extension of discrete event simulation that allows interactions between individuals (Hunt et al. 2013; Luke and Stamatakis 2012). This chapter focuses on models that are useful for noninfectious diseases. However, many of the issues associated with DES and state transition microsimulation also apply to agent-based models.

**In summary:** State transition models describe individual disease trajectories in discrete time, with time periods given by cycle lengths. Discrete event simulation (DES) models describe individual disease trajectories in continuous time. Either

modeling approach can be used to simulate individuals with specific characteristics (such as age, sex, or race).

### What Distinct Disease States and Characteristics Will the Model Describe?

All models require specification of a set of *mutually exclusive disease states* that reflect the disease processes of interest, such as the six states shown for colorectal cancer in Fig. 1. This basic model must be expanded to evaluate endoscopic tests because large adenomas are easier to detect than small adenomas. Both state transition and DES models could address the need for adenoma size information by expanding the model to include the size of the largest adenoma (e.g., diminutive (<5 mm), small ( $5 \leq 10$  mm), or large ( $\geq 10$  mm)). Alternatively, DES models can describe adenoma growth as a continuous process, which essentially describes the time to reach various sizes. Modeling continuous growth requires assumptions about the nature of adenoma growth but allows flexibility in how adenoma size is incorporated into an intervention examined in the decision problem.

### When (and How) Do Simulated Individuals Transition Between States?

Rules for moving individuals between states in a state transition model are based on cycle length, that is, how often state memberships are updated, and are given by probabilities for each possible transition.

Rules for moving individuals between states in DES models are based on time-to-event distributions, life tables that characterize the time between successive events or, possibly, continuous growth. Time-to-event distributions take positive values on and include distributions typically used in survival analysis, such as exponential and Weibull distributions. While state transition models have a single type of parameter (transition probabilities), DES models can incorporate a range of parameter types that are associated with different time-to-event distributions.

**Example: Colorectal Cancer Model**

The ColoRectal Cancer Simulated Population model for Incidence and Natural history (CRC-SPIN) (Rutter and Savarino 2010) is used as the primary example in this chapter, with assumptions described in section “Example: CRC Screening.”

**Will the Model Describe Events in Discrete or Continuous Time?**

This example compares two different screening tests for colorectal cancer. Test performance depends on the number and size of adenomas. In addition, cancer incidence and survival both depend on age and sex. The CRC-SPIN model describes events in continuous time (discrete event simulation) enabling description of adenoma size and number using a limited number of states and allowing transitions to depend on age and sex.

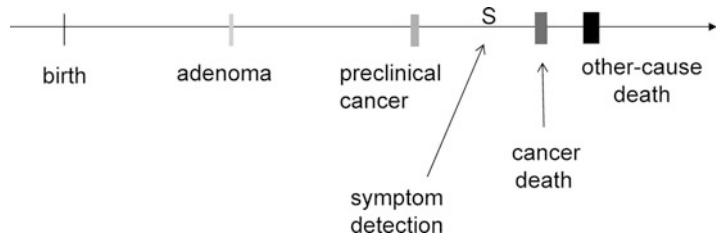
**What Distinct Disease States and Characteristics Will the Model Describe?**

The CRC-SPIN model (as shown in Fig. 1) describes six disease states. Individuals are

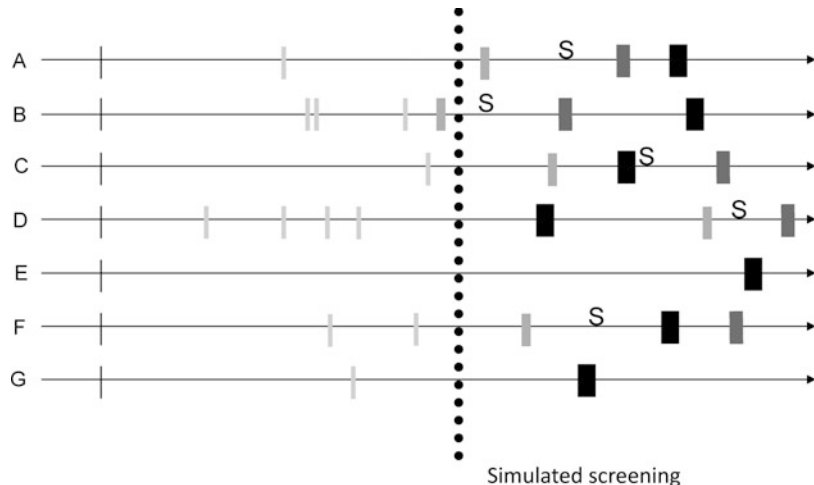
allowed to accumulate multiple adenomas while in the adenoma state and multiple adenomas and preclinical cancers while in the preclinical cancer state. In addition, the model simulates size and location characteristics for each adenoma. Figure 2 provides an example of the types of event histories the model will simulate, assigning date of birth as time zero across individuals. Figure 3 depicts a single screening event, at the same age for all individuals. For these hypothetical trajectories:

- Benefit is possible for trajectories A and B because screening has the potential to prevent disease through adenoma removal (A) or to detect cancer at a potentially earlier stage before it becomes symptomatic (B).
- Benefit is also possible for trajectory C, in terms of cancer incidence, because screening has the potential to avert symptomatic disease. However, for this trajectory screening does not improve survival because other-cause death is simulated to occur before cancer death.

**Fig. 2** Line graph showing a hypothetical sequence of simulated events in the colorectal cancer model



**Fig. 3** Line graph showing the simulated effect of screening in the colorectal cancer model, using symbols shown in Fig. 2



- There is no possible benefit for trajectories D, E, F, or G. Although screening has the potential to detect and remove adenomas (D) or to detect preclinical cancer (E), both trajectories simulate death before the cancer becomes symptomatic. For trajectory F, the simulated adenoma that could be detected at screening does not develop into preclinical cancer before other-cause death. For trajectory G, no disease events are simulated to occur.

**When (and How) Do Individuals Transitions Between States?**

This component of the model is made up of the mathematical functions and probability distributions that govern between state transitions. The following section describes between state transition rules for the CRC-SPIN model. Additional details are provided in Rutter and Savarino (2010).

The model describes the initiation of adenomas using on a nonhomogeneous Poisson process that allows adenoma risk to vary systematically by gender and age and to vary randomly across individuals. Under this model, the log-risk of developing an adenoma for the *i*th simulated individual is given by

$$\alpha_{0i} + \alpha_1 \text{sex}_i + \sum_{k=1}^4 \delta(A_k < \text{age}_i(t) \leq A_{k+1}) \times \left\{ \text{age}_i(t) \alpha_{2k} + \sum_{j=2}^k A_j (\alpha_{2,j-1} - \alpha_{2j}) \right\}$$

Here,  $\delta(\cdot)$  is an indicator function with  $\delta(x) = 1$  when *x* is true and  $\delta(x) = 0$ ; otherwise,  $\text{age}_i(t)$  is the *i*th individual’s simulated age at time *t*. Increases in adenoma risk with age are modeled with a piecewise linear function, with changes at  $A_1$ , with  $A_1 = 20, A_2 = 50, A_3 = 60, A_4 = 70, A_5 = 100$ .

Once an adenoma is initiated, the model assigns two characteristics: a location in the colorectum (colon/rectum) and a growth rate. Adenomas grow based on the Janoschek growth curve model, given by  $d_{ij}(t) = d_\infty - (d_\infty - d_0) \exp(-\lambda_{ij}t)$ , where  $d_{ij}(t)$  is the maximum diameter of the *j*th adenoma in the *i*th individual at time

*t* after initiation. The minimum detectable adenoma size is set to  $d_0 = 1$  mm, and the maximum adenoma size is set to  $d_\infty = 50$  mm.

Variation in growth across adenomas is allowed by varying the time it takes to reach 10 mm, given by  $t_{10} = -\ln((d_\infty - 10)/(d_\infty - d_0))/\lambda$ , allowing  $t_{10}$  to follow a type I extreme value distribution. Individuals can transition out of the adenoma state when adenomas are removed during colonoscopy. Individuals transition out of the adenoma state in two ways: (1) any adenoma transitioning to preclinical cancer, or (2) all adenomas are detected and removed during a colonoscopy exam.

Simulated individuals transition into the preclinical cancer state when any one of their adenomas becomes cancerous. For each adenoma, the model assigns a size at transition based on the lognormal distribution, with an expected size at transition that depends on location in the colon and rectum, gender, and age at initiation. Adenomas do not transition to preclinical cancer if the individual dies before the adenoma reaches transition size. Once in the preclinical cancer state, disease can be screen-detected, perhaps at an earlier stage than if it becomes clinical cancer, but the person cannot transition back to the disease-free or adenoma-only states.

Simulated individuals transition into the clinical cancer state when any preclinical cancer becomes clinically detected. Once the model simulates a preclinical cancer, the lesion is assigned a time to clinical cancer, based on a lognormal distribution that depends on location of the preclinical cancer (colon or rectum).

Once cancer is detected (clinically or through screening), the model assigns a stage at detection. For clinically detected cancers, stage is assigned using on the observed (SEER) stage distribution of clinically detected cancers. The model specifies that screen detection finds cancer at the same stage or earlier than clinical detection. Simulated individuals can only die from colorectal cancer after cancer is detected. Time from colorectal cancer diagnosis to death is based on survival probabilities based on analysis of SEER data and is a function of age at diagnosis, gender, stage at diagnosis, and year of diagnosis.

Individuals can transition to other-cause death from any health state (except cancer death). Data from national death registries were used to model other-cause death, as described in the following section.

### **Step 3: Identify Data Sources**

While a model's structure is driven by the decision problem at hand, data are required to inform the model so that it can be used for accurate predictions. The credibility of a model will be affected by the quality of the data that inform the model.

#### **Which Data Will Inform the Model?**

Which data will inform the model? Common data sources include registry data, published study results, unpublished study results, and cost data.

#### **Registry Data**

Registry data describe death and disease incidence. These data may directly inform transition probabilities. For example, national death registries provide good information about the time to other-cause death. Similarly, disease registries provide key information about incidence in a target population.

#### **Published Data**

Published data includes results from both randomized trials and observational studies of disease prevalence and characteristics and characteristics of modeled interventions. When selecting data sources, the modeler must consider potential biases. For example, individuals who choose to be screened for disease may be at higher risk because of their family history.

#### **Unpublished Data**

Unpublished data can provide a rich source of information at a greater level of detail than is possible from published sources or registry data. While useful for model development and evaluation, inclusion of unpublished data has the potential to reduce model transparency.

#### **Cost Data**

Cost data is incorporated into models that predict cost-effectiveness and may be needed for models that assume resource constraints.

#### **How Will Each Data Source Inform the Model?**

A model's credibility will also be affected by how data source are used to inform a model. Data can be incorporated into the model in three key ways: as an input, as a calibration point, or as a validation point. High-quality data sources are based on large sample sizes, are free from biases, and report on health states that are directly relevant to the model they inform. Few data sources meet these criteria, and so the modeler must decide how to best use limited data.

#### **Model Inputs**

Model inputs are set by the modeler and can include a range of basic information needed for simulations. Examples of model inputs include the percentage of simulated individuals who are female; characteristics of the intervention, such as the sensitivity and specificity of a screening test; and life tables that provide all-cause survival probabilities by sex and year of birth. Such model inputs are generally tied directly to data, with gender distributions coming from census information, sensitivity and specificity coming from published study results including meta-analysis, and life tables coming from registry data. Model inputs are pieces of information that can be directly integrated into the model.

#### **Calibration Targets**

Calibration targets are important statistics that cannot be directly integrated into the model. For example, it is important for a credible model to match observed disease rates as described by disease registry data. However, when disease incidence is the result of an accumulation of events, the modeler cannot directly incorporate this information as an input. Instead, model parameters are selected so that the model is able to reproduce calibration points.

### Validation Targets

Validation targets are similar to calibration targets. However, rather than being used to select model parameters, validation points are used to check the predictive ability of the model for new data, that is, data not used as inputs or for model calibration. Often, all data that are available at the time of model development are used for calibration, and validation points are obtained from new studies published after model development.

### Example: Colorectal Cancer Model

The CRC-SPIN model incorporates registry and published data. Two types of registry data inform the model. Data from the National Center for Health Statistics Databases is used to develop life tables that are used as an input to model other-cause death (National Center for Health Statistics 2000). Life table must be extrapolated to model life spans for individuals born more recently. The second type of registry data comes from the Surveillance Epidemiology and End Results (SEER) (U.S. Department of Health & Human Services 2012). This registry data provides observed incidence of colorectal cancer and the underlying SEER population in 1978, before the advent of colorectal cancer screening, to provide incidence the absence of screening by age, stage at diagnosis, and cancer location. The number of cancers within the target age range is used as calibration targets. Stage at diagnosis and survival information from SEER are used as a model inputs.

Several published data sources are used as calibration points in the model. These include results from studies describing adenoma prevalence (Rutter et al. 2007; Strul et al. 2006), adenoma count and size at detection (Lieberman et al. 2000; Pickhardt et al. 2003), preclinical cancer prevalence (Imperiale et al. 2000), studies of adenoma size, and presence of preclinical cancer (Church 2004; Odom et al. 2005). For example, data from adenoma case series (Church 2004) were used to inform the probability of transition of an adenoma to preclinical cancer as a function of size. These data describe the percentage of adenomatous lesions with preclinical cancer by size: among 666 lesions between 6 and 10 mm,

one was found to be a preclinical cancer, and among 673 lesions, over 10 mm 21 were found to be preclinical cancer.

The CRC-SPIN model does not use unpublished data or cost data.

## Step 4: Select Model Parameters

### Which Parameters Are “Inputs,” and Which Parameters Will Be Calibrated?

As mentioned previously, some model parameters are completely specified by the modeler. These are referred to as “inputs.” Inputs can include the age range of the target population, the percent of women in the population, or, for more detailed models, the distribution of risk factors in the target population. Model inputs are directly informed by data or, in the absence of data, expert opinion.

Other model parameters, which are the focus in the next section, are indirectly informed by observed data and may need to be inferred through a process called *model calibration*.

*Calibration* is used to select model parameters that result in predictions that are consistent with (or “fit”) calibration targets. Calibration is needed because calibration targets are not directly related to model parameters and therefore cannot be directly incorporated into the model, as inputs can be. Calibration may also be needed to reconcile multiple calibration targets, observed with error, that are not fully concordant. Finally, calibration provides a data-based approach to selection parameters that describe unobserved process. For example, information about the number and size of adenomas detected can provide information about two unobserved processes: the rate of adenoma initiation and the growth of adenomas size.

### Which Goodness of Fit Measure Will Be Used to Guide Calibration?

After setting calibration targets and identifying model parameters that will be calibrated, the modeler must select a calibration method. An important aspect of the calibration method is the measure of fit, that is, the statistic that will be

used to measure how close the model predictions are to the observed data. At least three measures can be used to measure goodness of fit (GOF): least squares, chi-squared, or likelihood methods (Vanni and Karnon 2011). Least squares minimize the sum of squared differences between predicted values,  $P_i$ , and observed values  $O_i$ . The chi-square approach scales these differences, for example, by dividing by the standard deviation of the observed data,  $\sigma_i : \sum \frac{(O_i - P_i)^2}{\sigma_i}$ . The goal of calibration is to minimize the distance between the observed and predicted values, that is, to minimize the least squares or chi-square statistics. A third common approach is to use the likelihood of the data at a specific parameter value,  $\hat{\theta}$ , that is, the probability of the observed data at  $\hat{\theta}$ . The goal of calibration is to maximize the likelihood. The likelihood approach requires specification of a probability distribution for observed data as a function of model parameters or simulation-based estimation of the likelihood at  $\hat{\theta}$  (Rutter et al. 2009).

### Which Calibration Method Will Be Used for Parameter Selection?

The next step in model calibration is selection of a search strategy. There are two primary approaches to model calibration: undirected and directed searches (Rutter et al. 2010).

#### Undirected Searches

Undirected searches involve exhaustive evaluation of the model at a defined set of points in the parameter space. Models with few parameters may be able to use a grid search. Using this approach, the modeler defines a grid of parameter values. The model is evaluated at every point on the grid. The best parameter set is chosen from these, as the parameter that provides the closest fit to the observed data. A related approach uses a randomly selected set of parameter values, with evaluation of the model at every point in this selected set. Undirected searches are theoretically easy to apply, but this approach is not computationally feasible for highly parameterized models, because the number of grid nodes grows exponentially with the number of model parameters. Furthermore, even a dense grid or a large random

sample of parameters might miss regions of good fit.

#### Directed Searches

Directed searches move through the parameter space by “hill climbing,” that is, moving in a direction of improving goodness of fit. If the functional form of the likelihood is available, then the algorithm can take steps in directions that are based on the derivative of the likelihood function, with movements in the direction of most rapid increase (“up the hill”). In general, micro-simulation models do not have closed form expression for these derivatives. This can be addressed by using approximations to the derivative or by using the Nelder-Mead algorithm, which does not require derivatives. Directed searches may find parameter values that provide locally, but not globally, good fit to calibration targets. To avoid this problem, directed searches should be initiated at multiple widely dispersed points within the parameter space. Directional searches for model calibration are generally more computationally efficient than grid search approaches, requiring fewer model runs for calibration.

---

### Implementation

Once the model is completely specified, it is ready to be used to address decision problems by generating predictions across a range of scenarios. A *model run* generally refers to a set predictions associated with a single set of model assumptions, including the parameters associated with transition probabilities and any interventions that the modeler has chosen to explore. A “base case” run generally refers to a run with assumptions that are believed to be most plausible.

### Example: Comparison of Two Tests to Screen for Colorectal Cancer

This section continues with the example comparing two approaches to screening for colorectal cancer: annual screening with a fecal

immunochemical stool test (FIT) and screening every 10 years with colonoscopy. Because these analyses focus on screening beginning at a particular age (50), rather than a screening program that begins in a particular year, the model is used to simulate a cohort of individuals who turned 50 in an arbitrarily selected year (2012). All simulated individuals were free of clinically detectable CRC on their 50th birthday. Model predictions are based on a single run of the model with ten million simulated individuals. Model parameters were calibrated using a likelihood-based approach (Rutter et al. 2009; Rutter and Savarino 2010).

Table 2 shows the predicted results for the no screening and two screening scenarios, focusing on the number of colorectal cancers detected and the number of colorectal cancer deaths. These outcomes were also used to predict the number of colorectal cancers prevented, the number of colorectal cancer deaths prevented, and life-years gained. Screening colonoscopies are defined to include primary screening exams, exams indicated because of a positive FIT result and exams that are part of short-interval follow-up.

The model predicts that screening annually with FIT or every 10 years with colonoscopy is both effective at reducing colorectal cancer incidence and deaths from colorectal cancer. Compared to FIT, for every 100,000 50-year-olds entering screening colonoscopy results in 0.22 fewer colorectal cancer cases, 0.13 fewer colorectal cancer deaths, and 1.6 more life-years gained but requires 255.7 more screening colonoscopies.

## Sensitivity Analysis

In some cases, a model parameter cannot be informed by data. In this case, the modeler may choose to select a specific value for the parameter and explore its effect on predictions through sensitivity analysis. Sensitivity analysis refers to model runs that systematically vary the values of model parameters, and modelers examine the sensitivity of the predictions to the choice of parameters values. Sensitivity analyses can also provide insight into the impact of specific model assumptions. For example, sensitivity analysis can be used to explore whether adenoma regression, which cannot be directly observed, is plausible by comparing predictions under specific scenarios, such as a model with no regression and model that assumes that 10% of adenomas regress (Lowe et al. 2004). Probabilistic sensitivity analysis places distributions on unknown parameters, providing a range of possible results. Parameters are sampled from specified distributions, and multiple model runs are used to infer variability in model predictions that result from variability in model parameters (Briggs et al. 2012; Cronin et al. 1998; Doubilet et al. 1985; Parmigiani 2002). Sensitivity analyses are common, largely because most models include unobservable components.

## Exploration and Description of Model Uncertainty

Models are used to predict unobserved outcomes based on imperfect knowledge, and

**Table 2** Simulated effect of screening for colorectal cancer, based on a cohort of individuals screened at age 50. The table below shows predictions per 100,000 individuals screened

	No screening	FIT every year	Colonoscopy every 10 years
Screen detected colorectal cancers	0	0.49	0.13
Clinically detected colorectal cancers	5.73	0.64	0.42
Colorectal cancer deaths	2.08	0.30	0.17
Colorectal cancers prevented	0	5.09	5.31
Colorectal cancer deaths prevented	0	1.78	1.91
Life-years gained	0	19.05	20.92
Number of screening Colonoscopies	0	173.4	429.1

these predictions are uncertain. Several sources of uncertainty have been identified and are described below (Briggs et al. 2012).

### **Stochastic, or “First-Order,” Uncertainty**

Stochastic, or “first-order,” uncertainty refers to the uncertainty that results from using a stochastic rather than deterministic decision model. Stochastic uncertainty is analogous to random error in regression models. Because modelers report average effects, simulating very large sample sizes can essentially eliminate stochastic error.

### **Parameter, or “Second-Order,” Uncertainty**

Parameter, or “second-order,” uncertainty refers to the uncertainty that results from having to calibrate model parameters and is related to the data that are available to inform parameters. Assessment of parameter uncertainty requires elimination of stochastic uncertainty, but parameter uncertainty is rarely reported because most model calibration is based on search strategies that do not directly provide standard error estimates. Instead, modelers sometimes use sensitivity analysis to address parameter variability, running the model at different parameter values and describing the relationship between parameter variability and variability in model predictions. Findings from this type of sensitivity analysis can be used to direct model improvement toward reducing variability of those parameters that have the greatest impact on prediction variability, for example, through additional data collection or, when appropriate, modifications to the model structure.

### **Systematic Variability or “Heterogeneity”**

Systematic variability or “heterogeneity” refers to variability that is built into the model. For example, a model may include systematic differences in the disease process or intervention effects that are a function of individual characteristics (age, sex, race, risk factors).

### **Structural Variability**

Structural variability refers to variability that results from the states selected and the rules for transitioning between states that are described by the model. Structural variability can be addressed using a single model through sensitivity analyses, focused on the most uncertain aspects of the model. This approach generally requires recalibration of each unique model. This “single model” approach is complicated because the model states and transition rules are often selected very deliberately and in consultation with experts. Structural variability can also be addressed through cross-validation or comparative modeling.

### **Model Validation**

Model validation is a critical component of model development. Validation is required to gain confidence in the model. There are five types of validity, outlined below: face validity, internal validity, cross-validity, external validity, and predictive validity (Eddy et al. 2012).

#### **Face Validity**

Face validity is subjective and refers to whether the model “makes sense.” Face validity of the model relates to the model structure and data used to inform the model. Face validity depends on model *transparency*, the clear description of the model structure and inputs.

To achieve face validity, models need both nontechnical and technical documentation. Nontechnical documentation should provide basic information about:

- **Model Structure:** The type of model, health states, and nontechnical descriptions of general rules for transitions between states.
- **Model Inputs:** This should include a description of inputs specified by the modeler to characterize the target population and inputs that are directly informed by observed data. Depending on the model, a description of the model parameters selected using calibration may or may not be useful. When models



include costs, the costs assigned to various actions and events in the model need to be clearly described.

- **Calibration Targets and Model Fit to Targets:** This provides information about observed information that the model is able to accurately simulate and how accurately the model simulates these data.

Technical documentation should be sufficiently detailed to enable others to reproduce the model, if they wish. This documentation should include:

- **Mathematical formulae for transition rules:** if the model is based entirely on fixed transition probabilities, then these should be provided.
- **Methods used for model calibration:** as this would enable others to reproduce the model.

While release of computer code is seemingly the most transparent approach, this strategy is time consuming and ultimately uninformative to the vast majority of end users so that code release may obscure rather than clarify the model assumptions.

### Internal Validity

Internal validity, or verification, refers to coding accuracy. Verification of code is a process that takes place within a modeling team and can be facilitated by modular programming to allow testing of specific blocks of code.

### Cross-Validation

Cross-validation, also known as comparative modeling, is based on comparing results obtained from different models and is the primary method for evaluating structural variability. Cross-validation provides a way to assess model predictions in the absence of observed or “gold standard” information and also provides a way of exploring unobserved or unobservable phenomena that are predicted by the model but cannot be validated against observed data such as predicted disease incidence in future years. Cross-validation may be reassuring when model predictions are similar, but when there are differences, cross-

validation does not provide a method for choosing the correct or best model.

The National Cancer Institute, through the CISNET group (National Cancer Institute), has championed the comparative modeling approach, by funding more than one modeling group to address policy questions. Examples of comparative modeling include estimation of the combined effects of screening and treatment on breast cancer mortality based on seven CISNET models for breast cancer (Berry et al. 2006) and the Mt. Hood Challenge comparing diabetes models (The Mount Hood 4 Modeling Group 2007). Each of these groups compared models only after standardizing the calibration targets. Without such cooperation, with each group simulating and presenting results under the same conditions, it can be difficult to directly compare model results. Cross-model comparisons can be very time consuming, involving coordination across modeling groups, and so are generally only practical for major policy questions.

### External Validation

External validation refers to how well the model is able to predict (or “fit”) existing data that was not used for model calibration. **Predictive validation** takes this idea a step further and refers to how well the model is able to predict study outcomes *before* they are observed. Among the validity measures discussed, external validity and predictive validity most closely correspond to the models’ purpose and therefore are critical to model confidence. Yet it is uncommon for models to carry out external or predictive validation exercises, largely because of data limitations.

Both external and predictive validation exercises require new data. For a model to be immediately validated after development, some data would have to be held out for validation. But because models are complex, modelers often need to use all available data to inform parameters. In some cases, modelers may validate to data that is partially dependent on calibration data, which represents a gray area between goodness of fit to calibration targets (sometimes referred to as internal calibration) and external validation. For example, a model may use overall disease

incidence rates by decades of age as a calibration target and then validate the model by predicting incidence rates by sex and age in years. To maintain trust in a model, it is critical that modelers be transparent about their validation approaches, clearly stating when partially dependent data are used for validation.

## In Conclusion

Microsimulation models are a powerful tool for systematically combining evidence from a variety of sources to provide critical information to health policy decision maker. Decision problems can be unconstrained, assuming unlimited resources, or they can be constrained to restrict resources such as total costs or treating physicians. The use of models to inform policy is increasing, partly due to increasing computational power but also because of increasing interest in evidence-based medicine. Yet there remain concerns about credibility of model predictions. These concerns are a natural consequence of the complexity of models and their focus on prediction, which requires extrapolation beyond available data. One way to build model credibility is to make model assumptions as transparent as possible. Another way to build credibility is through model predictions, that is, by comparing model predictions to observed data and, when possible, allowing end users to examine model predictions under different hypothetical scenarios.

## References

- Atkin WS, Edwards R, Kralj-Hans I, et al. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *Lancet*. 2010;375(9726):1624–33.
- Beck J, Pauker S. The Markov process in medical prognosis. *Med Decis Mak*. 1983;3:419–58.
- Berry DA, Inoue L, Shen Y, et al. Modeling the impact of treatment and screening on U.S. breast cancer mortality: a Bayesian approach. *J Natl Cancer Inst Monogr*. 2006; 36:30–6.
- Brauer F, Castillo-Chavez C. *Mathematical models for communicable diseases*. Philadelphia: Society for Industrial and Applied Mathematics; 2013.
- Briggs AH, Weinstein MC, Fenwick EA, et al. Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM modeling good research practices Task Force Working Group-6. *Med Decis Mak*. 2012;32(5):722–32.
- Cancer Incidence – Surveillance, Epidemiology, and End Results (SEER) Registries Research Data [database on the Internet]. National Cancer Institute, Surveillance Systems Branch. 2012. Available from: <http://seer.cancer.gov/data/seerstat/nov2011/>.
- Caro JJ, Briggs AH, Siebert U, et al. Modeling good research practices – overview: a report of the ISPOR-SMDM modeling good research practices Task Force-1. *Med Decis Mak*. 2012;32(5):667–77.
- Church JM. Clinical significance of small colorectal polyps. *Dis Colon Rectum*. 2004;47(4):481–5.
- CISNET. 2014. Available at: <http://cisnet.cancer.gov>. Accessed 30 Apr 2014.
- Cronin KA, Legler JM, Etzioni RD. Assessing uncertainty in microsimulation modelling with application to cancer screening interventions. *Stat Med*. 1998;17(21): 2509–23.
- Doubilet P, Begg CB, Weinstein MC, et al. Probabilistic sensitivity analysis using Monte Carlo simulation. A practical approach. *Med Decis Mak*. 1985;5(2): 157–77.
- Eddy D. Breast cancer screening for Medicare beneficiaries: effectiveness, costs to Medicare and medical resources required. Washington, DC: U.S. Congress, Health Program, Office of Technology Assessment; 1987.
- Eddy DM, Hollingworth W, Caro JJ, et al. Model transparency and validation: a report of the ISPOR-SMDM modeling good research practices Task Force-7. *Med Decis Mak*. 2012;32(5):733–43.
- Etzioni R, Penson DF, Legler JM, et al. Overdiagnosis due to prostate-specific antigen screening: lessons from U.S. prostate cancer incidence trends. *J Natl Cancer Inst*. 2002;94(13):981–90.
- Hardcastle JD, Chamberlain JO, Robinson MH, et al. Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. *Lancet*. 1996;348 (9040):1472–7.
- Hixson LJ, Fennerty MB, Sampliner RE, et al. Prospective study of the frequency and size distribution of polyps missed by colonoscopy. *J Natl Cancer Inst*. 1990; 82(22):1769–72.
- Hunt CA, Kennedy RC, Kim SH, et al. Agent-based modeling: a systematic assessment of use cases and requirements for enhancing pharmaceutical research and development productivity. *Wiley Interdiscip Rev Syst Biol Med*. 2013;5(4):461–80.
- Imperiale TF. Sigmoidoscopy screening: understanding the trade-off between detection of advanced neoplasia and diagnostic efficiency. *J Natl Cancer Inst*. 2013; 105(12):846–8.
- Imperiale TF, Wagner DR, Lin CY, et al. Risk of advanced proximal neoplasms in asymptomatic adults according to the distal colorectal findings. *N Engl J Med*. 2000;343(3):169–74.
- Johnson CD, Chen MH, Toledano AY, et al. Accuracy of CT colonography for detection of large adenomas and cancers. *N Engl J Med*. 2008;359(12):1207–17.
- Karnon J, Stahl J, Brennan A, et al. Modeling using discrete event simulation: a report of the ISPOR-SMDM

- modeling good research practices Task Force-4. *Value Health*. 2012;15(6):821–7.
- Knudsen AB, Lansdorp-Vogelaar I, Rutter CM, et al. Cost-effectiveness of computed tomographic colonography screening for colorectal cancer in the Medicare population. *J Natl Cancer Inst*. 2010; 102(16):1238–52.
- Kronborg O, Fenger C, Olsen J, et al. Randomised study of screening for colorectal cancer with faecal-occult-blood test. *Lancet*. 1996;348(9040):1467–71.
- Lansdorp-Vogelaar I, Kuntz KM, Knudsen AB, et al. Stool DNA testing to screen for colorectal cancer in the Medicare population. A cost-effectiveness analysis. *Ann Intern Med*. 2010;153(6):368–77.
- Levin B, Lieberman DA, McFarland BG, et al. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-society task force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology*. 2008; 134(5):1570–95.
- Lieberman DA, Weiss DG, Bond JH, et al. Use of colonoscopy to screen asymptomatic adults for colorectal cancer. Veterans affairs cooperative study group 380. *N Engl J Med*. 2000;343(3):162–8.
- Loeve F, Boer R, Zauber AG, et al. National Polyp Study data: evidence for regression of adenomas. *Int J Cancer*. 2004;111(4):633–9.
- Luke DA, Stamatakis KA. Systems science methods in public health: dynamics, networks, and agents. *Annu Rev Public Health*. 2012;33:357–76.
- Mandelblatt J, Schechter C, Levy D, et al. Building better models: if we build them, will policy makers use them? Toward integrating modeling into health care decisions. *Med Decis Mak*. 2012;32(5):656–9.
- Muller CM, Mandelblatt J, Schechter C. The cost and effectiveness of cervical cancer screening in elderly women. Washington, DC: Congress of the United States, Office of Technology Assessment; 1990.
- National Cancer Institute. Cancer Intervention and Surveillance Modeling Network (CISNET). n.d.. Available at: <http://cisnet.cancer.gov/>. Accessed 2008.
- National Center for Health Statistics. US Life Tables. 2000.; Available at: [www.cdc.gov/nchs/products/pubs/pubd/iftbls/life/1966.htm](http://www.cdc.gov/nchs/products/pubs/pubd/iftbls/life/1966.htm). Accessed 2013.
- Odom SR, Duffy SD, Barone JE, et al. The rate of adenocarcinoma in endoscopically removed colorectal polyps. *Am Surg*. 2005;71(12):1024–6.
- Parmigiani G. Measuring uncertainty in complex decision analysis models. *Stat Methods Med Res*. 2002;11(6): 513–37.
- Petitti DB. Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine. 2nd ed. New York: Oxford University Press; 2000. 306 p.
- Pickhardt PJ, Choi JR, Hwang I, et al. Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults. *N Engl J Med*. 2003;349(23):2191–200.
- Pitman R, Fisman D, Zaric GS, et al. Dynamic transmission modeling: a report of the ISPOR-SMDM modeling good research practices Task Force Working Group-5. *Med Decis Mak*. 2012;32(5): 712–21.
- Rex DK, Cutler CS, Lemmel GT, et al. Colonoscopic miss rates of adenomas determined by back-to-back colonoscopies. *Gastroenterology*. 1997;112(1): 24–8.
- Rex DK, Johnson DA, Anderson JC, et al. American College of Gastroenterology guidelines for colorectal cancer screening 2009 [corrected]. *Am J Gastroenterol*. 2009;104(3):739–50.
- Roberts M, Russell LB, Paltiel AD, et al. Conceptualizing a model: a report of the ISPOR-SMDM modeling good research practices Task Force-2. *Med Decis Mak*. 2012;32(5):678–89.
- Rutter CM, Savarino JE. An evidence-based micro-simulation model for colorectal cancer. *Cancer Epidemiol Biomark Prev*. 2010;19(8):1992–2002.
- Rutter CM, Yu O, Miglioretti DL. A hierarchical non-homogenous Poisson model for meta-analysis of adenoma counts. *Stat Med*. 2007;26(1):98–109.
- Rutter CM, Miglioretti DL, Savarino JE. Bayesian calibration of microsimulation models. *J Am Stat Assoc*. 2009;104(488):1338–50.
- Rutter CM, Zaslavsky AM, Feuer EJ. Dynamic micro-simulation models for health outcomes: a review. *Med Decis Mak*. 2010;31(1):10–8.
- Siebert U, Alagoz O, Bayoumi AM, et al. State-transition modeling: a report of the ISPOR-SMDM modeling good research practices Task Force-3. *Med Decis Mak*. 2012;32(5):690–700.
- Strul H, Kariv R, Leshno M, et al. The prevalence rate and anatomic location of colorectal adenoma and cancer detected by colonoscopy in average-risk individuals aged 40–80 years. *Am J Gastroenterol*. 2006;101(2):255–62.
- The Mount Hood 4 Modeling Group. Computer modeling of diabetes and its complication: a report on the 4th Mount Hood challenge meeting. *Diabetes Care*. 2007;30:1638–46.
- Towler B, Irwig L, Glasziou P, et al. A systematic review of the effects of screening for colorectal cancer using the faecal occult blood test, hemoccult. *BMJ*. 1998;317(7158):559–65.
- U. S. Preventive Services Task Force. Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med*. 2008;149(9):627–37.
- van der Akker-van Marle ME, van Ballegooijen M, van Ootmarsen GJ, et al. Cost-effectiveness of cervical cancer screening: comparison of screening policies. *J Natl Cancer Inst*. 2002;94:193–204.
- Vanni T, Karnon J, Madan J, et al. Calibrating models in economic evaluation: a seven-step approach. *PharmacoEconomics*. 2011;29(1):35–49.
- Vogelaar I, Van Ballegooijen M, Schrag D, et al. How much can current interventions reduce colorectal cancer mortality in the U.S.? *Cancer*. 2006; 107:1623–33.



Georgia Salanti, Deborah Caldwell, Anna Chaimani, and  
Julian Higgins

## Contents

<b>Introduction</b> .....	578
Example: Incident Diabetes with Antihypertensive Drugs .....	579
A Roadmap to the Chapter .....	579
<b>Meta-analysis of Head-to-Head Comparisons</b> .....	580
Types of Data that Feed into a Meta-analysis .....	580
Meta-analysis and Meta-regression as Linear Model .....	580
Meta-analysis as Hierarchical Model .....	581
Fitting the Meta-analysis Model .....	582
<b>Indirect and Mixed Comparison</b> .....	584
Theory and Formulae for Indirect Comparisons .....	584
Theory and Formulae for Mixed Comparisons .....	586
Assumptions Underlying Indirect and Mixed Comparisons .....	587
<b>Models for Network Meta-analysis</b> .....	591
Consistency Models .....	592
Assumptions of Network Meta-analysis .....	602
Statistical Methods to Detect Inconsistency in a Network of Interventions .....	603
Inconsistency Models .....	605

---

G. Salanti (✉) · A. Chaimani  
Department of Hygiene and Epidemiology, University of  
Ioannina School of Medicine, Ioannina, Greece  
e-mail: [georgia.salanti@ispm.unibe.ch](mailto:georgia.salanti@ispm.unibe.ch);  
[annachaimani@gmail.com](mailto:annachaimani@gmail.com)

D. Caldwell  
School of Social and Community Medicine, University of  
Bristol, Bristol, UK  
e-mail: [d.m.caldwell@bristol.ac.uk](mailto:d.m.caldwell@bristol.ac.uk)

J. Higgins  
MRC Biostatistics Unit, Cambridge, UK  
Centre for Reviews and Dissemination, University of York,  
York, UK  
e-mail: [julian.higgins@bristol.ac.uk](mailto:julian.higgins@bristol.ac.uk)

Exploring Heterogeneity and Inconsistency: Network Meta-regression .....	608
<b>Numerical and Graphical Presentation of Results from Network Meta-analysis</b> .....	611
<b>References</b> .....	613

### Abstract

The increasing number of alternative treatment options for the same condition created the need to undertake reviews that address complex policy-relevant questions and make inferences about many competing treatments. Such reviews collect data which, under conditions, can be statistically synthesized using network meta-analysis. This chapter presents the basic concepts of indirect and mixed comparison of treatments and presents the statistical models for network meta-analysis and their implementation both theoretically and in examples. The assumption underlying network meta-analysis is extensively discussed and extensions of the models to account for effect modifiers are presented.

limited. Moreover, although clinical and policy making interest lies in comparing active agents, new drugs are commonly compared with placebo in order to obtain marketing authorization. Given that clinical practice changes over time and that licensed or reference treatments differ across countries, it is unrealistic to expect that individual trials and pairwise meta-analyses can provide evidence of comparative effectiveness for every intervention of interest.

The need to compare multiple competing treatments to inform clinical guidelines and health technology appraisals has underpinned the development of network meta-analysis. Also known as a multiple treatment meta-analysis and mixed treatment comparisons, a network meta-analysis simultaneously combines direct and indirect information across a network of studies to make inferences regarding the relative effectiveness of multiple interventions. An indirect comparison, which underpins the method, is a simple idea: treatment A can be compared with treatment B via a common comparator C, by statistically combining the comparison A versus C (AC) and B versus C (BC) studies. Several applications and methods papers have outlined the benefits of combining direct and indirect evidence in a network meta-analysis (Caldwell et al. 2005; Cooper et al. 2011; Hoaglin et al. 2011; Mills et al. 2011). These include improvement in precision for the estimated effect sizes and the ability to compare treatments that have not been directly compared in any trial. Despite the increasing number of applications, network meta-analysis is far from being an established practice. Many authors emphasize the secondary or supplementary nature of the analyses, giving priority to direct evidence (NICE 2008; Edwards et al. 2009). Network meta-analyses are often considered controversial (Piccini and Kong 2011; Thijs et al. 2008), for example, a recent evaluation of the relative

## Introduction

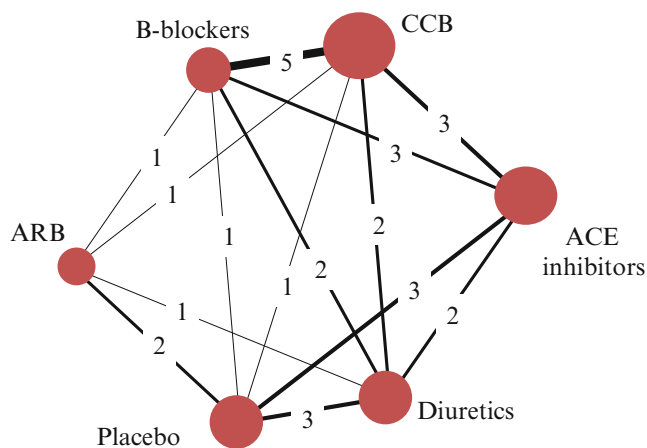
Meta-analyses of randomized controlled trials (RCTs) are often considered to provide the most reliable and valid evidence on which to base healthcare decisions, usually ranking above single RCTs in evidence-based medicine (EBM) hierarchies of evidence (Higgins and Green 2008). Meta-analysis is an integral part of EBM, used by international health organizations such as the World Health Organization and The Cochrane Collaboration, and is widely used to inform health technology assessment and clinical guidelines produced by organizations such as the Canadian Agency for Drugs and Technologies in Health (CADTH), the National Institute for Health and Clinical Excellence (NICE), and the Agency for Healthcare Research and Quality (AHRQ). However, as meta-analysis traditionally compares two treatments at a time (a pairwise comparison) its use in the presence of multiple competing treatment options is

effectiveness of twelve new-generation antidiabetic drugs attracted supporters as well as skeptics (Barbui et al. 2009; Cipriani et al. 2009).

### Example: Incident Diabetes with Antihypertensive Drugs

To exemplify all methodologies of this chapter, a published network meta-analysis (Elliott and Meyer 2007) will be used. It is based on a systematic review that aimed to compare antihypertensive drugs with respect to the incidence of diabetes. The review included 22 randomized controlled trials. The competing interventions are placebo (P),  $\beta$ -blockers (BB), diuretics (D), calcium channel blockers (CCB), angiotensin converting enzyme (ACE) inhibitors and angiotensin II receptor blockers (ARB).

Four studies include three arms and the rest are two-arm trials. All comparisons have been evaluated in at least one study except for the comparisons ARB versus ACE inhibitors for which no studies exist. Figure 1 shows the plot of this network of intervention comparisons.



**Fig. 1** Plot of network for incidence of diabetes. The size of the nodes is proportional to the number of studies that evaluate each intervention and the thickness of the lines is proportional to the frequency of each comparison in the

### A Roadmap to the Chapter

The chapter starts by setting up notation and the two most commonly applied models used for meta-analysis and meta-regression of pairwise comparison and discusses the frequentist and Bayesian implementation of the models (section “[Meta-analysis of Head-to-Head Comparisons](#)”). Section “[Indirect and Mixed Comparison](#)” describes the theory of indirect comparisons and the combination of these with direct comparisons (sometimes called “mixed” comparisons) in a simple three-treatment network consisting of trials that compare any two of the three treatments. The key assumption required to derive valid indirect and mixed estimates is extensively discussed in section “[Assumptions Underlying Indirect and Mixed Comparisons](#).” Section “[Models for Network Meta-analysis](#)” is more technical and describes the models used to fit network meta-analysis and discusses statistical methods to detect and account for violation of the key assumption. Section “[Models for Network Meta-analysis](#)” concludes by outlining extensions of network meta-analysis to account for the impact of effect modifiers. Section “[Numerical and Graphical Presentation of Results](#)”

network. The numbers represent the number of studies including each comparison (*CCB* calcium channel blockers, *ARB* angiotensin receptor blockers, *ACE* angiotensin converting enzyme)

from Network Meta-analysis” reviews numerical and graphical methods for presenting results from network meta-analysis to assist clinicians with interpretation of findings.

---

## Meta-analysis of Head-to-Head Comparisons

Pairwise meta-analysis summarizes the relative effectiveness of two interventions across  $N$  studies. Two basic parametric models are usually used: the fixed-effect model and the random effects model. Under the fixed-effect assumption, it is considered that all studies estimate the same underlying treatment effect. In the random effects model, it is assumed that there is a study-specific treatment effect underlying each study and that the observations from different studies estimate these different underlying effects. The study-specific underlying effects can be different yet related, and it is assumed that they “belong” to the same distribution. The variance of this distribution is the heterogeneity parameter describing the magnitude of the between-study variation. Meta-analysis can be viewed as special case of a weighted linear regression or as a hierarchical model. Both models are equivalent; though linear regression approaches are the most common approach in a frequentist implementation when treatment effect estimates are the starting point of the analysis (known as a “contrast-based” approach), and hierarchical approaches are usually encountered when summary data from each treatment group are the starting point of the analysis (an “arm-based” approach), often fitted in a Bayesian framework. These ideas are discussed in detail in the following three subsections.

### Types of Data that Feed into a Meta-analysis

The systematic review process first requires identification and appraisal of studies that address the research question of interest. Then, relevant data are extracted from the studies that

fulfill the predefined inclusion criteria. Consider that  $N$  studies, indexed with  $i = 1, \dots, N$  and comparing two treatments.  $A$  and  $B$  are included that contribute data for a particular outcome of interest. The data from each study can be *arm based* or *contrast based*. The first term refers to data that apply to each arm; for dichotomous outcomes these can be the number of successes  $r_{iA}, r_{iB}$  out of the total randomized  $n_{iA}, n_{iB}$  for the arms  $A$  and  $B$ , respectively. For continuous outcomes, the arm-based data are the outcome means  $m_{iA}, m_{iB}$ , standard deviations  $sd_{iA}, sd_{iB}$  and total numbers of participants  $n_{iA}, n_{iB}$  per arm.

Instead of presenting the outcome in each arm separately, a study can report the difference in the outcome between the two arms using a statistic. The *contrast-based* approach refers to study-specific statistics that compare the two arms. With dichotomous outcomes, the statistics are usually the odds ratios (OR), risk ratios, risk differences, or hazard ratios, whereas for continuous outcomes, it is usually mean differences, standardized mean difference, or ratios of means. The logarithmic transformation of ratio measures (e.g., odds and risk ratios) is typically applied in practice. Let  $y_{iAB}$  be generic notation for one of these statistics, which will be referred to as “the effect size.” The sample variance of an effect size will be denoted with  $s_{iAB}^2$ . Of course, arm-specific data can be transformed into contrast-based data before the start of the analysis. However, modeling arm-specific data is often an advantage in terms of model fit and therefore detailed data, if available, should be given preference.

### Meta-analysis and Meta-regression as Linear Model

Meta-analysis can be viewed as a linear regression model with no covariates. As each observation represents a study and these studies typically have different sample sizes, it is reasonable to weight the observations accordingly; hence, meta-analyses are fitted as a weighted linear

model. In a random effects meta-analysis, a study's effect size is given by

$$y_i = \mu + \delta_i + e_i \tag{1}$$

where  $\mu$  is the summary treatment effect and the random errors are assumed normally distributed

$$e_i \sim N(0, s_i^2) \tag{2}$$

The quantities  $\delta_i$  account for random variation in the treatment effects across studies (heterogeneity) and are assumed to be normally distributed as  $\delta_i \sim N(0, \tau^2)$ . Setting either the heterogeneity variance  $\tau^2$  to be zero or all  $\delta_i = 0$  reduces the model to the fixed-effect model.

Equation 1 can be extended into meta-regression in order to account for variability in the effect sizes with respect to a trial-specific variable  $x_i$ :

$$y_i = \mu_1 + \mu_2 x_i + \delta_i + e_i$$

When  $x_i$  is a categorical variable then the meta-regression model is equivalent to subgroup analysis. Consider, for example, that the systematic review comprises studies with appropriate and inappropriate blinding (subgroups 1 and 2). Then, using the dichotomous variable  $x_i$  as an index variable which takes values 0 for appropriate and 1 for inappropriate blinding, a subgroup analysis via meta-regression can be fitted. Then, the summary effect  $\mu_1$  would be the summary estimate from the subgroup of appropriately blinded studies and  $\mu_1 + \mu_2$  would be the summary estimate from the subgroup of inappropriately blinded studies. In a general framework of  $F$  subgroups indexed with  $f$ , a meta-regression model can be fitted without intercept:

$$y_i = \sum_{f=1}^F \mu_f x_{if} + \delta_i + e_i \tag{3}$$

where now the regression coefficients  $\mu_f$  are the summary estimates in subgroups.

### Meta-analysis as Hierarchical Model

An alternative representation of the random effects meta-analysis model is to consider two levels of estimation hierarchy: one level for the observation in each study that estimates the study-specific underlying effect and a second level for all the study-specific underlying effects that arise form a common distribution centered around the meta-analysis summary effect. Specifically, in each study the observed effect size  $y_i$  is assumed normally distributed with mean equal to the underlying effect size  $\theta_i$ , and uncertainty reflected by the sample variance:

$$y_i \sim N(\theta_i, s_i^2) \tag{4}$$

Then, it is assumed that the underlying  $\theta_i$  form a common distribution with expectation  $\mu$ . The variance of the distribution is the heterogeneity:

$$\theta_i \sim N(\mu, \tau^2) \tag{5}$$

The equivalence between the two alternative representations (linear and hierarchical model) is seen by identifying  $\theta_i$  with  $\delta_i + \mu$ . The fixed effects model can be obtained by substituting  $\theta_i = \mu$  into distribution (5).

The hierarchical model presented in this section can be used to model arm-specific data instead of effect sizes. This offers the advantage that the true likelihood of the data can be used and bypasses the assumption of normality for the observed effect sizes (as reflected in distributions (2) and (4)), often yielding better fit of the models. For example, when the outcome is dichotomous the normal likelihood in Eq. 5 can be substituted by two binomial likelihoods:

$$r_{iA} \sim B(p_{iA}, n_{iA})$$

$$r_{iB} \sim B(p_{iB}, n_{iB})$$

Then the probabilities of success in the two arms can be parameterized to derive contrast-specific parameters  $\theta_i$  using a link function  $\varphi$ :



$$\varphi(p_{iA}) = u_i$$

$$\varphi(p_{iB}) = u_i + \theta_i$$

For instance, the underlying study-specific treatment effect  $\theta_i$  can be the log-odds ratio if  $\varphi$  is the logit function or the log-risk ratio if  $\varphi$  is the logarithmic function. For more details, see (Warn et al. 2002).

When the outcome is continuous, distribution (4) is substituted by the two normal distributions for the means in the two arms:

$$m_{iA} \sim N\left(\lambda_{iA}, (sd_{iA})^2 / \sqrt{n_{iA}}\right)$$

$$m_{iB} \sim N\left(\lambda_{iB}, (sd_{iB})^2 / \sqrt{n_{iB}}\right)$$

Then, the effect size can be derived by parameterizing the two means  $\lambda_{iA}$  and  $\lambda_{iB}$ ; for example, the mean difference could be derived as  $\theta_i = \lambda_{iA} - \lambda_{iB}$

For either type of data (dichotomous or continuous) and for any statistic, the underlying effects  $\theta_i$  are assumed to arise from a common distribution as in (5).

## Fitting the Meta-analysis Model

The meta-analysis models above can each be fitted within a frequentist or a Bayesian framework. This section briefly summarizes the practical differences between the two approaches and the implications they might have for the summary estimates. For a more detailed overview of the Bayesian methodology, the reader should refer to Spiegelhalter et al. (2004) and Sutton and Abrams (2001). The choice between the different frameworks depends primarily on familiarity with the required software and methods.

The main practical differences between frequentist and Bayesian implementations relate to how the methods estimate the heterogeneity. In most frequentist implementations, the parameter  $\tau$  is assumed “known” and several estimation approaches have been proposed including the method of moments and restricted maximum

likelihood (see Viechtbauer 2007). Accounting for uncertainty in the estimation of heterogeneity is possible, but most existing software does not include uncertainty for  $\tau$  in standard meta-analysis routines. The frequentist estimates invariably perform poorly when few studies are included in the meta-analysis. In a Bayesian framework,  $\tau$  may easily be treated as a random variable and is given a prior distribution which, combined with likelihood statement, provides inference on the (posterior) distribution of the heterogeneity parameter. Therefore, uncertainty about the estimation of  $\tau$  is always introduced and impacts on the results. However, with few studies, Bayesian estimation of heterogeneity is also problematic because the choice of the prior distribution may have considerable impact on the results since little information is provided from the data (Lambert et al. 2005). In such cases, it is particularly advisable to carry out sensitivity analyses.

In a Bayesian framework the fit of the model to the data can be measured by calculating the posterior mean residual deviance  $\bar{D}$ . The model fits the data adequately when  $\bar{D}$  approximates the number of unconstrained data points (e.g., the number of studies when the contrast-based approach is used in a head-to-head meta-analysis). The deviance information criterion (*DIC*) is the sum of  $\bar{D}$  and the effective number of parameters,  $pD$ , and provides a measure of model fit penalized for model complexity (Spiegelhalter et al. 2002). It has an interpretation similar to the Akaike information criterion: lower values of the *DIC* suggest a better compromise between model fit and complexity. A difference in *DIC* of three units or more is usually considered important. *DIC* can be used to compare different models as long as they are applied to the same amount of data. For example, *DIC* can be used to select between different meta-regression models to choose between consistency and different inconsistency models, as will be discussed later.

An advantage of the Bayesian fitting of the models is that the posterior distribution can be directly interpreted as the probability distribution of the quantity of interest (e.g., summary effect, heterogeneity). Consequently, probabilistic statements follow naturally; it is straightforward to

calculate probabilities of one treatment being better than the other, or outperforming another by a specific magnitude. This is an important advantage when many treatments need to be compared and pairwise presentation of effect sizes becomes cumbersome. Calculation of probabilities is possible in a frequentist setting via resampling techniques, but this typically requires specialized routines or extra programming for the user.

Several software options exist that fit meta-analysis models in a frequentist setting. Freely available software includes RevMan and packages in R; a popular commercial option is STATA. The available routines and software frame the flexibility of models; for instance, it is not possible to fit arm-specific data using their exact likelihood with the existing meta-analysis specific routines.

With network meta-analysis increasing in popularity, Bayesian approaches have become popular as they offer greater flexibility, and WinBUGS is the most common software used. Meta-analysis can be fitted as a linear or hierarchical model and both arm-specific or contrast specific data can be modelled, giving Bayesian fitting a practical advantage compared to the frequentist approach.

**Example: Subgroup Meta-analysis for ACE Inhibitors and CCB Versus  $\beta$ -Blockers**

To exemplify the methods outlined above, consider the two comparisons CCB versus  $\beta$ -blockers and ACE inhibitors versus  $\beta$ -blockers from the network introduced earlier relating to incident diabetes. Firstly, a meta-regression model will be fitted,

with dummy variables, to carry out subgroup analysis on contrast-based data (the  $\ln(OR)$  for diabetes from each study), using the treatments being compared to define two subgroups. There are three studies comparing ACE inhibitors versus  $\beta$ -blockers and five comparing CCB versus  $\beta$ -blockers. Although a regression model is usually written with an intercept and one or more regression terms, it can also be written with no intercept as in Eq. 3. The eight observed  $\ln(OR)$  estimates are denoted  $asy_i$  using study indices,  $i = 1, 2, \dots, 8$ . Each  $y_i$  is then written as a function of the variables  $x_{iACE-BB}$  and  $x_{iCCB-BB}$ . These variables take values  $x_{iACE-BB} = 1$  if study  $i$  compares ACE inhibitors versus  $\beta$ -blockers and  $x_{iACE-BB} = 0$  otherwise, and  $x_{iCCB-BB} = 1$  for CCB versus  $\beta$ -blockers and zero otherwise. The meta-regression model that gives the summary effects for these two comparisons is

$$y_i = \mu_{ACE-BB}x_{iACE-BB} + \mu_{CCB-BB}x_{iCCB-BB} + \delta_i + e_i$$

where  $\delta_i$  is the study-specific random effect. Fitting this model in STATA using the command `metareg` and specifying the method of moments as the method to estimate the heterogeneity parameter produces the results shown in the upper part of Table 1.

The coefficients  $\mu$  of the regression are the subgroup-specific summary effects  $\mu_{ACE-BB}$ ,  $\mu_{CCB-BB}$  on the  $\ln(OR)$  scale. The heterogeneity parameter was estimated as  $\tau^2 = 0.01$  and the proportion of variability due to heterogeneity rather than sampling error (after accounting for subgroup differences) as  $I^2 = 59\%$ .

**Table 1** Results of subgroup analysis for ACE inhibitors versus  $\beta$ -blockers and CCB versus  $\beta$ -blockers. Log-odds ratios ( $\beta$ ) with their standard error  $SE(\beta)$  and odds ratios

( $OR$ ) with their 95% confidence or credible interval (CI/CrI) estimated from meta-regression and hierarchical models are reported

Model	Comparison	$\mu$	$SE(\mu)$	OR	95% CI/CrI for OR
Linear model in frequentist implementation	ACE inhibitors versus $\beta$ -blockers	-0.17	0.10	0.84	(0.69,1.03)
	CCB versus $\beta$ -blockers	-0.21	0.07	0.81	(0.71,0.93)
Hierarchical model in Bayesian implementation	ACE inhibitors versus $\beta$ -blockers	-0.18	0.12	0.84	(0.66,1.06)
	CCB versus $\beta$ -blockers	-0.21	0.09	0.81	(0.68,0.97)

The subgroup meta-analysis can also be fitted with the  $2 \times 2$  tables as the starting point rather than the  $\ln(OR)$ , and it is convenient to write this implementation as a hierarchical model. The outcome in each study is the number of patients diagnosed with diabetes and therefore the binomial likelihood can be used in a hierarchical model. This means that the number of events (patients with diabetes) in each study arm,  $r_{iBB}$  and  $r_{iACE}$  for the first three studies comparing ACE inhibitors to  $\beta$ -blockers or  $r_{iBB}$  and  $r_{iCCB}$  for the five studies comparing CCB to  $\beta$ -blockers, follow a specific binomial distribution with a respective probability of success:

$$r_{iBB} \sim B(p_{iBB}, n_{iBB}), \quad i = 1, \dots, 8.$$

$$r_{iACE} \sim B(p_{iACE}, n_{iACE}), \quad i = 1, \dots, 3$$

$$r_{iCCB} \sim B(p_{iCCB}, n_{iCCB}), \quad i = 4, \dots, 8$$

The log-odds ratios can be written as functions of the arm-specific probabilities; the two  $\ln(OR)$  are  $\text{logit}(p_{iACE}) - \text{logit}(p_{iBB})$  and  $\text{logit}(p_{iCCB}) - \text{logit}(p_{iACE})$ . In this case the parameterization of the model is:

$$\text{logit}(p_{iBB}) = u_i$$

$$\text{logit}(p_{iACE}) = u_i + \theta_{iACE-BB},$$

if study  $i$  compares ACE inhibitors versus  $\beta$ -blockers or

$$\text{logit}(p_{iBB}) = u_i$$

$$\text{logit}(p_{iCCB}) = u_i + \theta_{iCCB-BB},$$

if study  $i$  compares CCB versus  $\beta$ -blockers.

Then the study-specific underlying treatment effects  $\theta_{iACE-BB}$  and  $\theta_{iCCB-BB}$  are distributed normally with expectations  $\mu_{ACE-BB}$ ,  $\mu_{CCB-BB}$  and common heterogeneity  $\tau^2$  in the same way as the previous model. Using a half-normal prior distribution for the heterogeneity ( $\tau \sim N(0, 1)$ ,  $\tau > 0$ ) and fitting the model in WinBUGS produces the estimates presented in the lower part of Table 1.

The estimates obtained from the two approaches are very similar. The major difference between the two approaches is in the estimation of heterogeneity. The subgroup meta-analysis fitted within a Bayesian setting with the binomial likelihood gives a posterior median of  $\tau^2$  equal to 0.02 with 95% CrI (0.001, 0.12), slightly larger than the point estimate from the frequentist meta-regression.

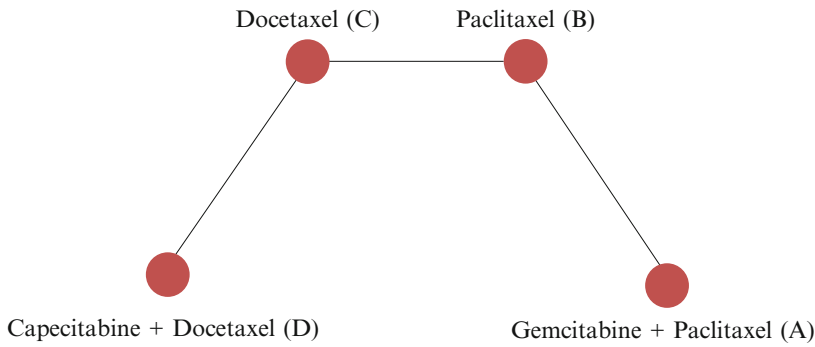
## Indirect and Mixed Comparison

### Theory and Formulae for Indirect Comparisons

In evidence-based medicine, estimates of treatment effect obtained from head-to-head RCTs and combined in a meta-analysis are widely considered the “best available” evidence with which to evaluate the effectiveness of medical interventions (Guyatt et al. 1995; McAlister et al. 1999). Consider two treatments, labelled B and C, which have been compared directly in RCTs and combined in a pairwise meta-analysis. The summary treatment effect estimate is denoted as  $\hat{\mu}_{BC}^D$ , where the superscript denotes the “direct” estimate and the subscript denotes the treatment comparison, where BC is the effect of C relative to B. In the absence of the “level one” evidence for B versus C, it has been suggested that an indirect estimate can be formed via a “common comparator” (Bucher et al. 1997; Song et al. 2003; Glenny et al. 2005), which is assumed to be treatment A. An indirect estimate  $\hat{\mu}_{BC}^I$  can be derived by combining the meta-analytic effect estimates of A versus B studies ( $\hat{\mu}_{AB}^D$ ) and A versus C studies ( $\hat{\mu}_{AC}^D$ ), such that,

$$\hat{\mu}_{BC}^I = \hat{\mu}_{AC}^D - \hat{\mu}_{AB}^D$$

This method is often referred to as an “adjusted indirect comparison,” so-called because randomization is respected by using the relative effect estimates  $\hat{\mu}_{AC}^D, \hat{\mu}_{AB}^D$  obtained from the meta-analyses. Here it is referred to simply as an indirect comparison.



**Fig. 2** Chain of comparisons network of chemotherapy treatments for second-line treatment of breast cancer

**Table 2** Findings from the manufacturer’s submission for gemcitabine STA. Median difference in survival and 95% confidence intervals (Adapted from: Eli Lilly 2006 and Jones et al. 2006)

Treatment comparison	Trials	Median difference (MD) (95% CI)	SE (MD)	Variance (MD)
Gemcitabine + paclitaxel (A) versus paclitaxel (B)	1	−2.8 (−0.01, −5.6)	1.42	2.02
Paclitaxel (B) versus docetaxel(C)	1	2.7 (0.3, 5.1)	1.24	1.54
Docetaxel (C) versus Capecitabine + docetaxel (D)	1	3.0 (0.6, 5.4)	1.20	1.44

The usual measures of statistical variability can be derived for the indirect estimate. As  $\hat{\mu}_{BC}^I$  is formed as a difference between two independent estimates its variance,  $\hat{v}_{BC}^I$ , is equal to the sum of the variances,  $\hat{v}_{AC}^D$  and  $\hat{v}_{AB}^D$ , estimated from the direct AC and AB comparisons:

$$\hat{v}_{BC}^I = \hat{v}_{AC}^D + \hat{v}_{AB}^D$$

A single head-to-head randomized trial is as precise as an indirect comparison based on four trials of the same size. To see this, suppose each trial produces an estimate with variance  $\sigma^2$ . A meta-analysis of  $s$  trials with direct estimates of  $A$  versus  $B$  will have variance  $\hat{v}_{AB}^D = \sigma^2/s$  (based on inverse variance weights). The indirect estimate of  $B$  versus  $C$  via  $A$  based on  $s$  AB and  $s$  AC trials will have variance  $\hat{v}_{BC}^I = \hat{v}_{AC}^D + \hat{v}_{AB}^D = \sigma^2/s + \sigma^2/s = 2\sigma^2/s$ .

A common misconception is that for an indirect comparison to be valid, every trial must include a common comparator (Hughes 2010). In truth, indirect estimates can be derived via many routes. The only requirement is that

the network is “connected” and not necessarily via a common comparator. Consider the network shown in Fig. 2 which is adapted from a 2006 submission to NICE which included the four distinct regimens for the second-line treatment of metastatic breast cancer (Eli Lilly 2006).

Table 2 reports the results for difference in median years survival. Note there are direct estimates available for gemcitabine + paclitaxel versus paclitaxel ( $\hat{\mu}_{AB}^D = -2.8$  years), paclitaxel versus docetaxel ( $\hat{\mu}_{BC}^D = 2.7$  years and docetaxel versus capecitabine + docetaxel ( $\hat{\mu}_{CD}^D = 3.0$  years), which forms a “chain” of evidence A-B-C-D. The comparison of interest to the decision-maker was gemcitabine + paclitaxel (A) versus capecitabine + docetaxel (D) (Jones et al. 2006), an indirect comparison of which can be formed as

$$\hat{\mu}_{AD}^I = \hat{\mu}_{AB}^D + \hat{\mu}_{BC}^D + \hat{\mu}_{CD}^D = 2.9 \text{ years}$$

$$SE(\hat{\mu}_{AD}^I) = \sqrt{\hat{v}_{AB} + \hat{v}_{BC} + \hat{v}_{CD}} = 2.23$$

### Theory and Formulae for Mixed Comparisons

If both direct and indirect estimates are available for the same comparison, they can be combined by taking the weighted average of  $\hat{\mu}_{BC}^D$  and  $\hat{\mu}_{BC}^I$ . This has been referred to as a *mixed* comparison and will be denoted here as  $\hat{\mu}_{BC}^M$ . However, it should not be confused with a “mixed-treatment comparison” (Lu 2004), which refers to the simultaneous comparison of multiple treatments in a single analysis and is synonymous to network meta-analysis. A simple and intuitive approach for combining direct and indirect evidence is the inverse variance method, where

$$\hat{\mu}_{BC}^M = \frac{\frac{1}{\hat{v}_{BC}^D} \hat{\mu}_{AB}^D + \frac{1}{\hat{v}_{BC}^I} \hat{\mu}_{BC}^I}{\frac{1}{\hat{v}_{BC}^D} + \frac{1}{\hat{v}_{BC}^I}}$$

with variance

$$\hat{v}_{BC}^M = \frac{1}{\frac{1}{\hat{v}_{BC}^D} + \frac{1}{\hat{v}_{BC}^I}}$$

A 95% confidence interval for the mixed estimate can be obtained as  $\hat{\mu}_{BC}^M \pm 1.96\sqrt{\hat{v}_{BC}^M}$ . Note that in the case of a dichotomous outcome, where  $\mu$  is the  $\ln(OR)$  or  $\ln(RR)$ , mean effect size and confidence intervals for the  $OR$  and  $RR$  can be obtained by exponentiation.

#### Example: Indirect and Mixed Comparison for ACE inhibitors Versus CCB

Using the  $\ln(OR)$  for the comparisons of ACE inhibitors and CCB each versus  $\beta$ -blockers presented in Table 1, an indirect estimate for ACE inhibitors versus CCB can be obtained. The indirect  $\ln(OR)$  estimate  $\hat{\mu}_{ACE-CCB}^I$  is calculated as the difference between the direct  $\ln(OR)$  for CCB versus  $\beta$ -blockers and direct  $\ln(OR)$  for ACE inhibitors versus  $\beta$ -blockers. Using the estimates from the frequentist subgroup analysis described earlier, the indirect estimate is

$$\begin{aligned} \hat{\mu}_{ACE-CCB}^I &= \hat{\mu}_{ACE-BB}^D - \hat{\mu}_{CCB-BB}^D \\ &= -0.17 - (-0.21) = 0.04 \end{aligned}$$

The variance of the indirect estimate  $\hat{\mu}_{ACE-CCB}^I$  is the sum of the variances of  $\hat{\mu}_{ACE-BB}^D$  and  $\hat{\mu}_{CCB-BB}^D$ :

$$\begin{aligned} \hat{v}_{ACE-CCB}^I &= \hat{v}_{ACE-BB}^D + \hat{v}_{CCB-BB}^D \\ &= 0.10^2 + 0.07^2 = 0.0149 \end{aligned}$$

Therefore, the indirect OR of ACE inhibitors versus CCB is  $\exp(\hat{\mu}_{ACE-CCB}^I) = 1.04$  with 95% CI  $\exp(\hat{\mu}_{ACE-CCB}^I \pm 1.96\sqrt{\hat{v}_{ACE-CCB}^I}) = (0.82, 1.32)$ .

Since there are also three studies that directly compare ACE inhibitors with CCB, they can be combined with the indirect estimate to produce a mixed estimate. Synthesis of the studies provides a direct estimate for the  $\hat{\mu}_{ACE-CCB}^D$  equal to  $-0.22$  with standard error 0.11. Then, the mixed estimate can be obtained as the weighted average of the direct and indirect  $\ln(OR)$ :

$$\begin{aligned} \hat{\mu}_{ACE-CCB}^M &= \frac{\frac{1}{\hat{v}_{ACE-CCB}^D} \hat{\mu}_{ACE-CCB}^D + \frac{1}{\hat{v}_{ACE-CCB}^I} \hat{\mu}_{ACE-CCB}^I}{\frac{1}{\hat{v}_{ACE-CCB}^D} + \frac{1}{\hat{v}_{ACE-CCB}^I}} \\ &= \frac{\frac{1}{0.0121}(-0.22) + \frac{1}{0.0149}0.04}{\frac{1}{0.0121} + \frac{1}{0.0149}} = -0.10 \end{aligned}$$

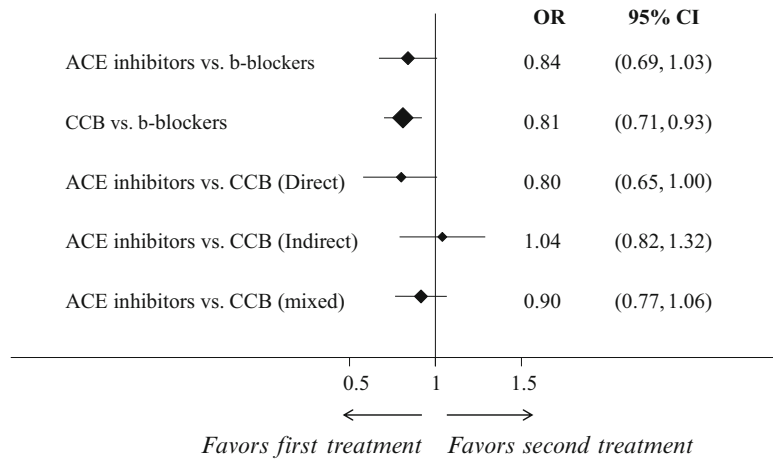
The variance of this estimate is

$$\begin{aligned} \hat{v}_{ACE-CCB}^M &= \frac{1}{\frac{1}{0.0121} + \frac{1}{0.0149}} \\ &= \frac{1}{\frac{1}{0.0225} + \frac{1}{0.0149}} = 0.009 \end{aligned}$$

The mixed OR is  $\exp(\hat{\mu}_{ACE-CCB}^M) = 0.90$  with 95% CI (0.77, 1.06).

The indirect and mixed ORs for ACE inhibitors versus CCB are presented in Fig. 3. Note that

**Fig. 3** Summary odds ratios (OR) for each comparison using direct, indirect, and mixed evidences. Diamonds represent the point estimates and the horizontal lines the corresponding 95% confidence intervals



the confidence interval for the mixed estimate is narrower than the confidence interval for the direct estimate. By combining direct evidence with indirect evidence, the variance is reduced by 45%.

This approach is intuitive and will be familiar to meta-analysts; however, it is labor-intensive. In a three treatment network, A-B-C, the meta-analytic estimates  $\hat{\mu}_{AB}^D, \hat{\mu}_{AC}^D, \hat{\mu}_{BC}^D$  must be obtained, the indirect estimates  $\hat{\mu}_{AB}^I, \hat{\mu}_{AC}^I, \hat{\mu}_{BC}^I$  derived, and then the mixed estimates  $\hat{\mu}_{AB}^M, \hat{\mu}_{AC}^M, \hat{\mu}_{BC}^M$  computed. As the number of treatments increases and the network expands, this approach quickly becomes untenable and more sophisticated approaches can be used (Caldwell et al. 2005). Section “Models for Network Meta-analysis” discusses methods for simultaneously combining direct and indirect evidence in a single analysis. The next section discusses the underlying assumptions needed to undertake an indirect or mixed comparison.

### Assumptions Underlying Indirect and Mixed Comparisons

Current hierarchies of evidence place indirect and “mixed” comparisons below direct evidence regardless of whether the constituent effect estimates have been obtained from meta-analyses of RCTs (currently “level one” evidence). Several

HTA organizations have expressed doubts about indirect comparisons and state that if direct evidence exists, it should take precedence. For example, the *Cochrane Handbook* (Higgins and Green 2008) states “indirect comparisons may suffer the biases of observational studies” and advises that direct and indirect evidence only be combined as a supplemental analysis. In England and Wales, NICE (2008) uses direct evidence as the reference case for appraisals of new technologies, only allowing indirect and “mixed” comparisons as a supporting analysis. Similarly, CADTH (Wells et al. 2009) adopts a cautious stance and the Pharmaceutical Benefits Advisory Committee (PBAC) in Australia have expressed skepticism about the use of indirect evidence (PBAC 2008).

This caution is based on concerns regarding the key assumption underpinning indirect comparisons, which is reflected mathematically in the consistency equation  $\mu_{BC} = \mu_{AC} - \mu_{AB}$ . The consistency equation relates to the true (or average) effectiveness of B versus C rather than to each individual study. It states that the effect of B versus C can be estimated either indirectly via A (right part of the equation) or directly (left part of the equation) and that these two pieces of evidence will, on average, give the same result. Rearranging the parts of the equation shows that one consistency equation is sufficient to reflect consistency for all three comparisons in a three treatment network. Such that

$$\mu_{BC} = \mu_{AC} - \mu_{AB}, \mu_{AB} = \mu_{AC} - \mu_{BC},$$

$$\mu_{BC} = \mu_{AC} - \mu_{AB}$$

The validity of the consistency equation is fundamental to the validity of indirect comparisons. In considering the validity of the assumption for the combination of direct and indirect evidence, some authors have found it instructive to separate the notion of *similarity* (Song et al. 2009; Donegan et al. 2010) or *transitivity* (Baker and Kramer 2002) from the notion of *consistency*. In the current chapter, these notions are interpreted as the distinction between clinical or epidemiological considerations on the one hand, and statistical considerations on the other. Transitivity refers to the genuine ability to learn about a pairwise comparison via an intermediate treatment via indirect comparison. As will be discussed below, it requires the intermediate treatment to be equivalent when compared against each of the treatments of interest and that the actual studies contributing to the indirect comparison do not differ in important ways. Specifically, when  $\mu_{AB}^I$  is calculated, it is assumed that we can learn about B versus C via A. The common comparator A might be said to be “transitive” when it allows valid comparison of the treatments to which it is linked. Note that transitivity is not a property of the common treatment A alone but of the two sets of studies it links.

Consistency is a statistical notion that can be considered at the level of the parameters or the level of the data. The consistency equation defines relationships among the parameters. The validity of the assumption embodied in the equation can be assessed only when data from different sources form a “closed loop” of evidence in the network (a path that starts and ends at the same node treatment). When the consistency assumption does not hold or when there is evidence of disagreement between direct and indirect evidence, then the evidence is said to be *inconsistent* (or show *inconsistency*).

When transitivity holds and there are multiple sources of evidence, the consistency equation should hold. The consistency equation may hold in a statistical sense, however, even when the

studies do not allow valid indirect comparisons, due to important differences between studies that prevent an assumption of transitivity from holding. If there is inconsistency in the data, the assumption of transitivity is clearly challenged. For an “open loop,” that is one for which there is indirect evidence but no direct evidence, consistency cannot be evaluated statistically, and the validity of the indirect comparison must rest entirely on clinical and epidemiological judgments regarding the plausibility of transitivity. It can be shown mathematically (Lu and Ades 2009) that consistency is a consequence of the assumption of exchangeability that forms the basis of the Bayesian network meta-analysis models which is, in turn, an extension of the usual assumption made in a pairwise meta-analysis (Dias et al. 2010). The assumption of transitivity is essentially equivalent to the assumption of exchangeability in this sense, since it relates to similarity of studies. The term “transitivity” might be preferred to “similarity” (Donegan et al. 2010); however, because (i) it better describes the aim of the assumption to compare two treatments via a third one; (ii) it clearly refers to more than two comparisons whereas the term “similarity” reduces to homogeneity when we refer to a single head-to-head comparison; and (iii) “similarity” may be misinterpreted as necessitating *all* trial and patient characteristics to be similar, when in truth a valid indirect comparison can be obtained even when studies are dissimilar, so long as such characteristics do not modify treatment effect.

### Requirements for Transitivity

Transitivity requires some particular characteristics of the studies contributing to the indirect comparison, as follows:

- The two sets of trials AB and AC do not differ with respect to the distribution of effect modifiers.

In order for an indirect comparison to be valid, the distribution of treatment effect modifiers should be similar in AB and AC trials. Before

conducting an indirect or mixed comparison, the analyst should therefore ensure they have identified a priori possible effect modifiers and should compare their distributions across treatment comparisons. For example, in a network of treatments for childhood nocturnal enuresis, Caldwell et al. (2010) hypothesize that the age of the children could be a potential effect modifier since a 6-year-old suffering from nighttime bedwetting might have a different underlying pathology from a 12-year-old.

Note that the consistency assumption holds at the level of the mean effect sizes and as such an effect modifier that differs within studies of one comparison but has a similar distribution across both comparisons will not violate the assumption. For example, if age is an effect modifier and AC trials differ in terms of mean age of participants (which will manifest as heterogeneity in AC studies), but the same variability is observed in the set of AB trials then transitivity could still hold. In contrast, if the distribution of age differs across comparisons such that children in the AC studies tended to be younger and those in AB studies tended to be much older, then the assumption would not hold.

Adjustment using regression techniques can be used to account for small differences in the distribution of effect but note that covariates must be carefully selected. Only effect modifiers and not “colliders” (variables that influence the choice of comparison and the effectiveness) should be considered for adjustment. Adjustment for colliders, as in classical epidemiology, will introduce bias rather than improve the plausibility of the transitivity assumption (Jansen et al. 2012).

- The interventions are being given for the same indications.

Transitivity could be violated if interventions have different indications, so it is important that sets of studies contributing to the indirect comparison are using the treatments for the same underlying condition. A particularly useful way to think about this requirement is to consider whether the participants included in the network could, in principle, be randomized to any of the three treatments A, B, and C. For example, if treatment

A is a chemotherapy regimen typically administered as a second-line treatment, whereas treatments B and C can be used either as first or second line, it cannot be assumed that participants in a BC trial could have been randomized in an AC trial. Although this consideration is a fundamental one and should be addressed when building the evidence network, it might be the case that treatments are comparable in theory but not in practice. For example, interferon, glatiramer acetate, or natalizumab are commonly used in clinical practice for patients with relapsing-remitting multiple sclerosis whereas mitoxantrone, methotrexate, cyclophosphamide, or azathioprine are more frequent for patients with a progressive disease. Evidence to support this clinical “tradition” is not solid, and it would be appealing to compare all these treatments. In practice, however, transitivity will be violated as comparisons will differ in disease severity.

Another way to conceptualize this requirement is to consider treatments not included in each study as missing data (Lu and Ades 2006). Thus, AB trials are missing C arms, and AC trials are missing B arms. The transitivity assumption is likely to hold if these arms are missing in an entirely random way, which guarantees that the choice of treatments is unrelated to the indications for which they are given. In practice, the selection of treatments to be included in a trial is not random. In many clinical trials, the choice of comparator is placebo or an older, suboptimal intervention rather than a realistic alternative such as an established effective treatment. If the choice of comparator is associated, directly or indirectly, with the relative effectiveness of the interventions then the key assumption will be violated.

- The treatment C is similar when it appears in AC and BC trials.

The transitivity assumption is violated when the treatments in question differ *systematically* between trials. The definition of the nodes in the treatment network is a challenging issue as very often treatments are given at various doses, administrations routes, frequencies, etc. For



example, consider that the common comparator A is a treatment which can be given at different doses, but there is no systematic difference on the average dose of A between AC and AB trials. In this case the assumption can hold although there could be heterogeneity within AC and AB comparisons. Consequently, the “anchor” treatment A can be represented by a single node allowing the indirect comparison of B and C. If, however, A is given via a different administration route in all AC and AB trials, then it is questionable whether the two types of A can form a common node and an indirect comparison of B versus C via A would be impossible. For example, when comparing different fluoride treatments, comparison between fluoride toothpaste and fluoride rinse can be made via placebo. However, placebo toothpaste and placebo rinse might not be comparable as the mechanical function of brushing might have a different effect on the prevention of caries. If this is the case, the transitivity assumption is doubtful (Salanti et al. 2009).

**Estimating Inconsistency in Mixed Comparisons**

In theory, the consistency equation  $\mu_{BC} = \mu_{AC} - \mu_{AB}$  must hold if transitivity is deemed to hold. However, in practice, there may be inconsistency in the evidence base. In a three-treatment network, three independent direct estimates,  $\hat{\mu}_{AB}^D, \hat{\mu}_{AC}^D$  and  $\hat{\mu}_{BC}^D$  (assuming there are no trials with more than two arms), and three indirect estimates,  $\hat{\mu}_{AB}^I, \hat{\mu}_{AC}^I$  and  $\hat{\mu}_{BC}^I$ , can be obtained. Assuming the treatment comparison of interest is B versus C, the discrepancy (difference) between the direct and indirect estimates forms the measure of inconsistency. This discrepancy is called the inconsistency factor (IF) which is estimated as

$$\hat{IF}_{ABC} = |\hat{\mu}_{BC}^D - \hat{\mu}_{BC}^I|$$

Note that the direction of the difference might be clinically important but mathematically is unimportant for the statistical evaluation of consistency. Consequently only absolute differences are taken. In a three-treatment network, only one measure of inconsistency is possible (and hence

the subscript denoting the loop) (Lu and Ades 2006) as it can be shown that the same inconsistency factor will be obtained whichever edge of the triangle is of interest.

The variance of the inconsistency factor is

$$\text{var}(\hat{IF}_{ABC}) = \hat{v}_{BC}^D + \hat{v}_{BC}^I$$

A 95% confidence interval can be obtained for the inconsistency factor as  $\hat{IF}_{ABC} \pm 1.96 \sqrt{\text{var}(\hat{IF}_{ABC})}$ . The null hypothesis of evidence consistency  $\hat{IF} = 0$  can then be tested by deriving a z-test (Bucher et al. 1997).

$$z = \frac{\hat{IF}_{ABC}}{\sqrt{\text{var}(\hat{IF}_{ABC})}}$$

If consistency holds, it is reasonable to combine across  $\hat{\mu}_{BC}^D$  and  $\hat{\mu}_{BC}^I$  to form  $\hat{\mu}_{BC}^M$ . However, if there is evidence of a “statistically significant” discrepancy ( $p \leq 0.05$ ), the fundamental assumption is not fulfilled, and one may say that there is evidence of inconsistency.

Claims have been made that indirect comparisons may systematically over- (Bucher et al. 1997; Mills et al. 2011) or underestimate treatment effects compared with direct comparisons. Since inconsistency is a property of a “loop” of evidence apparent overestimation of a treatment effect on one side of a triangle network (e.g.,  $\hat{\mu}_{BC}^I$ ) corresponds to underestimation of another (e.g.,  $\hat{\mu}_{AC}^I$ ). Thus, any assessment of consistency needs to take account of the particular circumstances of the problem. Until recently, empirical investigation of the extent of inconsistency has been limited. In a recent review, (Song et al. 2011) examined 112 independent three-treatment networks and detected 16 cases of statistically significant discrepancies between direct and indirect estimates. However, there was no consistent direction as to over- or underestimation. Of course, the test for inconsistency may have low power to detect true inconsistency should it exist, as with other interaction effects. The analyst must therefore be extremely cautious in their interpretation even if inconsistency is not detected.

Note that the discovery of inconsistency does not necessarily mean that all indirect comparisons in the loop are invalid. For example, suppose that AC and BC trials are similar regarding the distribution of effect modifiers (e.g., all studies are carried out in adults with a similar distribution in age), so that  $\hat{\mu}_{AB}^I$  is a valid estimate of the relative effectiveness of A versus B for the given setting and population. If now the AB studies have all being carried out in younger populations (e.g., in adolescents) then the consistency assumption does not hold; both  $\hat{\mu}_{AB}^I$  and  $\hat{\mu}_{AB}^D$  are valid but answer different questions; hence computation of a mixed estimate,  $\hat{\mu}_{AB}^M$ , would be inappropriate.

### Example: Inconsistency in the Evidence Triangle ACE Inhibitors Versus CCB Versus $\beta$ -Blockers

Inconsistency can be evaluated by calculating the difference between direct and indirect estimate for the same comparison. In the case of ACE inhibitors versus CCB, the inconsistency factor reflects the disagreement in the triangle formed by the three sets of trials ACE inhibitors versus CCB versus  $\beta$ -blockers and is calculated as

$$\begin{aligned} \hat{IF}_{ACE-CCB-BB} &= |\hat{\mu}_{ACE-CCB}^D - \hat{\mu}_{ACE-CCB}^I| = |-0.22 - 0.04| = 0.26 \end{aligned}$$

The standard error of the inconsistency factor is obtained as

$$\begin{aligned} SE(\hat{IF}_{ACE-CCB-BB}) &= \sqrt{\hat{v}_{ACE-CCB}^D + \hat{v}_{ACE-CCB}^I} \\ &= \sqrt{0.0144 + 0.0225} \\ &= 0.192 \end{aligned}$$

A 95% CI for the inconsistency factor is obtained as  $\hat{IF} \pm 1.96 \times SE(\hat{IF}) = (-0.12, 0.64)$ . The z-test for the hypothesis  $H_0 : \hat{IF}_{ACE-CCB-BB} = 0$  is

$$z = \frac{\hat{IF}_{ACE-CCB-BB}}{SE(\hat{IF}_{ACE-CCB-BB})} = \frac{0.26}{0.192} = 1.35$$

leading to a p-value equal to 0.91. Note that this result applies to the entire triangle: the same inconsistency factor and p-value could have been obtained by calculating the difference between direct and indirect evidence for the ACE versus  $\beta$ -blockers or CCB versus  $\beta$ -blockers comparisons. As the 95% CI includes zero, there is no indication of important statistical inconsistency between direct and indirect estimate, which is also supported by the p-value.

## Models for Network Meta-analysis

Extensions of the ideas above to more than three treatments lead to a general framework for network meta-analysis. Consider a set of  $T$  treatments of interest that we want to evaluate according to their relative effectiveness on a single outcome measure. The treatments are studied collectively in  $N$  studies. Each study may provide evidence about some of the treatments; it will include only a subset of  $T$ ,  $T_i \in T$ . The study data can be arm-based or contrast-based. In the contrast-based approach, the effect sizes  $y_{ijk}$  from each study are available, and they refer to the relative effectiveness of a treatment  $k$  relative to  $j$  with  $j, k \in T_i$ . Network meta-analysis can be viewed as a special case of meta-regression (linear model), as a hierarchical model or as a multivariate meta-analysis model. The estimation methods that arise from these approaches are essentially equivalent and can be employed under the assumption of consistency or under assumptions that impose fewer restrictions.

A key issue in all methods for fitting network meta-analysis is the minimization of the parameters' space by selecting a minimum set of *basic parameters*. This is a set of comparisons (as many as the total number of treatments minus one) that are sufficient to generate all possible comparisons between the treatments via the consistency equations. Under consistency, the choice of the basic parameters does not affect the results but typically the basic parameters are defined by taking the comparisons of all treatments versus a common reference to simplify interpretation. Examples to follow should make this clear.

## Consistency Models

### Network Meta-analysis as a Linear Model

Consider the simplest case of having three treatments of interest  $T = \{A, B, C\}$  and studies that compare all possible pairs of those treatments; i.e., there are  $AB, AC,$  and  $BC$  studies. For now it is assumed that only two-arm trials are available. In general,  $y_{ijk}$  refers to the relative effectiveness of two interventions  $j$  and  $k$  within study  $i$ . When each study has only two treatments, the treatment indices can be dropped and the observed effect written as  $y_i$ . The treatment indices will be reemployed in section “Network Meta-analysis as a Hierarchical Model.”

The two-step process described in sections “Theory and Formulae for Indirect Comparisons” and “Theory and Formulae for Mixed Comparisons” is a simple network meta-analysis. For a given comparison, say  $B$  versus  $C$ , an indirect estimate  $\hat{\mu}_{BC}^I$  is derived by combining  $AC$  and  $AB$  studies. Then, the indirect estimate and direct estimate  $\hat{\mu}_{BC}^D$  are synthesized to obtain the mixed summary estimate  $\hat{\mu}_{BC}^M$ . This first step of this process was described in the context of a meta-regression in section “Meta-analysis and Meta-regression as Linear Model.” This can be done by creating two dummy variables to identify the  $AC$  and  $BC$  studies and omitting the intercept:

$$y_i = \mu_{AC}x_{iAC} + \mu_{AB}x_{iAB} + \delta_i + e_i$$

The same model can be used for both stages of the mixed comparison analysis, by careful specification of the covariate values in a way that forces the consistency equation into the analysis as a constraint. As above, if study  $i$  compares  $A$  and  $C$ , then  $x_{iAC} = 1$ ,  $x_{iAB} = 0$ , and if study  $i$  compares  $A$  and  $B$  then  $x_{iAC} = 0$ ,  $x_{iAB} = 1$ . Now if study  $i$  compares  $B$  and  $C$ , the consistency equation  $\mu_{BC} = \mu_{AC} - \mu_{AB}$  can be introduced by setting  $x_{iAC} = 1$ ,  $x_{iAB} = -1$ . Note that because of the assumption of consistency, only two comparisons need to be included in the model (here  $AC$  and  $AB$ ), and consequently there are two explanatory variables to be included in the meta-regression model. The two

contrasts  $AC$  and  $AB$  are called the *basic contrasts* and the parameters  $\mu_{AC}, \mu_{AB}$  the *basic parameters*, whereas the  $\mu_{BC}$  is a functional parameter and can be derived as a linear function of the two basic parameters. The choice of the two out of the three contrasts that enter the meta-regression is arbitrary and does not impact on the parameters estimation; e.g.,  $x_{iAC}$  and  $x_{iBC}$  could have been chosen as covariates.

Extending the idea to the case of more than three treatments results in a full network meta-analysis. For example, with  $T = \{A, B, C, D, E\}$  treatments included, there are  $T(T - 1)/2 = 10$  possible head-to-head comparisons.  $T - 1$  basic parameters are selected, such as all treatment comparisons  $A_j$  of treatment  $j = B, C, D, E$  versus treatment  $A$ , relating to regression coefficients  $\mu_{Aj}$ . The meta-regression model would be

$$y_i = \sum_{j=B,C,D,E} \mu_{Aj}x_{iAj} + \delta_i + e_i \quad (6)$$

with  $e_i \sim N(0, s_i^2)$ . The variable  $x_{iAj} = 1$  if study  $i$  compares  $A$  and  $j$ ,  $x_{iAj} = 0$  if study  $i$  compares  $A$  and  $k$ ,  $k \neq j$ . If a study compares treatment  $s$ ,  $j$  and  $k$  and does not include  $A$ , then the consistency equations are used to derive the values of  $x_{iAj}$ . If, for example, a study compares  $B$  and  $D$ , then  $x_{iAj} = 0, j = C, E, x_{iAB} = -1$  and  $x_{iAD} = 1$  because  $\mu_{BD} = \mu_{AD} - \mu_{AB}$ .

In summary, all observed comparisons are reexpressed using the regression covariates  $x_{iAj}$ . This gives the model  $N_{\text{comp}} - (T - 1)$  degrees of freedom (number of functional parameters), where  $N_{\text{comp}}$  is the number of comparisons observed in the network. For example, if the network consists of  $AC, AB, BC, CD, BE, BD$  studies, then there are  $6 - (5 - 1) = 2$  degrees of freedom. This can be also visualised by the number of independent closed loops in the network diagram ( $ABC$  and  $BCD$ ).

This model can be fitted using any meta-regression software (such as the `metareg` command in STATA). The estimated regression coefficients  $\hat{\mu}_{Aj}$  are network meta-analysis estimates for all treatments versus the reference treatment  $A$  and their uncertainty is conveyed by  $SE(\hat{\mu}_{Aj})$ . Network

meta-analysis summary effects for all other comparisons, say  $B$  versus  $D$ , can be obtained by considering the consistency equations relating the  $\hat{\mu}_{Aj}$  to the functional parameters. Their variances can be obtained by combining standard errors and covariances (from the variance-covariance matrix for the estimated regression coefficients). For instance,  $\hat{\mu}_{BD} = \hat{\mu}_{AD} - \hat{\mu}_{AB}$  and  $SE^2(\hat{\mu}_{BD}) = SE^2(\hat{\mu}_{AD}) + SE^2(\hat{\mu}_{AB}) - 2\text{COV}(\hat{\mu}_{AD}, \hat{\mu}_{AB})$ .

Note that the random effects follow a normal distribution  $\delta_i \sim N(0, \tau^2)$ , with heterogeneity variance assumed to be equal for every comparison. This may be a strong assumption as different comparisons might include studies with different between-study variability. Assuming a common heterogeneity might impose an inappropriate  $\tau^2$  value for some comparisons. Although assuming comparison-specific heterogeneities can be desirable in many cases, it presents practical difficulties. Estimation of the parameter  $\tau^2$  can be challenging if few studies are available. Even with large network meta-analyses including many treatments, it is often the case that some of the comparisons include only a few studies; some comparisons might even be informed by a single study. Nevertheless, assuming a common heterogeneity parameter allows comparisons to “borrow strength” from each other in the estimation of the common  $\tau^2$ , overcoming computational problems that are encountered both with frequentist and Bayesian fitting of models.

**Application: Network Meta-analysis Using Meta-regression for Incident Diabetes**

Standard meta-regression methods can be only be applied to networks that contain two-arm studies. The following analysis treats the 30 pairwise comparisons in the incident diabetes data set as if they came from 30 (rather than the true 22) independent studies. A meta-regression model is employed where again the different comparisons define the subgroups. First the  $T - 1$  “basic contrasts” need to be selected, to be included as covariates in the model. Several combinations of basic contrasts are possible, and for  $T = 6$ , five parameters need to be selected. For ease of interpretation, it is convenient to choose the comparisons of each treatment versus

a common reference treatment. Here, placebo (P) is chosen to be the reference treatment and basic contrasts are defined for each treatment versus placebo. Then, to specify the design matrix all comparisons in the network need to be written as functions of the basic parameters. The first two columns of Table 3 list all comparisons in the network for which direct estimates are available and the number of studies involving each comparison. Then, for the five comparisons belonging to the basic contrasts (e.g.,  $\beta$ -blockers (BB) vs. P), the respective variable  $x_i$  ( $x_{iBB-P}$ ) takes the value 1 and the variables of the other four basic contrasts take the value 0. For any other treatment comparison (e.g., diuretics (D) vs. BB)  $x_i$  takes value -1 for the first treatment ( $x_{iD-P}$ ) and -1 for the second treatment based on the consistency equations ( $\mu_{D-BB} = \mu_{D-P} - \mu_{BB-P}$ ).

The full meta-regression model is

$$y_i = \mu_{BB-P}x_{iBB-P} + \mu_{D-P}x_{iD-P} + \mu_{CCB-P}x_{iCCB-P} + \mu_{ACE-P}x_{iACE-P} + \mu_{ARB-P}x_{iARB-P} + \delta_i + e_i.$$

Fitting the model in STATA using **metareg** produces the regression coefficients in Table 4.

The common heterogeneity parameter of the network was estimated as 0.02. The variance-covariance matrix of the regression-coefficients is saved by STATA as the “e(v)” matrix and can be obtained after fitting the meta-regression model (Table 5).

Then any head-to-head comparison can be derived applying again the consistency equations to the point estimates. For example, the  $\ln(OR)$  of diuretics versus  $\beta$ -blockers is  $\hat{\mu}_{D-BB} = \hat{\mu}_{D-P} - \hat{\mu}_{BB-P} = 0.32 - 0.24 = 0.08$ , and its standard error is

$$\begin{aligned} SE(\hat{\mu}_{D-BB}) &= \sqrt{\hat{v}_{D-P} + \hat{v}_{BB-P} - 2\text{Cov}(\hat{\mu}_{D-P}, \hat{\mu}_{BB-P})} \\ &= \sqrt{0.008 + 0.0076 - 2 \times 0.004} = 0.09 \end{aligned}$$

All other functional contrasts estimates are derived the same way. The network meta-analysis estimates for all comparisons are presented in the black diamonds in Fig. 4.

**Table 3** Parameterization of design matrix for the five basic contrasts when placebo is the reference treatment for incident diabetes

Comparison in study $i$	Number of studies	$x_{iBB-P}$	$x_{iD-P}$	$x_{iCCB-P}$	$x_{iACE-P}$	$x_{iARB-P}$
		$\beta$ -blockers versus placebo	diuretics versus placebo	CCB versus placebo	ACE inhibitors versus placebo	ARB versus placebo
$\beta$ -blockers versus placebo	1	1	0	0	0	0
diuretics versus placebo	3	0	1	0	0	0
diuretics versus $\beta$ -blockers	2	-1	1	0	0	0
CCB versus placebo	1	0	0	1	0	0
CCB versus $\beta$ -blockers	5	-1	0	1	0	0
CCB versus diuretics	2	0	-1	1	0	0
ACE inhibitors versus placebo	2	0	0	0	1	0
ACE inhibitors versus $\beta$ -blockers	3	-1	0	0	1	0
ACE inhibitors versus CCB	2	0	0	-1	1	0
ARB versus placebo	3	0	0	0	0	1
ARB versus $\beta$ -blockers	3	-1	0	0	0	1
ARB versus diuretics	1	0	-1	0	0	1
ARB versus CCB	1	0	0	-1	0	1

**Table 4** Results of network meta-analysis as meta-regression for incident diabetes. Log-odds ratios ( $\hat{\mu}$ ) with their standard error  $SE(\hat{\mu})$  and odds ratios (OR) with their 95% confidence interval (CI) for all basic and functional contrasts are reported

Comparison	$\hat{\mu}$	$SE(\hat{\mu})$	OR	95% CI for OR
$\beta$ -blockers versus placebo	0.24	0.09	1.27	(1.07, 1.52)
diuretics versus placebo	0.32	0.09	1.38	(1.15, 1.64)
CCB versus placebo	0.08	0.08	1.08	(0.93, 1.27)
ACE inhibitors versus placebo	-0.11	0.08	0.90	(0.77, 1.05)
ARB versus placebo	-0.17	0.10	0.84	(0.69, 1.03)

Note that the confidence intervals for the comparison ACE inhibitors versus CCB have further reduced compared with the direct or mixed estimate previously calculated. Also, an estimate for the comparison ARB versus ACE inhibitors is obtained for which no studies exist. Figure 4 also shows the ranking of the treatments by

ranking the mean OR of each treatment versus placebo.

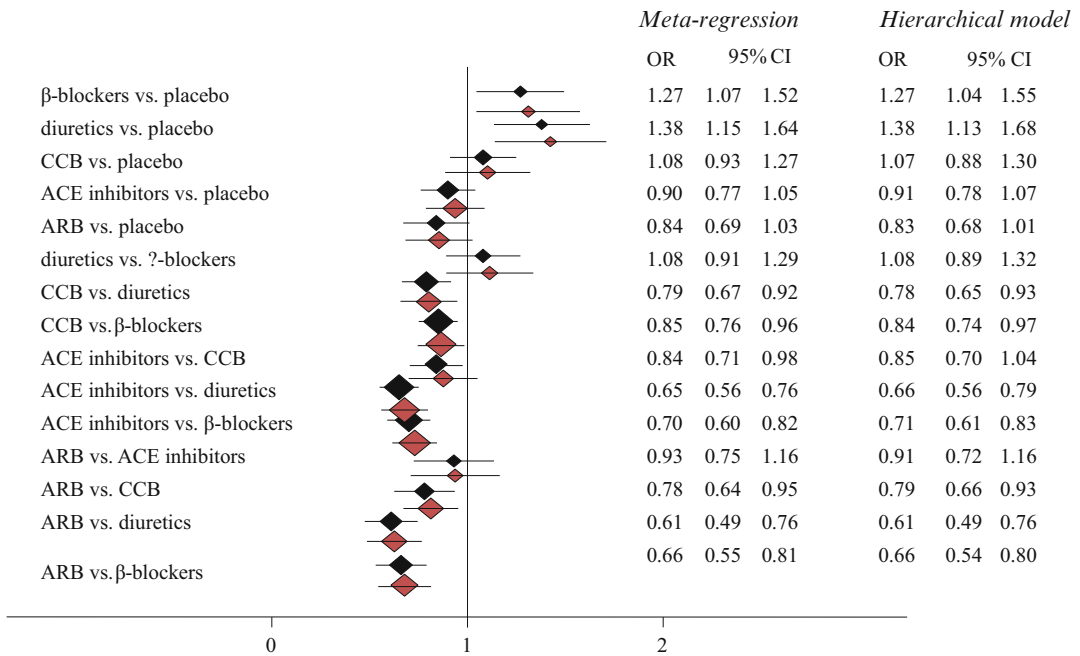
**Network Meta-analysis as a Hierarchical Model**

An alternative way to fit the network meta-analysis model is by extending the hierarchical

**Table 5** Variance-covariance matrix of the five basic parameters used in meta-regression approach for network meta-analysis for incident diabetes. The diagonal includes

	$\hat{\mu}_{BB-P}$	$\hat{\mu}_{D-P}$	$\hat{\mu}_{CCB-P}$	$\hat{\mu}_{ACE-P}$	$\hat{\mu}_{ARB-P}$
$\hat{\mu}_{BB-P}$	0.0076				
$\hat{\mu}_{D-P}$	0.0040	0.0080			
$\hat{\mu}_{CCB-P}$	0.0052	0.0040	0.0070		
$\hat{\mu}_{ACE-P}$	0.0038	0.0034	0.0035	0.0058	
$\hat{\mu}_{ARB-P}$	0.0037	0.0024	0.0037	0.0022	0.0098

the variances of the parameters and the other cells the covariances between the two corresponding parameters



**Fig. 4** Results from network meta-analysis conducted as meta-regression in STATA (black) ignoring correlations from multi-arm trials and as hierarchical model in WinBUGS that account for correlations (red). Diamonds

are the point estimates of summary odds ratios (OR) and the horizontal lines represent the corresponding 95% confidence intervals (CI)

model. For the simplest case of three treatments  $A, B, C$ , assume there are studies that inform all possible comparisons. The effect size for a study that compares  $A$  versus  $B$  is denoted by  $y_{iAB}$ . When only two-arm trials are included in the network, the likelihood for the observations is specific to the comparison being presented, i.e.,

$$y_{iAC} \sim N(\theta_{iAC}, s_{iAC}^2), y_{iAB} \sim N(\theta_{iAB}, s_{iAB}^2),$$

$$y_{iBC} \sim N(\theta_{iBC}, s_{iBC}^2),$$

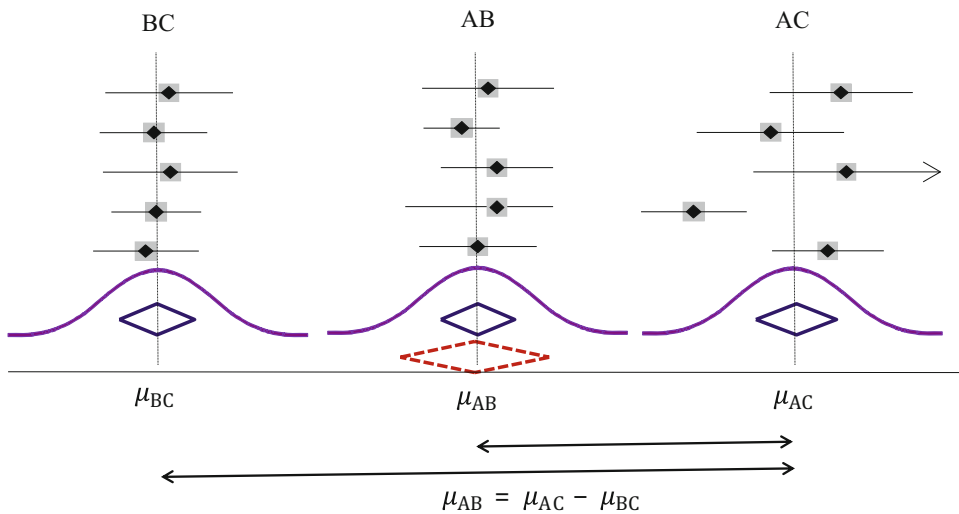
and similarly for the random effects

$$\theta_{iAC} \sim N(\mu_{AC}, \tau^2), \theta_{iAB} \sim N(\mu_{AB}, \tau^2),$$

$$\theta_{iBC} \sim N(\mu_{BC}, \tau^2).$$

This situation is depicted in Fig. 5.

The model so far is a collection of three independent meta-analyses. The three distributions relate to three different means, one mean per comparison. The consistency assumption claims that the three means are related via  $\mu_{BC} = \mu_{AC} - \mu_{AB}$ . This constraint results in the indirect estimation of  $B$  versus  $C$  and, if there are studies making the comparison directly, results also in the synthesis of the indirect



**Fig. 5** Hypothetical example from three sets of meta-analyses that form a closed loop of evidence. The diamonds represent the summary effects using a random

effects meta-analysis and a common heterogeneity parameter. The diamond in dashed line is the indirect estimate for the AB comparison

evidence with the direct evidence. Indirect estimation is represented by the dashed diamond in Fig. 5. Of course the consistency relationship works by estimating indirect and mixed estimates for all comparisons, not just *B* versus *C*.

Extending the idea to more than three treatments is straightforward. For any two treatments  $j, k = \{A, B, C, D, E\}$  compared in study  $i$ , the likelihood and the random effects distributions can be written as in a subgroup analysis, i.e., specific to the comparison  $j$  versus  $k$

$$y_{ijk} \sim N(\theta_{ijk}, s_{ijk}^2) \tag{7}$$

$$\theta_{ijk} \sim N(\mu_{jk}, \tau^2) \tag{8}$$

Assuming consistency, the means of the random effects distributions are related. Selecting again  $T - 1$  basic parameters  $\mu_{Aj}$ , all means are related via  $\mu_{jk} = \mu_{Ak} - \mu_{Aj}$ . There are as many consistency equations as comparisons-specific meta-analysis, that is,  $N_{\text{comp}}$ , each equation expressing every comparison that appears in the data as a combination of the basic parameters. This gives the model  $N_{\text{comp}} - (T - 1)$  degrees of freedom.

Note that the formulation above assumes that the different comparisons have the same heterogeneity variance, as in the meta-regression model in section “[Network Meta-analysis as a Linear Model](#).” The impact of this assumption can be better visualized in Fig. 5, where the three forest plots appear to reflect different degrees of heterogeneity. Nevertheless, the three random effects distributions have the same dispersion as a result of imposing a common heterogeneity parameter of their variance  $\tau^2$ . This will “inflate” the uncertainty in the summary estimates in the more homogeneous sets of studies *BC* and *AB* and “deflate” the uncertainty in the *AC* estimate by attaching a lower  $\tau^2$  value for this comparison. A notable consequence is that the network meta-analysis summary effect size for a particular comparison can be less precise than the summary effect size from direct evidence alone. This can happen when a comparison with very low or no heterogeneity enters a network that consists of heterogeneous comparisons. Then, the estimated common heterogeneity parameter (which will be higher than the true for the homogeneous comparison) will impose greater uncertainty in the estimate for the homogeneous comparison,

which might outweigh the gain in precision offered by including the indirect evidence.

Fitting the hierarchical model within a Bayesian framework makes the use of the true likelihood for the data easier. In the case of dichotomous outcomes, each study reports numbers of successes per arm and the likelihood (Eq. 7) is substituted by two arm-specific binomial distributions

$$r_{ij} \sim B(p_{ij}, n_{ij})$$

$$r_{ik} \sim B(p_{ik}, n_{ik})$$

Then the probabilities  $p_{ik}$ ,  $p_{ij}$  are parameterized to produce a treatment effect measure  $\theta_{ijk}$  (e.g., for  $\log(\text{OR})$ ,  $\theta_{ijk} = \text{logit}(p_{ij}) - \text{logit}(p_{ik})$ ). The hierarchical network meta-analysis model is mathematically equivalent to the meta-regression as long as the contrast-specific data are used: they both have the same number of degrees of freedom and the same number of parameters.

**Application: Network Meta-analysis as Hierarchical Model for Incident Diabetes**

As in the meta-regression approach, all comparisons included in the four three-arm trials are assumed to be evaluated in three independent two-arm studies, and this results in 30 comparisons indexed with  $i$ . The same five basic parameters  $\mu_{A_j}$  are chosen, with placebo as reference:  $\mu_{BB - P}$ ,  $\mu_{DD - P}$ ,  $\mu_{CCB - P}$ ,  $\mu_{ACE - P}$ ,  $\mu_{ARB - P}$ . Arm-specific data will be modelled using the binomial likelihood. A categorical covariate needs to be specified for each arm, with  $x_{ij}$  showing the intervention given to arm  $j$  of study  $i$  and  $x_{ik}$  the intervention of arm  $k$  ( $x_{ij}, x_{ik} = \{P, BB, D, CCB, ACE, ARB\}$ ). Fitting the model in WinBUGS, and using a half-normal prior distribution  $\tau \sim N(0, 1)$ ,  $\tau > 0$  for the common heterogeneity, gives the results in Table 6.

These estimates are comparable to the effect sizes obtained by the meta-regression approach (Table 4, Fig. 4). The most important difference is, as in subgroup analysis, in the estimation of heterogeneity. Although both meta-regression and the hierarchical model result in the same point estimate for heterogeneity of  $\hat{\tau}^2 = 0.02$ , the

**Table 6** Results of network meta-analysis as hierarchical model for incident diabetes. Log-odds ratios ( $\hat{\mu}$ ) with their standard error  $SE(\hat{\mu})$  and odds ratios (OR) with their 95% credible interval (CrI) for all basic and functional contrasts are reported

Comparison	$\hat{\mu}$	$SE(\hat{\mu})$	OR	95% CrI for OR
$\beta$ -blockers versus placebo	0.24	0.10	1.27	(1.04, 1.55)
diuretics versus placebo	0.32	0.10	1.38	(1.13, 1.68)
CCB versus placebo	0.07	0.10	1.07	(0.88, 1.30)
ACE inhibitors versus placebo	-0.09	0.08	0.91	(0.78, 1.07)
ARB versus placebo	-0.19	0.10	0.83	(0.68, 1.01)

Bayesian approach accounts for uncertainty in this value with a 95% CrI (0.01, 0.07) and provides estimates of the ORs with slightly wider confidence intervals.

One advantage of conducting network meta-analysis as a hierarchical model compared with a meta-regression approach is that ranking of all interventions included in the network is easier. This will be discussed in the next section, on the results from the model that accounts properly for the multi-arm trials.

**Models for Data that Include Multi-arm Trials**

When trials involve more than two arms, the network meta-analysis models described in sections “Network Meta-analysis as a Linear Model” and “Network Meta-analysis as a Hierarchical Model” are further complicated for two reasons. The first is the need to account for correlations induced by the fact that multi-arm trials inform more than one comparison. The second is that multi-arm studies are inherently consistent; if A, B, and C are all included within the same study  $i$  then, it is plainly the case that  $y_{iBC} = y_{iAC} - y_{iAB}$  where  $y_{ikj}$  is the effect size in study  $i$  for the contrast  $k$  versus  $j$ . This means that if a study has  $\alpha_i$  arms, then only  $\alpha_i - 1$  of the  $\alpha_i(\alpha_i - 1)/2$  possible comparisons are linearly independent, and so only  $\alpha_i - 1$  need to be modelled. This inherent consistency also



**Table 7** Data for network meta-analysis with a three-arm trial

Study $i$	No. arms $\alpha_i$	Arms/design	Data	Comparison
1	2	$A, C$	$y_{1AC}, s_{1AC}^2$	$AC$
2	2	$B, C$	$y_{2BC}, s_{2BC}^2$	$BC$
3	2	$A, B$	$y_{3AB}, s_{3AB}^2$	$AB$
4	3	$A, B, C$	$y_{4AC}, s_{4AC}^2$ $y_{4BC}, s_{4BC}^2$ $\text{cov}(y_{4AC}, y_{4AB}) = c$	$AC$ $AB$

makes the calculation of the number of degrees of freedom difficult and the formula  $N_{\text{comp}} - (T - 1)$  no longer holds (see also section “Statistical Methods to Detect Inconsistency in a Network of Interventions”).

Consider the case of three treatments and four studies as presented in Table 7. In study four, only two of the three contrasts need to be included in the model as the third effect size  $y_{iBC}$  can be simply computed as  $y_{iAC} - y_{iAB}$ . Thus, the study will contribute *directly* to two out of the three meta-analyses in Fig. 5. The two observed effect sizes  $y_{iAC}$ ,  $y_{iAB}$  are correlated as they both include the common treatment  $C$ . This covariance needs to be taken into account in the analysis.

Note that in Table 7, the data for study 4 includes the sample covariance  $\text{cov}(y_{4AC}, y_{4AB})$ , denoted also as  $c$ . The covariance can be estimated from the data as the variance of the outcome in the common arm. For example, if the outcome is continuous and the effect size is the mean difference, it turns out that  $c$  is the sample variance of the outcome in the common arm  $C$ , that is,  $sd_{iC}^2/n_{iC}$ . When the outcome is dichotomous and the effect size is the  $\ln(OR)$ , the covariance is  $c = 1/r_C + 1/(n_C - r_C)$ . When the outcome is measured on the risk ratio scale (RR), then the covariance for  $\ln(RR)$  is  $c = 1/r_C - 1/n_C$  and for risk difference it is  $c = r_C(n_C - r_C)/n_C^3$ .

The meta-regression model as presented in section “Network Meta-analysis as a Linear Model” does not account for the dependence between the observations in study 4. Moreover, correlations are present not only in the observations  $y_{4AC}, y_{4AB}$  but also in their underlying random effects  $\delta_{4AC}, \delta_{4AB}$ .

Using matrix notation, the meta-regression model in section “Network Meta-analysis as a Linear Model” will have the form

$$\begin{pmatrix} y_{1AC} \\ y_{2BC} \\ y_{3AB} \\ y_{4AC} \\ y_{4AB} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & -1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{AC} \\ \mu_{AB} \end{pmatrix} + \begin{pmatrix} \delta_{1AC} \\ \delta_{2BC} \\ \delta_{3AB} \\ \delta_{4AC} \\ \delta_{4AB} \end{pmatrix} + \begin{pmatrix} e_{1AC} \\ e_{2BC} \\ e_{3AB} \\ e_{4AC} \\ e_{5AB} \end{pmatrix} \quad (9)$$

To account for the fact that the random errors and the random effects that belong to the same study are correlated, it is assumed that  $e \sim N(0, \mathbf{S}^2)$  and  $\delta \sim N(0, \mathbf{T}^2)$  where  $e, \delta$  are the vectors of the random errors and random effects,  $\mathbf{S}^2$  is the within-studies variance-covariance matrix (estimated from the data), and  $\mathbf{T}^2$  is the between-studies variance-covariance matrix (and consists of unknown parameters to be estimated from the model). For the data in Table 7, the within-studies variance-covariance matrix is

$$\mathbf{S}^2 = \begin{pmatrix} s_{1AC}^2 & 0 & 0 & 0 & 0 \\ 0 & s_{2BC}^2 & 0 & 0 & 0 \\ 0 & 0 & s_{3AB}^2 & 0 & 0 \\ 0 & 0 & 0 & s_{4AC}^2 & c \\ 0 & 0 & 0 & c & s_{5AB}^2 \end{pmatrix}$$

whereas the between-studies variance-covariance matrix is

$$\begin{aligned}
 \mathbf{T}^2 &= \begin{pmatrix} \tau_{AC}^2 & 0 & 0 & 0 & 0 \\ 0 & \tau_{BC}^2 & 0 & 0 & 0 \\ 0 & 0 & \tau_{AB}^2 & 0 & 0 \\ 0 & 0 & 0 & \tau_{AC}^2 & \text{cov}(\delta_{4AC}, \delta_{4AB}) \\ 0 & 0 & 0 & \text{cov}(\delta_{4AC}, \delta_{4AB}) & \tau_{AB}^2 \end{pmatrix} \\
 &= \begin{pmatrix} \tau^2 & 0 & 0 & 0 & 0 \\ 0 & \tau^2 & 0 & 0 & 0 \\ 0 & 0 & \tau^2 & 0 & 0 \\ 0 & 0 & 0 & \tau^2 & \tau^2/2 \\ 0 & 0 & 0 & \tau^2/2 & \tau^2 \end{pmatrix}
 \end{aligned}$$

As discussed before, it is often the case that a common heterogeneity parameter is assumed; that is,  $\tau_{AC}^2 = \tau_{BC}^2 = \tau_{AB}^2 = \tau^2$ . This assumption offers an advantage in the case of multi-arm studies as it considerably simplifies the between-studies variance-covariance matrix  $\mathbf{T}^2$ . It can be shown that when heterogeneity is equal across comparisons, the covariance of any two random effects is  $\tau^2/2$ . Then the matrix  $\mathbf{T}^2$  has  $\tau^2$  in the diagonal and  $\tau^2/2$  in the cells that refer to pairs of effects from the same study.

Similar considerations need to be made for the hierarchical model. The distributions (7) and (8) apply only to studies  $i = 1, 2, 3$ . For  $i = 4$  the likelihood of the two-dimensional vector of effect sizes is

$$\begin{pmatrix} y_{4AC} \\ y_{4AB} \end{pmatrix} \times \sim \text{MVN} \left( \begin{pmatrix} \theta_{4AC} \\ \theta_{4AB} \end{pmatrix}, \begin{pmatrix} s_4^2 & \text{cov}(y_{4AC}, y_{4AB}) \\ \text{cov}(y_{4AC}, y_{4AB}) & s_5^2 \end{pmatrix} \right)$$

and the random effects are distributed assuming equal heterogeneities as

$$\begin{pmatrix} \theta_{4AC} \\ \theta_{4AB} \end{pmatrix} \sim \text{MVN} \left( \begin{pmatrix} \mu_{AC} \\ \mu_{AB} \end{pmatrix}, \begin{pmatrix} \tau^2 & \tau^2/2 \\ \tau^2/2 & \tau^2 \end{pmatrix} \right). \tag{10}$$

The consistency equations remain as presented in section “Network Meta-analysis as a Hierarchical Model”.

With arm-specific data and a hierarchical structure, no correlations are needed to account in the likelihood as the observations in arms are independent. For example, if study 4 presents the number of successes  $r_{4A}, r_{4B}, r_{4C}$  for a dichotomous outcome out of the total  $n_{4A}, n_{4B}, n_{4C}$  randomized, then the

likelihood of the data consists of three binomial distributions with event probability parameters  $p_{4A}, p_{4B}, p_{4C}$  which parameterized will give two effect sizes  $\theta_4, \theta_5$  that correspond to underlying relative effects for treatments A and B compared to C (see section “Meta-analysis as Hierarchical Model”). So, for studies  $i = 1, 2, 3$ , the underlying random effects  $\theta_{ijk}$  follow independent distributions as described in Eq. 8, but the random effects  $\theta_{4AC}, \theta_{4AB}$  from the fourth study will follow the multivariate normal distribution (10).

*Technical note:* the multivariate normal distribution above can be decomposed into a series of conditional distributions; this offers computational advantages. Distribution (10) can be written as a set of one unconditional and one conditional distribution:

$$\begin{aligned}
 \theta_{4AB} &\sim N(\mu_{AC}, \tau^2) && \text{and} \\
 \theta_{4AC} | \theta_{4AB} &\sim N\left(\mu_{AC} + \frac{1}{2}(\theta_{4AB} - \mu_{AB}), \frac{3\tau^2}{4}\right)
 \end{aligned}$$

More generally, if a study  $i$  has  $a_i$  arms that correspond to treatments  $T_i = \{A, B, C, D, \dots\}$  in this presented order, the  $(a_i - 1)$ -dimensional normal distribution of all treatments versus A can be “decomposed” by writing the independent distribution for  $\theta_{iAB}$ , then the conditional  $\theta_{iAC} | \theta_{iAB}$ , then  $\theta_{iAD} | \theta_{iAB}, \theta_{iAC}$ , and so on. The distribution of the random effect  $\theta_{iAj}$  conditional on all “previous” comparisons  $\theta_{iAk}$  has mean:

$$\mu_{Aj} + \frac{1}{a_i - 1} \sum_{k < j} (\theta_{iAk} - \mu_{Ak})$$

with variance

$$\frac{a_i}{a_i - 1} \frac{\tau^2}{2}$$

where  $k < j$  means that comparison  $Ak$  was been modelled before  $Aj$ .

**Application: Network Meta-analysis with Multi-Arm Trials as Hierarchical Model for Incident Diabetes**

In this application index  $i$  refers to studies ( $i = 1, \dots, 22$ ). There are 18 studies that compared only two interventions and thus have only

**Table 8** Results of network meta-analysis as hierarchical model for incident diabetes taking into account multi-arm trials. Log-odds ratios ( $\hat{\mu}$ ) with their standard error  $SE(\hat{\mu})$  and odds ratios (OR) with their 95% credible interval (CrI) for all basic and functional contrasts are reported

Comparison	$\hat{\mu}$	$SE(\hat{\mu})$	OR	95% CrI for OR
$\beta$ -blockers versus placebo	0.24	0.10	1.27	(1.04,1.55)
diuretics versus placebo	0.32	0.10	1.38	(1.13,1.68)
CCB versus placebo	0.07	0.10	1.07	(0.88,1.30)
ACE inhibitors versus placebo	-0.09	0.08	0.91	(0.78,1.07)
ARB versus placebo	-0.19	0.10	0.83	(0.68,1.01)

two arms ( $\alpha_i = 2$ ), and there are four studies with three arms ( $\alpha_i = 3$ ). The variable  $\alpha_i = \{2, 3\}$  needs to be specified for each study  $i$  and then a binomial likelihood is assumed for the number of patients in all arms of each study. Using the index  $j$  to show the arm (treatment) within a study, the binomial likelihood is written as

$$r_{ij} \sim B(p_{ij}, n_{ij}), \quad i = 1, \dots, 22, \\ j \in \{P, BB, D, CCB, ACE, ARB\}$$

The probabilities  $p_{ij}$  can be parameterized to model  $\alpha_i - 1$  effect sizes as

$\text{logit}(p_{i1}) = u_i$  for the “first” (reference) arm in each study that pertains to treatment  $j$

$\text{logit}(p_{ij}) = u_i + \theta_{ijk}$  for the other arms in the study

The underlying  $\ln(OR), \theta_{ijk}$ , compares treatments  $k$  and  $j$  (reported in the first arm) where  $j, k \in \{P, BB, D, CCB, ACE, ARB\}$ . For the multi-arm trials the correlation between  $\theta_{ijk}$  and  $\theta_{ijl}, l \neq j \neq k$  in the same trial is taken into account by the conditional mean and variance of their distribution. Table 8 shows the results of fitting this model in WinBUGS against placebo (basic parameters and prior distribution for

heterogeneity are as in the application of hierarchical model that does not account for multi-arm trials).

The estimate of common heterogeneity is 0.02 with 95% CrI (0.01, 0.07). Very little change is observed compared with the analyses above in which the correlations between multiple arms were ignored; this is probably due to the fact that multi-arm trials represent only the 18% of our data. All pairwise ORs are presented in Fig. 4.

The posterior deviance from the analysis is  $\bar{D} = 53.26$  which, when compared to the number of data points (48), suggests a rather poor fit of the model to the data. The *DIC* of the model was estimated as 91.4.

### Network Meta-analysis as a Multivariate Meta-analysis

Multivariate meta-analysis is an extension of meta-analysis that simultaneously synthesizes data on more than one outcome per study. For example, studies which compare antihypertensive interventions might measure the two related outcomes fatal stroke and nonfatal stroke. Some studies will only report fatal or only nonfatal stroke, others will report both. Because these two outcomes are correlated, there are important benefits in analyzing them jointly via multivariate meta-analysis, including improved precision and calculation of confidence regions for both outcomes (Jackson et al. 2011; Riley 2009).

Multiple treatment comparisons reported by multi-arm studies may be viewed in a similar way to multiple outcomes. Specifically, the *basic contrasts* can be considered analogous to different outcomes, where the basic contrasts are the set of necessary comparisons to represent all comparisons under the consistency assumption (e.g., the contrasts  $A_j$  of each treatment versus a common reference treatment  $A$ ). Studies may report on many, all or a single basic contrast. In the example of Table 7, the basic contrasts are the contrasts  $AC$  and  $AB$ . So, study 1 reports on the first “outcome”  $AC$ , study 3 reports on the second “outcome”  $AB$ , and study 4 reports on both “outcomes.”

A departure from the analogy arises for study 2, which compares  $B$  and  $C$ . This study gives

**Table 9** Data for network meta-analysis assuming data augmentation for study 2

Study <i>i</i>	No. arms $\alpha_i$	Arms/ design	Data	Contrast
1	2	A,C	$y_{1AC}, s_{1AC}^2$	$AC$
2	2	A imputed, B,C	$y_{2AC}, s_{2AC}^2$ $y_{2BC}, s_{2BC}^2$ cov $(y_{2AC}, y_{2AB}) = c_2$	$BC$
3	2	A,B	$y_{3AB}, s_{3AB}^2$	$AB$
4	3	A,B,C	$y_{4AC}, s_{4AC}^2$ $y_{4BC}, s_{4BC}^2$ cov $(y_{4AC}, y_{4AB}) = c_4$	$AC$ $AB$

information about a combination of the two “outcomes.” To model the  $BC$  study, the assumption of consistency is employed. As presented in section “Assumptions Underlying Indirect and Mixed Comparisons” transitivity suggests that the missing arm in a study is missing at random. If study 2 had reported arm A, then the two “outcomes”  $y_{2AC}, y_{2AB}$  could have been derived. This suggests a simple imputation strategy, whereby data in the “missing” arm can be created via a data augmentation technique (White 2011). The imputed data are designed to provide minimal information, for example, by giving them a very large variance (for continuous outcomes) or a very small sample size less than one (for dichotomous outcomes). The two effect sizes  $y_{2AC}, y_{2AB}$  are correlated, and together they give information about the direct observed  $BC$  contrast. The data can be rewritten as in Table 9:

Then, following standard multivariate meta-regression techniques and assuming equal heterogeneities:

$$\begin{pmatrix} y_{1AC} & \cdot \\ y_{2AC} & y_{2BC} \\ \cdot & y_{3AB} \\ y_{4AC} & y_{4BC} \end{pmatrix} = \begin{pmatrix} 1 & \cdot \\ 1 & 1 \\ \cdot & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_{AC} \\ \mu_{AB} \end{pmatrix} + \begin{pmatrix} \delta_{1AC} & \cdot \\ \delta_{2AC} & \delta_{2BC} \\ \cdot & \delta_{3AB} \\ \delta_{4AC} & \delta_{4BC} \end{pmatrix} + \begin{pmatrix} e_{1AC} & \cdot \\ e_{2AC} & e_{2BC} \\ \cdot & e_{3AB} \\ e_{4AC} & e_{4BC} \end{pmatrix}$$

$$\text{with } \begin{pmatrix} \delta_{1AC} & \cdot \\ \delta_{2AC} & \delta_{2BC} \\ \cdot & \delta_{3AB} \\ \delta_{4AC} & \delta_{4BC} \end{pmatrix} \sim NMV \left( \begin{pmatrix} 0 & \cdot \\ 0 & 0 \\ \cdot & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & 0 & 0 & 0 \\ 0 & M & 0 & 0 \\ 0 & 0 & \tau^2 & 0 \\ 0 & 0 & 0 & M \end{pmatrix} \right)$$

$$\text{and } \begin{pmatrix} e_{1AC} & \cdot \\ e_{2AC} & e_{2BC} \\ \cdot & e_{3AB} \\ e_{4AC} & e_{4BC} \end{pmatrix}$$

$$\sim NMV \left( \begin{pmatrix} 0 & \cdot \\ 0 & 0 \\ \cdot & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} s_{1AC}^2 & 0 & 0 & 0 \\ 0 & L_1 & 0 & 0 \\ 0 & 0 & s_{3AB}^2 & 0 \\ 0 & 0 & 0 & L_2 \end{pmatrix} \right)$$

$$L_1 = \begin{pmatrix} s_{2AC}^2 & c_2 \\ c_2 & s_{2BC}^2 \end{pmatrix}, L_2 = \begin{pmatrix} s_{4AC}^2 & c_4 \\ c_4 & s_{5BC}^2 \end{pmatrix} \text{ and}$$

$$M = \begin{pmatrix} \tau^2 & \tau^2/2 \\ \tau^2/2 & \tau^2 \end{pmatrix}.$$

It has been shown that the choice of the basic contrasts and the data augmentation parameters do not impact on the estimation of the effects. The model can be fitted using the `mvmeta` command in STATA, providing estimates of the means and standard errors for the basic parameters  $\mu_{AC}$  and  $\mu_{BC}$ . As described earlier, combinations of the basic parameters can give estimates of all functional parameters through application of the consistency equations and uncertainty in these can be obtained by incorporating covariances between the estimates.

**Application: Network Meta-analysis for Incident Diabetes as a Multivariate Meta-analysis**

The use of standard multivariate meta-regression requires that all studies have data for the treatment that has been chosen as reference. When none of the treatments of the network is common to all studies (as in the current example), one of the treatments can be chosen to be the common reference treatment and then the data augmentation technique is applied. More specifically, choosing

**Table 10** Results of network meta-analysis as multivariate meta-analysis for incident diabetes. Log-odds ratios ( $\hat{\mu}$ ) with their standard error  $SE(\hat{\mu})$  and odds ratios (OR) with their 95% confidence interval (CI) for all basic contrasts are reported

Comparison	$\hat{\mu}$	$SE(\hat{\mu})$	OR	95% CI for OR
$\beta$ -blockers versus placebo	0.21	0.08	1.24	(1.05,1.44)
diuretics versus placebo	0.28	0.08	1.32	(1.12,1.56)
CCB versus placebo	0.04	0.08	1.04	(0.89,1.21)
ACE inhibitors versus placebo	-0.12	0.07	0.88	(0.77,1.10)
ARB versus placebo	-0.19	0.09	0.83	(0.70,0.98)

placebo as reference implies that in studies without a placebo arm, we need to “impute” data for a very small sample size for an assumed placebo arm, and here the values  $r_{iP} = 0.001$  and  $n_{iP} = 0.01$  are used. Then all studies will report on the relative effectiveness of the included treatments versus placebo,  $y_{iPj}$  where  $j = \{BB, D, CCB, ACE, ARB\}$ . The sample variance-covariance matrix  $S$  of all  $y_{iPj}$  needs to be specified. As the outcome is measured using the (OR), the variances of all observations are calculated using the formula:

$$s_{iPj}^2 = \frac{1}{r_P} + \frac{1}{n_P - r_P} + \frac{1}{r_j} + \frac{1}{n_j - r_j}$$

and the covariances are calculated as

$$\text{cov}(y_{iPj}, y_{iPk}) = \frac{1}{r_P} + \frac{1}{n_P - r_P}$$

The variance-covariance matrix of the random effects can be modelled in various ways. The most flexible structure is to estimate different heterogeneity variances  $\tau_{Pj}^2$  for each comparison ( $j$  vs.  $P$ ). In the analyses that follow, a much more restricted structure is used, following the assumption of a common heterogeneity variance as has been used

in previous analyses in the chapter. This sets  $\tau_{Pj}^2 = \tau^2$  so all covariances between random effects are  $\tau^2/2$ . This model can be implemented in STATA using the `mvmeta` command with the option `bscov()`, which gives the results of Table 10.

Estimates for all functional comparisons can be derived with the use of consistency equations. There are small differences between the results of this approach with the corresponding results of the hierarchical model. Using the restricted maximum likelihood estimator in `mvmeta` results in  $\hat{\tau}^2 = 0.01$ , which is the same as the heterogeneity estimated in the hierarchical model.

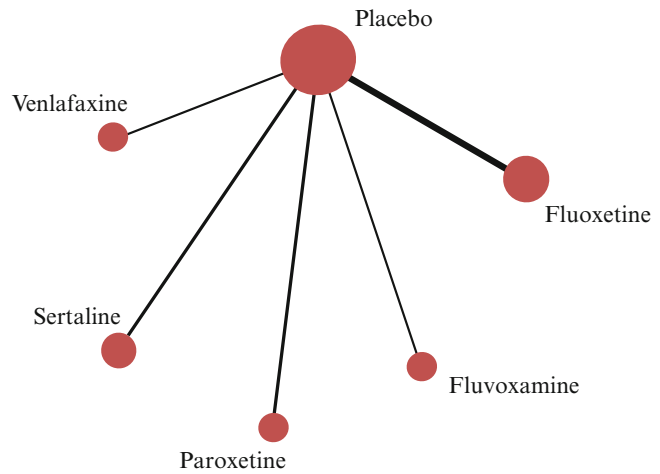
### Assumptions of Network Meta-analysis

As presented in section “[Estimating Inconsistency in Mixed Comparisons](#),” inconsistency in a network can manifest as a disagreement between different sources of evidence for the same comparison and can be identified statistically. For example, an indirect estimate of  $A$  versus  $B$  via a treatment  $C$  can be in conflict with the direct estimate or with another indirect estimate, e.g.,  $A$  versus  $B$  via a treatment  $C$ .

Both the likelihood of transitivity (based on clinical and epidemiological considerations) and any evidence of (in)consistency (based on statistical considerations) should be evaluated in a network as part of a network meta-analysis. Conceptual evaluation involves a priori judgements about the comparability of the studies across comparisons with respect to the distribution of potential confounders, considering whether treatments were all given for the same indication and considering whether anchor treatments are equivalent. Such judgements should be made ideally before the outcome data are extracted but after the studies and their characteristics are collected.

Although transitivity and consistency are interwoven concepts and are often thought of as one, it can be useful to consider them separately for ease of evaluation. Consider, for instance, the network presented in Fig. 6 where all treatments have been compared with placebo but not with each other. In

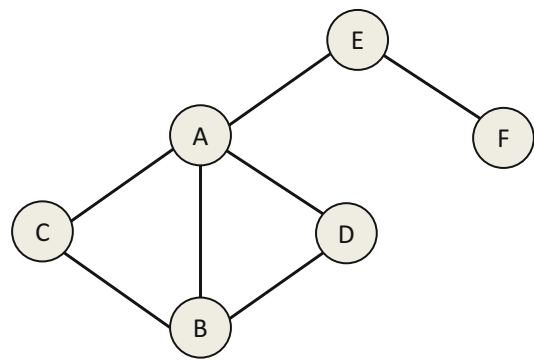
**Fig. 6** Plot of a network for efficacy of pharmacotherapeutic agents for anxiety disorders in children and adolescents. The size of the nodes is proportional to the number of studies that evaluate each intervention and the thickness of the lines is proportional to the frequency of each comparison in the network (Uhtman and Abdulmalik 2010)



this network, each given comparison is informed by a single source of evidence (there are no closed loops in the network), and therefore it is impossible to observe inconsistency by statistical means. However, network meta-analysis can be performed and will yield estimates for the relative effects of all active treatments (which in this case are only indirect estimates). These estimates are valid only if transitivity can be assumed for the common “anchor” treatment (placebo). Therefore, judgements about the plausibility of transitivity should be done for the entire network as indirect evidence is derived for all treatments irrespectively of whether they belong to “closed loops” or not.

### Statistical Methods to Detect Inconsistency in a Network of Interventions

The assumption of consistency can be evaluated statistically in a full network by extending the idea outlined in section “[Estimating inconsistency in Mixed comparisons](#)”. A network often comprises several closed loops (triangles, quadrilaterals, etc.) which bring together evidence for the same comparison from direct and various indirect routes. Within each one of these closed loops, the consistency assumption, seen as agreement between direct and indirect estimates, can be evaluated. A first approach is therefore to evaluate



**Fig. 7** Fictional network of interventions with two triangular loops

inconsistency in all loops by calculating the loop-specific inconsistency factors  $\hat{IF}$ , their confidence intervals, and a z-test for each one. The  $\hat{IF}$ s of a network can be presented in a forest plot like graph where deviations from the consistency assumption would be reflected in loops with confidence intervals incompatible with zero. Note that not all loops need to be presented and tested; for example, if a quadrilateral consists of two triangles inconsistency needs to be evaluated only in the triangles. Consider, for example, the network in Fig. 7. Inconsistency can be evaluated in  $ABC$  and  $ADB$  loops; if these are consistent then the quadrilateral  $ADBC$  would be consistent as well.

The loop-based approach is simple and can be useful for identifying loops that deviate from consistency, but has important limitations. An obvious

problem is that the loop-specific tests are not independent as they share groups of studies. Consider, for example, the network in Fig. 7 and imagine that the  $AB$  comparison is informed by a single study in which an unobserved characteristic produced an estimate very different from what would be expected in the other studies. Then both  $ABC$  and  $ADB$  loops will present inconsistency because the respective  $\hat{IF}$ s share the same deviant  $AB$  study.

The loop-based approach does not provide a network-specific estimate of the inconsistency. The multiple dependent tests cannot be summarized into a global network-specific test. It is also unclear how to treat multi-arm trials, which are inherently consistent. Because of the dependence between the loops and the multiple testing nature of the approach, the results should be interpreted with caution; the absence of inconsistent loops may be reassuring for the assumption of consistency (notwithstanding the lack of power of such tests), but the presence of statistically significant loops cannot be used to infer the magnitude of inconsistency in a network.

In the special case where the loops share a single comparison as in Fig. 7, a chi-squared test can be applied (Caldwell et al. 2010). For the same comparison  $AB$ , there are three estimates; the direct estimate  $\hat{\mu}_{AB}^D$ , the two indirect estimates via  $C$  and  $D$ ,  $\hat{\mu}_{ABviaC}^I$  and  $\hat{\mu}_{ABviaD}^I$ , respectively, with their estimated variances noted as  $\hat{v}_{AB}^D$ ,  $\hat{v}_{ABviaC}^I$ ,  $\hat{v}_{ABviaD}^I$ . The mixed estimate  $\hat{\mu}_{AB}^M$  is the weighted average of the three estimates with weights being the inverse of the variance. To test both  $ABC$  and  $ADB$  loops (and therefore provide a global test for the network) the following chi-squared test can be applied

$$\frac{(\hat{\mu}_{AB}^D - \hat{\mu}_{AB}^M)^2}{\hat{v}_{AB}^D} + \frac{(\hat{\mu}_{ABviaC}^I - \hat{\mu}_{AB}^M)^2}{\hat{v}_{ABviaC}^I} + \frac{(\hat{\mu}_{ABviaD}^I - \hat{\mu}_{AB}^M)^2}{\hat{v}_{ABviaD}^I} \sim \chi_2^2$$

This can be generalized to combine testing for disagreement between direct estimate and  $l$  independent indirect sources; the weighted sum of the difference of each estimate from the mixed

estimate will follow a chi-squared distribution with  $l$  degrees of freedom.

The results of the loop-based approach can vary substantially depending on the method used to derive the pairwise estimates and their variances. In the presence of heterogeneity, the uncertainty of  $\hat{IF}$  will be larger in a random effects analysis compared with a fixed-effect analysis, and therefore there will be less chance of identifying statistically significant inconsistencies. The random effects approach will also give different results depending on which method will be used to estimate the heterogeneity parameter  $\tau$  (e.g., method of moments, restricted maximum likelihood). Some approaches will give larger estimates than others, resulting in different estimates for the uncertainty of  $\hat{IF}$ . Moreover, the estimated pairwise variances will change depending on whether the same or different heterogeneity parameters are assumed in the loop.

There is currently limited empirical evidence about the occurrence of statistical inconsistency. A study evaluated 112 triangular networks of which only 16 were found inconsistent (Song et al. 2011). O'Regan et al. (2009) empirically evaluated the agreement between indirect and mixed estimates that appear in networks of at least four treatments. Using a fixed-effect approach, they concluded that the two indirect and mixed estimates did not show important differences, although the 51 comparisons they examined came from only seven reviews.

Approaches that evaluate inconsistency globally in a network rather than testing each loop have gained in popularity but are typically cumbersome to apply and have limitations. For network models fitted within a Bayesian framework, the consistency assumption can be evaluated by comparing a model that assumes consistency with one that does not, using the  $DIC$  (Spiegelhalter et al. 2002). The model without consistency is the model described in section "Consistency Models" but without the consistency equations to derive indirect and mixed estimates. The inconsistency model relies only on direct evidence and is equivalent to a

series of pairwise meta-analyses (usually assuming, however, that they share the same heterogeneity parameter). The assumption of consistency is challenged when the inconsistency model presents, for the same data, a better trade-off between model fit and complexity; this is the case when the DIC for the inconsistency model is lower to the DIC for the consistency model by more than three units. An important drawback with this method is that results may depend on the parameterization of the multi-arm trials, from which only some of the study-specific effect sizes enter the model. Approaches that simultaneously test and account for inconsistency are discussed in the section “[Inconsistency models](#)”.

### Application: Statistical Evaluation of Inconsistency in Each Loop of Incident Diabetes Network

The network includes 16 “triangles” that can be evaluated for inconsistency. For the calculation of all inconsistency factors, the formulae of section 3.3.3 is employed. Then the estimates with their 95% CI can be plot in a forest plot. The pairwise effect sizes were estimated using the random effects model assuming different and loop-common heterogeneity parameters. There are no important differences between the two forest plots; both include two inconsistent loops.

The hierarchical model is fitted as described in section “[Network Meta-analysis as a Hierarchical Model](#)” but omitting the consistency equations (i.e., an inconsistency model); i.e., this is essentially a sequence of pairwise meta-analyses. The value of the posterior deviance was  $\bar{D} = 50.85$  and  $DIC = 93.6$ . Comparing the  $\bar{D}$  value to that obtained from the consistency model, since the difference in DIC is smaller than three points, this suggests that the inconsistency model fits the data better and might also be the most parsimonious model.

### Inconsistency Models

Two major approaches have been proposed so far to address inconsistency. The first approach was

proposed in (Lu and Ades 2006) and is based on the idea that inconsistency is a property of closed loops and a network can have as many inconsistencies as functional parameters. Recently, an approach has been proposed which extends the idea of inconsistency: it does not apply only to the disagreement between direct and indirect estimates in a loop but also disagreement between studies that report the same comparison but include different sets of treatments. The two approaches are outlined below, starting from the data in Table 7.

The loop-based inconsistency model assumes that inconsistency arises when the consistency equations between functional and basic parameters do not hold. Hence, an obvious solution is to “relax” the assumption by adding an extra term to account for inconsistencies. In the example of Table 7, there are two basic parameters  $\mu_{AC}$ ,  $\mu_{AB}$  and one functional  $\mu_{BC} = \mu_{AC} - \mu_{AB}$ . This reflects the closed loops  $ABC$ . Inconsistency in this loop can be accounted for if it is assumed that

$$\mu_{BC} = \mu_{AC} - \mu_{AB} + w_{ABC}$$

where  $w_{ABC}$  measures the amount of inconsistency in the loop. The term is also called an inconsistency factors and in fact in the absence of multiple correlated loops is analogous to the simple  $\hat{IF}$ . In complex networks where many inconsistency factors exist, the parameters  $w_{jkf}$  are assumed to be randomly distributed with expectation zero:

$$w_{jkf} \sim N(0, \sigma^2)$$

The variance  $\sigma^2$  is often referred to as the *inconsistency variance* in analogy with the heterogeneity variance  $\tau^2$  in the distribution of the study-specific random effects  $\delta_i \sim N(0, \tau^2)$ . The inconsistency  $\sigma^2$  describes the amount of variability across loops in the conflict between direct and indirect evidence. Monitoring the individual  $w_{jkf}$ s for large values will reveal loops with important inconsistency, whereas comparison of  $\sigma^2$  to  $\tau^2$  will show how much inconsistency exists compared with the heterogeneity.

As the degrees of freedom in a network describe the number of functional parameters, there are



$N_{\text{com}} - (T - 1)$  many inconsistency factors. Problems arise with this approach when there are multi-arm trials. The *ABC* trials in Table 7 are inherently consistent, and therefore the *BC* comparison reported in these studies does not contribute to the inconsistency as much as the *BC* comparison in an independent study. Lu and Ades suggested adjusting the inconsistency degrees of freedom to  $ICDF = N_{\text{com}} - (T - 1) - S$  where  $S$  is the number of independent inconsistency relations in which the corresponding parameters are supported by no more than two independent sources of evidence. In practice,  $S$  is the number of functional comparisons where two out of the three parameters are only estimated in multi-arm trials.

The difficulties in fully defining loop inconsistency when there are multi-arm studies motivated the concept of “design inconsistency.” Design inconsistency reflects the belief that studies which include different treatments might give different estimates for the same comparison. For example, an *AB* study and *ABC* study might provide different estimates because of their different design. Design inconsistency can be thought of as a special case of source-specific heterogeneity: variation between the estimates for the same comparison due to differences in the total treatments included. In the data of Table 7, this means adding an inconsistency factor for the disagreement between the three estimates in the *ABC* study and the *AB*, *AC*, and *BC* studies. The model with both loop and design inconsistency has  $N_{\text{Comp} \times \text{Design}} - (T - 1)$  inconsistency factors, where  $N_{\text{Comp} \times \text{Design}}$  is the number of independent comparisons per design. In Table 7 there is one independent comparison for each two-arm trial and two independent comparisons for the three-arm trials. This results in a total of three inconsistency factors for the network. These inconsistency factors are comparison-specific and are attached to every study reporting that comparison. For instance, one inconsistency factor is attached to each *AB*, *AC*, and *BC* study, respectively, ( $w_{AB}$ ,  $w_{AC}$ ,  $w_{BC}$ ). As the inconsistency factors derived in this way are independent, they can be summarized in a single test for the entire network (see White 2011 for details).

One further approach for detecting inconsistency in a network meta-analysis is “node splitting” (Dias et al. 2010) where a “node” refers to each summary effect generated from the network meta-analysis. This approach is based on the separation of the information contributing to each node into the direct and indirect evidence, within a single model. The node-splitting approach allows the analyst to split the network-wide information contributing to the summary estimate into the evidence directly comparing *B* versus *C* ( $\hat{\mu}_{BC}^D$ ) and all the remaining “indirect” evidence for *B* versus *C* ( $\hat{\mu}_{BC}^I$ ) after the studies directly comparing *B* to *C* have been removed. The extent of agreement between the direct and indirect estimates defines the magnitude of consistency. Note that this is a computationally intensive approach involving models that can be difficult to parameterize; care should be taken to ensure that multi-arm trials are handled correctly and to ensure that split nodes are actually from contrasts contributing to suspect loops.

#### **Application: Hierarchical Inconsistency Model for Network Meta-analysis in Incident Diabetes**

The application of a hierarchical inconsistency model requires careful choice of the basic parameters  $\mu_{Aj}$  and the inconsistency factors  $w_{j|k}$ , as well as the appropriate parameterization of multi-arm trials. First, all basic contrasts need to be informed directly from at least one study. Choosing placebo as reference treatment (*A*) satisfies this condition, because all other treatments are compared directly with placebo in at least one study. Second, the four multi-arm trials included in the data may modify the number of ICDF that should be included in the model. However, as all consistency equations are informed by at least three independent sources of evidence, it is

$$\begin{aligned} ICDF &= N_{\text{com}} - (T - 1) - S \\ &= 14 - (6 - 1) - 0 = 9 \end{aligned}$$

The consistency relations can be relaxed to include the nine inconsistency parameters:

**Table 11** Results of inconsistency hierarchical model for network meta-analysis for incident diabetes. Inconsistency factors ( $w$ ), log-odds ratios ( $\hat{\mu}$ ) with their standard error  $SE(\hat{\mu})$ , and odds ratios ( $OR$ ) with their 95% credible interval

(CrI) for all basic and functional contrasts are reported. Missing values of  $w$  correspond to basic contrasts or functional contrasts without direct estimates available

Comparison	$w_{Pjk}$	$\hat{\mu}$	$SE(\hat{\mu})$	OR	95% CI for OR
$\beta$ -blockers versus placebo	—	0.23	0.11	1.26	(1.03, 1.62)
diuretics versus placebo	—	0.31	0.10	1.36	(1.13, 1.71)
CCB versus placebo	—	0.06	0.10	1.06	(0.89, 1.32)
ACE inhibitors versus placebo	—	-0.13	0.08	0.88	(0.75, 1.03)
ARB versus placebo	—	-0.20	0.10	0.82	(0.66, 1.00)
diuretics versus $\beta$ -blockers	-0.02	0.08	0.12	1.08	(0.86, 1.36)
CCB versus diuretics	0.00	-0.25	0.11	0.78	(0.62, 0.97)
CCB versus $\beta$ -blockers	-0.01	-0.17	0.11	0.84	(0.68, 1.04)
ACE inhibitors versus CCB	0.01	-0.19	0.11	0.83	(0.65, 1.00)
ACE inhibitors versus diuretics	0.01	-0.44	0.11	0.65	(0.50, 0.78)
ACE inhibitors versus $\beta$ -blockers	0.04	-0.36	0.12	0.70	(0.53, 0.85)
ARB versus ACE inhibitors	—	-0.07	0.12	0.93	(0.73, 1.18)
ARB versus CCB	0.00	-0.26	0.12	0.77	(0.59, 0.95)
ARB versus diuretics	-0.01	-0.51	0.13	0.60	(0.45, 0.76)
ARB versus $\beta$ -blockers	0.02	-0.43	0.13	0.65	(0.49, 0.81)

$$\mu_{D-BB} = \mu_{D-P} - \mu_{BB-P} + w_{P-D-BB}$$

$$\mu_{CCB-D} = \mu_{CCB-P} - \mu_{D-P} + w_{P-CCB-D}$$

$$\mu_{CCB-BB} = \mu_{CCB-P} - \mu_{BB-P} + w_{P-CCB-BB}$$

$$\mu_{ACE-CCB} = \mu_{ACE-P} - \mu_{CCB-P} + w_{P-ACE-CCB}$$

$$\mu_{ACE-D} = \mu_{ACE-P} - \mu_{D-P} + w_{P-ACE-D}$$

$$\mu_{ACE-BB} = \mu_{ACE-P} - \mu_{BB-P} + w_{P-ACE-BB}$$

$$\mu_{ARB-CCB} = \mu_{ARB-P} - \mu_{CCB-P} + w_{P-ARB-CCB}$$

$$\mu_{ARB-D} = \mu_{ARB-P} - \mu_{D-P} + w_{P-ARB-D}$$

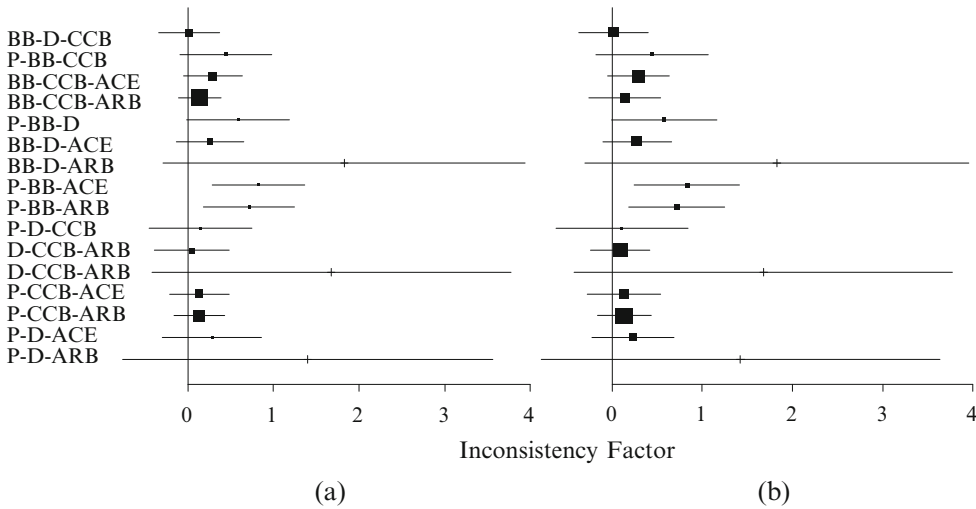
$$\mu_{ARB-BB} = \mu_{ARB-P} - \mu_{BB-P} + w_{P-ARB-BB}$$

where  $w_{Pjk} \sim N(0, \sigma^2)$  for  $j, k = \{BB, D, CCB, ACE, ARB\}$ . The rest of the model is the same with the consistency hierarchical model (accounting for multi-arm trials). Moreover, contrasts that are informed only from multi-arm trials need to be expressed in model parameters. Such contrasts are  $\beta$ -blockers versus placebo, included in a D-BB-P study, and ACE inhibitors versus CCB, included in two ACE-CCB-BB, and one ACE-CCB-D studies.

Since the model considers treatment 1 as baseline treatment ( $A$ ) of each study, we need in the data for the D-BB-P trial placebo to be the first treatment and for the other three studies CCB or ACE inhibitors.

Table 11 shows the results of fitting this model in WinBUGS employing a half-normal prior distribution on the inconsistency variance  $\sigma^2$  (the same as for the heterogeneity  $\tau^2$ ).

Heterogeneity and inconsistency variances were estimated as 0.02 and 0.01, respectively, with 95% CrI (0, 0.06) and (0, 0.13), respectively. Some  $w$ -factors are quite large in relation to the treatment effect estimates, indicating that there is probably some inconsistency in the network. Note this is in agreement with the loop-specific approach. The loop placebo versus ACE inhibitors versus  $\beta$ -blockers presents the largest inconsistency value (0.04) followed by the loops placebo versus ARB versus  $\beta$ -blockers (0.02) and placebo versus diuretics versus  $\beta$ -blockers (-0.02). The first two loops were also identified as inconsistent in Fig. 8, and the last was appeared marginally consistent. There are no large differences in the point estimates of the summary ORs compared to those from the consistency model. However, the 95% CrI from the inconsistency model are wider to account for inconsistency.



**Fig. 8** Inconsistency factors of all triangles of incident diabetes network with (a) a different heterogeneity estimate for each comparison and (b) with a common

heterogeneity estimate within each triangle. Triangles with statistically significant inconsistency factors (their 95% CI does not include 0) are considered as inconsistent

The DIC of the model was 92.1 and  $\bar{D} = 53.15$  showing that accounting for inconsistency does not improve the fit of the model as the consistency model resulted in almost the same values.

### Exploring Heterogeneity and Inconsistency: Network Meta-regression

When heterogeneity is found in a pairwise meta-analysis, subgroup analysis or meta-regression are employed to explore possible sources. Network meta-regression is an extension of network meta-analysis to include covariates and can be used to explore heterogeneity and/or inconsistency. Covariates typically include study-specific variables such as setting or length of follow-up, within-trial bias characteristics such as the quality of randomization, of allocation concealment and blinding, or patient-level characteristics such as age or sex. Meta-regression is equivalent to subgroup analysis for dichotomous or categorical explanatory variables. Characteristics such as differences in baseline risk (if there is a common comparator) and sample size (as a single proxy for study quality) can also be considered.

The network meta-regression model as a hierarchical model is

$$y_{ijk} \sim N(\theta_{ijk}^*, s_{ijk}^2)$$

$$\theta_{ijk}^* = \theta_{ijk} + b_{ijk}C_{ijk}$$

$$\theta_{ijk} \sim N(\mu_{jk}, \tau_r^2)$$

where  $b_{ijk}$  are the regression coefficients for study  $i$  and comparison  $jk$  and  $C_{ijk}$  the explanatory variable. The regression coefficients can be assumed to be fixed across studies ( $b_{ijk} = \beta_{jk}$ ) or, if there are many studies per comparison, as exchangeable across studies ( $b_{ijk} \sim N(\beta_{jk}, \gamma^2)$ ). The model can be applied to multi-arm trials and also extended to account for inconsistency as described in previous sections.

Consistency can be imposed for the regression coefficients by choosing a reference treatment  $A$  and defining  $\beta_{jk} = \beta_{Ak} - \beta_{Aj}$  (Cooper et al. 2009). To improve power, the independent  $\beta_{Aj}$  can be assumed exchangeable;  $\beta_{Aj} \sim N(B, \varphi^2)$ . Adjusting for factors that can vary across comparisons may reduce heterogeneity and improve the likelihood of transitivity. The

importance and impact of the adjustment can be judged by monitoring changes in the heterogeneity variance (compare  $\tau_r^2$  to  $\tau^2$ ) and inconsistency variance (compare  $\sigma_r^2$  to  $\sigma^2$ ), by monitoring the magnitude and significance of the coefficients  $\beta_{jk}$  and by comparing the goodness of fit and parsimony of adjusted and unadjusted models using  $DIC$  and  $\bar{D}$ .

Network meta-regression suffers from the same problems with simple meta-regression. These include ecological bias when aggregated patient-level data are used as covariates, low power with few studies and high false-positive rates if heterogeneity not explained by the covariates is ignored (Higgins and Thompson 2004).

Adjusting for bias in a network of interventions offers the advantage of increased power compared with traditional meta-analysis sensitivity analysis, because the regression coefficients share information via the consistency equations. Suppose, for example, that comparison  $B$  versus  $C$  is informed by very few studies, or by studies that all have the same characteristic (e.g., they all have poor allocation concealment). Then, conducting sensitivity analysis or adjusting the meta-analysis result of  $BC$  for allocation concealment is suboptimal or impossible. However, when these studies are part of a network meta-regression model, the bias coefficient  $\beta_{BC}$  for allocation concealment is linked to the other regression coefficients via  $\beta_{BC} = \beta_{AC} - \beta_{AB}$  and  $\beta_{A_j} \sim N(B, \varphi^2)$ .

A special application of network meta-regression is to address small study effects in a network of interventions. The association between sample size, effect size heterogeneity, and the probability of publication (which is often manifested as funnel plot asymmetry) has long been a challenging issue in meta-analysis. In a pairwise meta-analysis, the presence of small study effects (possibly due to publication bias) has been explored by regressing the underlying effect on a measure of the study precision. The same approach applies to networks of interventions to explore situations where comparisons that do not give significant results may be underrepresented or missing in the network and their relative effectiveness will be informed primarily by the indirect evidence. The covariate  $C_{ijk}$  can

be the sample standard error, variance, or inverse of sample size (references). However, significant associations between effect sizes and precision can be taken only as an indication of publication bias, as other explanations, including genuine heterogeneity, are possible. As publication bias and selective reporting will affect interventions and comparisons in different ways depending on the clinical context, the problem of selection bias in the network should be considered carefully. Further methodological development is needed to better address selection bias in network meta-analysis.

Because network meta-analysis combines studies that compare a treatment against a variety of comparators, it enables researchers to explore biases that are not identifiable in a head-to-head meta-analysis. “Optimism bias” associated with the use of novel interventions has been a concern difficult to address in a pairwise meta-analysis (Djulbegovic et al. 2011; Heres et al. 2006; Soares et al. 2005). However, in a network of interventions, the same treatment  $C$  can be the newer and hence the “favored” in a comparison  $A$  versus  $C$  but the older in another comparison  $B$  versus  $C$ . This enables us to explore apparent changes in the effectiveness of  $C$  because of optimism (Salanti et al. 2010).

#### **Application: Network Meta-regression for Incident Diabetes Using Year of Publication as Covariate**

An network meta-regression analysis of the incident diabetes data set will investigate whether differences in the publication year of included studies have an impact on the estimated treatment effects, and hence whether they can explain any of the heterogeneity and inconsistency of this network. Two meta-regression models will be used; one estimating a common fixed coefficient across all studies and all treatment comparisons and a second imposing consistency in coefficients. More specifically, in the first model, it is assumed that  $b_{ijk} = B (i = 1, \dots, 22)$ , and a vague normal prior distribution on the fixed coefficient  $\sim N(0, 10000)$  is employed. In the second model, the coefficients are assumed to be consistent  $b_{ijk} = \beta_{Ak} - \beta_{Aj}$  ( $A = P$ ) and exchangeable  $\beta_{Aj} \sim N(B, \varphi^2)$  ( $j = \{BB, D,$

**Table 12** Medians and 95% CrI of regression coefficients for comparisons of all treatments versus placebo estimated by network meta-regression model for incident diabetes with consistent and exchangeable coefficients

Comparison	$\beta$	95% CrI
$\beta$ -blockers versus placebo	-0.02	(-0.06, 0.01)
diuretics versus placebo	-0.02	(-0.06, 0.01)
CCB versus placebo	-0.03	(-0.07, 0.02)
ACE inhibitors versus placebo	-0.03	(-0.08, 0.01)
ARB versus placebo	-0.03	(-0.10, 0.03)

**Table 13** Results of network meta-regression model with a common fixed coefficient for incident diabetes using year of publication as covariate. Log-odds ratios ( $\hat{\mu}$ ) with their standard error  $SE(\hat{\mu})$  and odds ratios (OR) with their 95% credible interval (CrI) for all comparisons are reported

Comparison	$\hat{\mu}$	$SE(\hat{\mu})$	OR	95% CrI for OR
$\beta$ -blockers versus placebo	0.23	0.09	1.26	(1.06,1.50)
diuretics versus placebo	0.26	0.09	1.29	(1.07,1.56)
CCB versus placebo	0.06	0.09	1.07	(0.90,1.26)
ACE inhibitors versus placebo	-0.10	0.07	0.91	(0.78,1.05)
ARB versus placebo	-0.16	0.10	0.86	(0.70,1.04)

**Table 14** Results of network meta-regression model with consistent and exchangeable coefficients for incident diabetes using year of publication as covariate. Log-odds ratios ( $\hat{\mu}$ ) with their standard error  $SE(\hat{\mu})$  and odds ratios (OR) with their 95% credible interval (CrI) for all comparisons are reported

Comparison	$\hat{\mu}$	$SE(\hat{\mu})$	OR	95% CrI for OR
$\beta$ -blockers versus placebo	0.25	0.10	1.28	(1.06, 1.55)
diuretics versus placebo	0.30	0.10	1.34	(1.11, 1.64)
CCB versus placebo	0.05	0.10	1.05	(0.86, 1.29)
ACE inhibitors versus placebo	0.09	0.10	1.09	(0.91, 1.33)
ARB versus placebo	-0.06	0.09	0.94	(0.79, 1.12)

$CCB, ACE, ARB\}$ ), where  $B$  and  $\varphi^2$  are the mean and variance, respectively, of the distribution of all  $\beta_{Aj}$  with normal ( $B \sim N(0, 10000)$ ) and half-normal prior distributions. In both models a covariate  $(C_i - \bar{C}_i)$  is used instead of  $C_i$  (the year of publication of study  $i$ ) for computational reasons (e.g., convergence of the models), where  $\bar{C}_i$  is the mean publication year.

The estimate of the fixed regression coefficient from the first model ( $B$ ) was  $-0.01$ , corresponding to an odds ratio that is  $e^{-0.01} = 0.99$  times smaller for each 1 year later of publication. However, the 95%

CrI of  $B$  is  $(-0.03, 0.01)$  implying that there is no statistically significant effect of study publication year on treatments' effectiveness.

The same inference is derived from the second meta-regression model, which estimates the mean ( $B$ ) of distribution of regression coefficients' to be  $-0.02$   $(-0.07, 0.01)$  with variance  $\varphi^2 < 0.001$ . Table 12 shows the consistent coefficients ( $\beta_{Aj}$ ) of all treatments versus placebo.

The estimated treatment effects by the two models are presented in Tables 13 and 14.

Both meta-regression models resulted in heterogeneity estimates  $\hat{\tau}^2 = 0.02$ , the same as for the consistency hierarchical model (accounting for multi-arm trials), showing that year of publication as a covariate does not explain adequately the heterogeneity in the network.

The meta-regression with a fixed coefficient also does not improve the fit of the model ( $\bar{D} = 53.85$ ,  $DIC = 92.2$ ) compared with the hierarchical consistency model without any covariates, while the model with consistent coefficients shows a slightly better fit ( $\bar{D} = 51.57$ ,  $DIC = 91.3$ ).

The inconsistency model (as described in section “Consistency Models” but omitting the consistency equations) was also fitted with a fixed coefficient to investigate if differences in year of publication can explain the identified inconsistency. The value of the posterior deviance was  $\bar{D} = 50.43$  and  $DIC = 93.4$ , same with the inconsistency model not including any covariates. However, using the estimates of this model, the two inconsistent loops ACE-BB-P and ARB-BB-P become consistent with  $IF = 0.20$  ( $-0.44, 0.85$ ) and  $0.18$  ( $-0.53, 0.89$ ) implying that year of publication is a possible explanation of inconsistency.

## Numerical and Graphical Presentation of Results from Network Meta-analysis

Network meta-analysis involves many treatments and consequently results in a plethora of pairwise effect sizes. When presenting results from a network meta-analysis, it is useful to show both the direct and the mixed estimates along with their 95% confidence intervals and comment on any disagreements between them (e.g., as in Fig. 3). In a consistency model, all pairwise comparisons are possible and the effect sizes are often presented in the form of a “league table” or in a forest plot against a common comparator (see, e.g., Fig. 4). Presentation of the results using predictive intervals, though infrequent, best conveys the uncertainty due to heterogeneity.

Ranking measures and probabilities have become popular as they provide an understandable gateway to the results, particularly when there are many competing treatments. The probability of each treatment being the best is often calculated when the network model is fitted within the Bayesian framework. Methods are also available for similar ranking of treatments in a frequentist framework (White 2011). The probability of being the best treatment has the disadvantage that it does not reflect spread of rankings for the treatments and may thus be misleading. An obvious solution is to calculate the probabilities for all ranks. The probability of each treatment to achieve each possible rank can be plotted to yield “rankograms.” Presentation of the cumulative ranking curves in a single plot and a numerical summary of the area below the cumulative ranking curve for each treatment is useful as it gives a clear ordering of all treatments based on a summary of the rank probabilities. A review of graphical and numerical methods along with software code are presented in (Chaimani et al. 2013).

### Application: Presentation of Results for Incident Diabetes

The results of the consistency hierarchical model (accounting for multi-arm trials) will be used to illustrate the use of rankograms. The hierarchical model is fitted, and the ordering of the treatments according to their effectiveness is collected in each MCMC cycle using the equation:

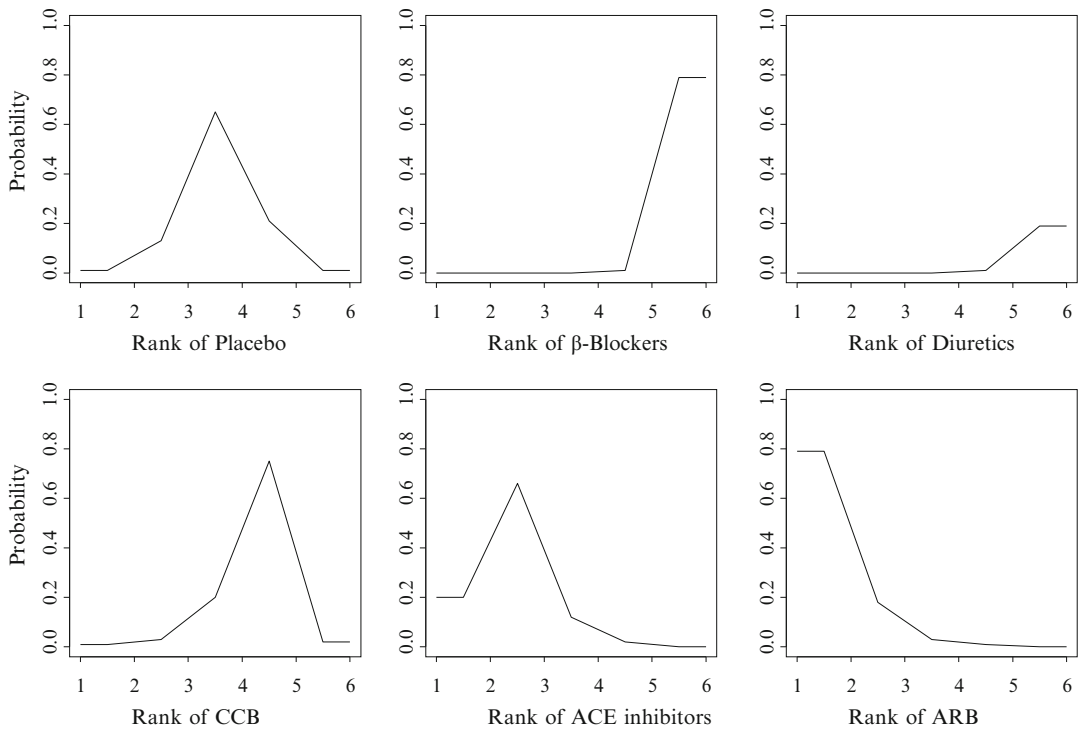
$$\text{order}_k = \sum_{j=1}^6 I(\mu_{Aj} \leq \mu_{Ak})$$

where  $I(\mu_{Aj} \leq \mu_{Ak}) = 1$  if  $\mu_{Aj} \leq \mu_{Ak}$  and 0 otherwise,  $A = P$  and  $j, k = \{BB, D, CCB, ACE, ARB\}$

Then the probability for each treatment  $k = \{P, BB, D, CCB, ACE, ARB\}$  of being the  $j$ th ( $j = 1, \dots, 6$ ) order ( $P_k^j$ ) is the ratio of MCMC simulations for which  $\text{order}_k = j$  over the total number of simulations. Table 15 includes the values of the ranking probabilities and Fig. 9 the corresponding rankograms.

**Table 15** Ranking probabilities for all treatments of incident diabetes. Results are based on the consistency hierarchical model (accounting for multi-arm trials)

Order	Placebo	β-Blockers	Diuretics	CCB	ACE inhibitors	ARB
1	0.01	0.00	0.00	0.00	0.22	0.77
2	0.07	0.00	0.00	0.02	0.71	0.20
3	0.65	0.00	0.00	0.27	0.06	0.02
4	0.27	0.01	0.01	0.70	0.01	0.00
5	0.01	0.79	0.20	0.01	0.00	0.00
6	0.00	0.20	0.80	0.00	0.00	0.00



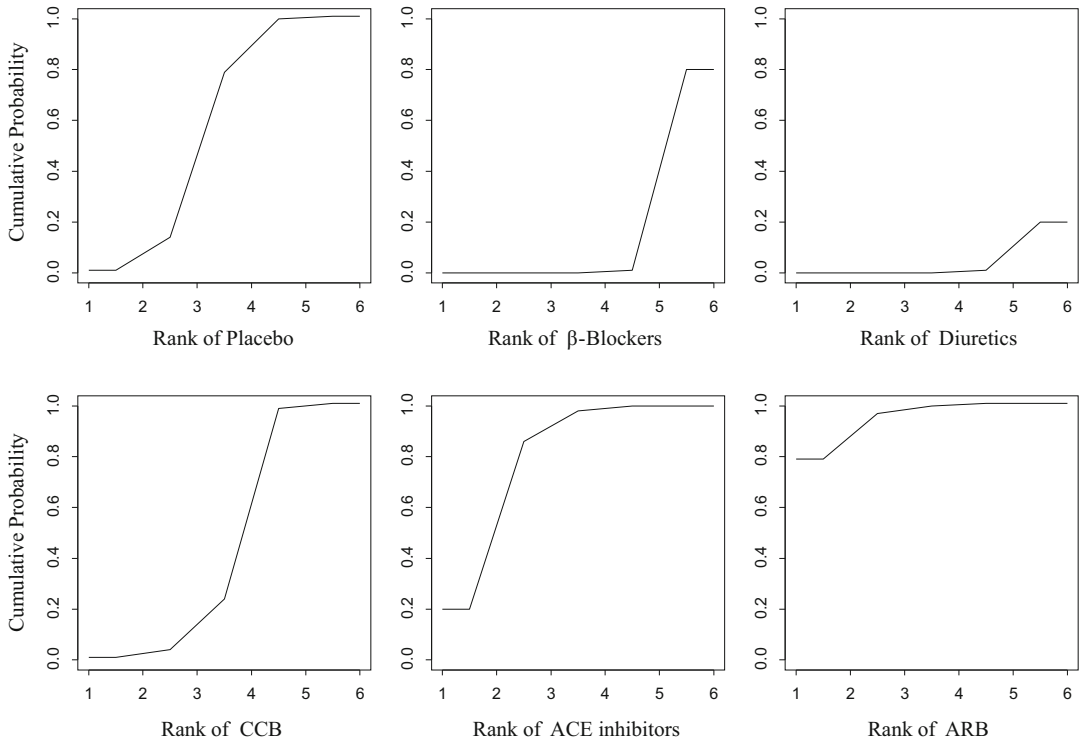
**Fig. 9** “Rankograms” for all treatments of incident diabetes. Results are based on the consistency hierarchical model (accounting for multi-arm trials)

**Table 16** Numerical summary of area below the cumulative raking curve for all treatments of incident diabetes. Results are based on the consistency hierarchical model (accounting for multi-arm trials)

Placebo	0.59
β-Blockers	0.16
Diuretics	0.04
CCB	0.46
ACE inhibitors	0.81
ARB	0.96

The numerical summary of area below the cumulative raking curve for each treatment  $k$  is calculated as  $\left(\sum_{j=1}^5 \text{cum. } P_k^j\right)/5$ .

The results are presented in Table 16 and the plots in Fig. 10. These results suggest that the best treatment appears to be ARB followed by ACE inhibitors, placebo, CCB, β-blockers, and last diuretics.



**Fig. 10** Plot of area below the cumulative raking curve for all treatments of incident diabetes. Results are based on the consistency hierarchical model (accounting for multi-arm trials)

**Acknowledgments** GS and AC received funding from the European Research Council (ERC starting grant IMMA 260559). DC is supported by an UK MRC Population Health Scientist Fellowship (G0902118). JPTH is funded by Medical Research Council grant U105285807.

## References

- Baker SG, Kramer BS. The transitive fallacy for randomized trials: if A bests B and B bests C in separate trials, is A better than C? *BMC Med Res Methodol.* 2002;2:13.
- Barbui C, Cipriani A, Furukawa TA, et al. Making the best use of available evidence: the case of new generation antidepressants: a response to: are all antidepressants equal? *Evid Based Ment Health.* 2009;12:101–4.
- Bucher HC, Guyatt GH, Griffith EL, et al. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol.* 1997;50(6):683–91.
- Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ.* 2005;331:897–900.
- Caldwell DM, Welton NJ, Ades AE. Mixed treatment comparison analysis provides internally coherent treatment effect estimates based on overviews of reviews and can reveal inconsistency. *J Clin Epidemiol.* 2010;6(8):875–82.
- Chaimani A, Higgins JP, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. *PLoS One.* 2013;8(10):e76654.
- Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet.* 2009;373:746–58.
- Cooper NJ, Sutton AJ, Morris D, et al. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Stat Med.* 2009;28(14):1861–81.
- Cooper NJ, Peters J, Lai MC, et al. How valuable are multiple treatment comparison methods in evidence-based health-care evaluation? *Value Health.* 2011;14:371–80.
- Dias S, Welton NJ, Caldwell DM, et al. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med.* 2010;29:932–44.
- Djulgovic B, Kumar A, Magazin A, et al. Optimism bias leads to inconclusive results—an empirical study. *J Clin Epidemiol.* 2011;64:583–93.



- Donegan S, Williamson P, Gamble C, et al. Indirect comparisons: a review of reporting and methodological quality. *PLoS One*. 2010;5:e11054.
- Edwards SJ, Clarke MJ, Wordsworth S, et al. Indirect comparisons of treatments based on systematic reviews of randomised controlled trials. *Int J Clin Pract*. 2009;63:841–54.
- Eli Lilly and Company. Gemcitabine for the treatment of metastatic breast cancer: Single technology appraisal submission to the National Institute for health and Clinical Excellence. 2006. Available from <http://www.nice.org.uk>
- Elliott WJ, Meyer PM. Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. *Lancet*. 2007;369:201–7.
- Glenny AM, Altman DG, Song F, et al. Indirect comparisons of competing interventions. *Health Technol Assess*. 2005;9:26.
- Guyatt GH, Sackett DL, Sinclair JC, et al. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. *JAMA*. 1995;274:1800–4.
- Heres S, Davis J, Maino K, et al. Why olanzapine beats risperidone, risperidone beats quetiapine, and quetiapine beats olanzapine: an exploratory analysis of head-to-head comparison studies of second-generation antipsychotics. *Am J Psychiatry*. 2006;163:185–94.
- Higgins JPT, Green S. *Cochrane handbook for systematic reviews of interventions*. 5.0.1 ed. The Cochrane Collaboration; 2008; John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, England.
- Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression. *Stat Med*. 2004;23:1663–82.
- Hoaglin DC, Hawkins N, Jansen JP, et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR task force on indirect treatment comparisons good research practices-part 2. *Value Health*. 2011;14:429–37.
- Hughes S. First "comparison" of prasugrel and ticagrelor. 2010 Sep16. Available from <http://www.theheart.org/article/1122713.do>. Accessed 27 Apr 2011.
- Jackson D, Riley R, White IR. Multivariate meta-analysis: potential and promise. *Stat Med*. 2011;30:2481–98.
- Jansen JP, Schmid CH, Salanti G. Directed acyclic graphs can help understand bias in indirect and mixed treatment comparisons. *J Clin Epidemiol*. 2012;65:798–807.
- Jones A, Takeda A, Tan SC, Cooper K, Loveman E, Clegg A, Murray N. Gemcitabine for metastatic breast cancer: evidence review group report. 2006. Available from [www.nice.org.uk](http://www.nice.org.uk)
- Lambert PC, Sutton AJ, Burton PR, Abrams KR, et al. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat Med*. 2005;24:2401–28.
- Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med*. 2004;23(20):3105–24. PMID: 15449338"
- Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *J Am Stat Assoc*. 2006;101:447–59.
- Lu G, Ades AE. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics*. 2009;10(4):792–805.
- McAlister FA, Laupacis A, Wells GA, et al. Users' guides to the medical literature: XIX. Applying clinical trial results B. Guidelines for determining whether a drug is exerting (more than) a class effect. *JAMA*. 1999;282:1371–7.
- Mills EJ, Ghement I, O'Regan C, et al. Estimating the power of indirect comparisons: a simulation study. *PLoS One*. 2011;6:e16237.
- NICE. *Methods for the development of NICE public health guidance*. 2nd ed. Evidence Synthesis National Institute of Health and Clinical Excellence; 2008.
- O'Regan C, Ghement I, Eyawo O, et al. Incorporating multiple interventions in meta-analysis: an evaluation of the mixed treatment comparison with the adjusted indirect comparison. *Trials*. 2009;10:86.
- PBAC. Report of the indirect comparisons working group to the pharmaceutical benefits advisory committee: assessing indirect comparisons. Pharmaceutical Benefits Advisory Committee; 2008. [http://www.health.gov.au/internet/main/publishing.nsf/Content/B11E8EF19B358E39CA25754B000A9C07/\\$File/ICWG%20Report%20FINAL2.pdf](http://www.health.gov.au/internet/main/publishing.nsf/Content/B11E8EF19B358E39CA25754B000A9C07/$File/ICWG%20Report%20FINAL2.pdf)
- Piccini JP, Kong DF. Mixed treatment comparisons for atrial fibrillation: evidence network or bewildering entanglement? *Europace*. 2011;13:295–6.
- Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. *J R Stat Soc Ser A*. 2009;172:789–811.
- Salanti G, Marinho V, Higgins JP. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. *J Clin Epidemiol*. 2009;62:857–64.
- Salanti G, Dias S, Welton NJ, et al. Evaluating novel agent effects in multiple-treatments meta-regression. *Stat Med*. 2010;29:2369–83.
- Soares HP, Kumar A, Daniels S, et al. Evaluation of new treatments in radiation oncology: are they better than standard treatments? *JAMA*. 2005;293:970–8.
- Song F, Altman D, Glenny AM, et al. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ*. 2003;326:472.
- Song F, Loke YK, Walsh T, et al. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *BMJ*. 2009;338:b1147.
- Song F, Xiong T, Parekh-Bhurke S, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. *BMJ*. 2011;343:d4909.
- Spiegelhalter DJ, Best NG, Bradley PC, et al. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B*. 2002;64:583–639.

- Spiegelhalter DJ, Abrams KR, Myles PJ. Bayesian approaches to clinical trials and health-care evaluation. Chichester: Wiley; 2004.
- Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res.* 2001;10:277–303.
- Thijs V, Lemmens R, Fieuws S. Network meta-analysis: simultaneous meta-analysis of common antiplatelet regimens after transient ischaemic attack or stroke. *Eur Heart J.* 2008;29:1086–92.
- Uhtman OA, Abdulmalik J. Comparative efficacy and acceptability of pharmacotherapeutic agents for anxiety disorders in children and adolescents: a mixed treatment comparison meta-analysis. *Cur Med Res Opin.* 2010;26(1):53–9.
- Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat Med.* 2007;26:37–52.
- Warn DE, Thompson SG, Spiegelhalter DJ. Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. *Stat Med.* 2002;21:1601–23.
- Wells GA, Sultan SA, Chen L, et al. Indirect evidence: indirect treatment comparisons in meta-analysis. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2009.
- White IR. Multivariate random-effects meta-regression: updates to mvmeta. *Stata J.* 2011;11(2):255–70.



# Introduction to Social Network Analysis

# 26

Alistair James O'Malley and Jukka-Pekka Onnela

## Contents

<b>Part I: Introduction and Background</b> .....	618
Historical Note .....	619
<b>Representation of Networks</b> .....	620
Network Data .....	620
Representation of Network Data .....	621
<b>Descriptive Measures</b> .....	623
Unipartite or One-Mode Networks .....	623
Bipartite or Two-Mode Networks .....	625
<b>Part II: Statistical Models</b> .....	627
Network Influence Models .....	627
Relational Analyses .....	631
<b>Part III: Network Science</b> .....	638
Generative Models of Network Formation .....	639
Network Communities .....	644
<b>Part IV: Discussion and Glossary</b> .....	649
<b>Glossary of Terms</b> .....	650
Terms Used in Social Networks .....	651
Terms Used in Network Science .....	655
<b>References</b> .....	657

A. J. O'Malley (✉)

The Dartmouth Institute for Health Policy and Clinical Practice, Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

Department of Health Care Policy, Harvard Medical School, Boston, MA, USA  
e-mail: [James.OMalley@Dartmouth.edu](mailto:James.OMalley@Dartmouth.edu)

J.-P. Onnela

Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA  
e-mail: [onnela@biostat.hsph.edu](mailto:onnela@biostat.hsph.edu);  
[onnela@hsph.harvard.edu](mailto:onnela@hsph.harvard.edu)

## Abstract

This chapter introduces statistical methods used in the analysis of social networks and in the rapidly evolving parallel-field of network science. Although several instances of social network analysis in health services research have appeared recently, the majority involve only the most basic methods and thus scratch the surface of what might be accomplished. Cutting-edge methods using

relevant examples and illustrations in health services research are provided.

---

## Part I: Introduction and Background

Social network analysis is the study of the structure of relationships linking individuals (or other social units, such as organizations) and of interdependencies in behavior or attitudes related to configurations of social relations. The observational units in a social network are the relationships between individuals and their attributes. Whereas studies in medicine typically involve individuals whose observations can be thought of as statistically independent, observations made on social networks may be simultaneously dependent on all other observations due to the social ties and pathways linking them. Accordingly, different statistical techniques are needed to analyze social network data. The focus of this chapter is *socio-centric data*, the case when relational data is available for all pairs of individuals, allowing a fully-fledged review of available methods.

Two major questions in social network analysis are: (1) do behavioral and other mutable traits spread from person-to-person through a process of induction (also known as *social influence*, *peer effects*, or *social contagion*); (2) what exogeneous factors (e.g., shared actor traits) or endogeneous factors (e.g., internal configurations of actors such as triads) are important to the overall structure of relationships among a group of individuals.

The first problem has affinity to medical studies in that individuals are the observational units. In medicine, the health of an individual is paramount and so individual outcomes have historically been used to judge the effectiveness of an intervention. A study of social influence in medicine may involve the same outcome, but the treatment or intervention is the same variable evaluated on the peers of the focal individual (referred to as *alters*). An important characteristic of studies of social influence is that individuals may partly or fully share treatments and one individual's treatment may depend on the outcome of another. For example, an intervention that encourages person A to exercise in order to lose weight

might also influence the weight of A's friends (B and C) because they exercise more when around A. Hence, A's weight intervention may also affect the weight of B and C. A consequence is that the total effect of A's treatment must also consider its effect on B and C, the benefit to individuals to whom B and C are connected, and so on. Such *interference* between observations violates the stable-unit treatment value assumption (SUTVA) that one individual's treatment not affect another's outcome (Rubin 1978), which presents challenges for identification of causal effects. Interference is likely to result in an incongruity between a regression parameter and the causal effect that would be estimated in the absence of interference.

The second problem is important in sociology as social networks are thought to reveal the structure of a group, organization, or society as a whole (Freeman 2004). For example, there has always been great interest in determining whether the triad is an important social unit (Simmel 1908; Heider 1946). If the existence of network ties A-B and A-C makes the presence of network tie B-C more likely than the network exhibits *transitivity*, commonly described as "a friend of a friend is a friend." Thus, just as an individual may influence or be influenced by multiple others, the relationship status of one dyad (pair of individuals) may affect the relationship status of another dyad, even if no individuals are common to multiple dyads. Accounting for between dyad dependence is a core component of many social network analyses and has entailed much methodological research.

Network science is a parallel field to social network analysis in that there is very little overlap between researchers in the respective fields despite the similarity of the problems. Whereas solutions to problems in social networks have tended to be data-oriented in that models and statistical tests are based on the data, those in network science have tended to be phenomenon-oriented with analogies to problems in the physical sciences often providing the backbone for solutions. Methods for social network analysis often have causal hypotheses (e.g., does one individual have an effect on another, does the presence of a common friend make friendship

formation more likely) motivating them and involving microlevel modeling. In contrast, methods in network science seek models generated from some theoretical basis that reproduce the network at a global or system level and in so doing reveal features of the data-generating process (e.g., is the network scale-free, does the degree-distribution follow a power-law). One of the goals of this chapter is to address the lack of interaction between the social network and network science fields by providing the first joint review of both. By enlarging the range of methods at the disposal of researchers, advances at the frontier of networks and health will hopefully accelerate.

The computer age has enabled widespread implementation of methods for social network and network science analysis, particularly statistical models. At the same time, a diverse range of applications of social network analysis have appeared, including in medicine (Keating et al. 2007; Pham et al. 2009; Barnett et al. 2012a; Iwashyna et al. 2002; Pollack et al. 2012). Because many medical and health-related phenomena involve interdependent actors (e.g. patients, nurses, physicians, and hospitals), there is enormous potential for social network analysis to advance health services research (O'Malley and Marsden 2008).

The layout of the remainder of the chapter is as follows. This introductory section concludes with a brief historical account of social networks and network science is given. The major types of networks and methods for representing networks are then discussed (section “[Representation of Networks](#)”). In section “[Descriptive Measures](#)” formal notation is introduced and descriptive measures for networks are reviewed. Social influence and social selection are studied in sections “[Network Influence Models](#)” and “[Relational Analyses](#)”, respectively. Our focus switches to methods akin with network science in section “[Generative Models of Network Formation](#)”, where descriptive methods are discussed. The review of network science methods continues with community detection methods in section “[Network Communities](#)”. The chapter concludes in “[Discussion and Glossary](#)”.

## Historical Note

In the 1930s, a field of study involving human interactions and relationships emerged simultaneously from sociology, psychology, and anthropology. Moreno is credited for inventing the sociogram (Moreno 1934), a visual display of social structure. The appeal of the sociogram led to Moreno being considered a founder of sociometry, a precursor to the field known as *social networks*. A number of mathematical analyses of network-valued random variables in the form of sociograms followed (Festinger 1949; Katz 1947, 1953; Katz and Powell 1955). Other important contributions were to structural balance (Heider 1946; Newcomb 1953; Cartwright and Harrary 1956), the diffusion of medical innovations (Coleman et al. 1957, 1966), structural equivalence (Lorrain and White 1971), and social influence (Marsden and Friedkin 1993). Refer to Wasserman and Faust (1994, chapter 1) for a detailed historical account.

Early network studies involved small networks with defined boundaries such as students in a classroom, or a few large entities such as countries engaging in international trade. Because the typical number of individuals in such studies was small (e.g.,  $\leq 100$ ), relationships could be determined for all possible pairs of individuals yielding complete *sociocentric* datasets. Furthermore, the often enclosed nature of the system (e.g., a classroom or commune) reduced the risk of confounding by external factors (e.g., unobserved actors).

Sociological theory developed over time as sociologists provided intuitive reasoning to support various hypotheses involving social networks and society (Freeman 2004). In the specific area of individual health, at least five principal mediating pathways through which social relationships and thus social networks may influence outcomes have been posited (Berkman and Glass 2000). Prominent among these is social support, which has emotional, instrumental, appraisal (assistance in decision making), and informational aspects (House and Kahn 1985). Beyond social support, networks may also offer access to tangible resources such as financial assistance or transportation.

They can also convey social influence by defining norms about such health-related behaviors as smoking or diet, or via social controls promoting (for example) adherence to medication regimes (Marsden 2006). Networks are also channels through which certain communicable diseases, notably sexually transmitted ones, spread (Klovdahl 1985) and certain network structures have been hypothesized to reduce exposure to stressors (Haines and Hurlbert 1992).

A field known as mathematical sociology complemented social theory by attempting to derive results using mathematical rather than intuitive arguments. In particular, statistical and probability methods are used to test for the presence of various structural features in the network. Other key areas of mathematics that have been used in network analysis include graph theory and algebraic models. Katz and Powell (1955) develop tests of dependence within dyads (pairs of actors) while Harary (1953) and Harary (1955) develop tests of triadic dependence. In general, results were descriptive or based on simple models making strong assumptions about the network. With the advent of powerful computers, mathematical contributions have taken on more importance as so much more can be implemented than in the past. For example, computer simulation has recently been used to test and develop theoretical results (Centola 2009).

In the mid-late 1990s, network science emerged as a discipline. Whereas social networks were the domain of social scientists and a growing number of statisticians, network scientists typically have backgrounds in physics, computer science, or applied mathematics. The use of physical concepts to generate solutions to problems is common as evinced by the large domains of research focusing on the adaptation of (e.g.,) a particular physical equation to network data. For example, several procedures for partitioning a network into disjoint groups of individuals ("communities") rely on the *modularity* equation, which was developed in the context of spin-theory to model the interaction of electrons. While much of the initial work focused on the properties of the solution at different values of the parameters, there recently has been increased attention to using these

methods to provide valuable insight on important practical problems.

---

## Representation of Networks

Social networks are comprised of units and the relationships between them. The units are often individuals (also referred to as *actors*) but can include larger (e.g., countries, companies) and smaller (e.g., organisms, genes) entities.

## Network Data

In sociocentric studies, data is assembled on the ties linking all units or actors within some bounded social collective (Laumann et al. 1983). For example, the collection of data on the network of all children in a classroom or on all pairs of physician collaborations within a medical practice constitutes a sociocentric study. Relationships can be shared or directional, and quantified by binary (tie exists or not), scale (or valued), or multivariate variables. By measuring all relationships, sociocentric data constitutes the highest level of information collection and facilitates an extensive range of analyses including accounting for the effects of multiple actors on actor outcomes or the structure of the network itself to be studied (O'Malley and Marsden 2008). A weaker form of relational data is collected in egocentric studies where individuals ("egos") are sampled at random and information is collected on at least a sample of the individuals with direct ties to the egos ("alters"). Because standard statistical methods such as regression analysis can generally be used to analyze egocentric data (O'Malley et al. 2012), herein egocentric data are not featured.

Relational data is often binary (e.g., friend or nonfriend). One reason is that other types of relational data (e.g., nominal, ordinal, interval-valued) are often transformed to binary due to the convenience of displaying binary networks. Another is the greater range of models available for modeling binary data.

Many studies involve two distinct types of units, such as patients and physicians, or physicians and hospitals, authors and journal articles or books, etc. In these two-mode networks, the elementary relationships of interest usually refer to affiliations of units in one set with those in the other, e.g., of patients with the physician(s) responsible for their care, or of physicians with the hospital(s) at which they are admitted to practice. Two-mode networks are also known as affiliation or *bipartite* networks. They can be viewed as a special case of general sociocentric network data in that the relationship of interest is between heterogeneous types of actors.

The advent of high-powered computers has enabled the analysis of large networks, which has benefitted fields such as health services research that regularly encounter large data sets. A challenge facing analyses of large networks is that it may be infeasible for all actors to be exposed to each other actor and thus for a relationship to have formed. Therefore, statistical analyses for large networks essentially use relational data representing the joint event of individuals meeting and then forming a tie, not the network of ties that would be observed if all pairs of individuals actually met. Accordingly, analyses of large networks may underestimate effect sizes unless information on the likelihood of two individuals meeting is incorporated.

## Representation of Network Data

Let the status of the relationship from  $i$  to  $j$  be denoted by  $a_{ij}$ , element  $ij$  of the adjacency matrix  $A$ . In a directed network  $a_{ij}$  may differ from  $a_{ji}$  while in a nondirected network  $a_{ij} = a_{ji}$ , implying  $A = A^T$ . A network constructed from friendship nominations is likely to be directed while a network of coworkers is nondirected. In the case of immutable relationships (e.g., siblings),  $A$  will only change as actors are added or removed (e.g., through birth or death), as relationship status is otherwise invariant. In the following, assume the network is binary unless otherwise stated (Fig. 1).

Matrices and graphs are two common ways of representing the status of a network at a fixed time. In a matrix representation, rows and columns correspond to units or actors; the matrix is

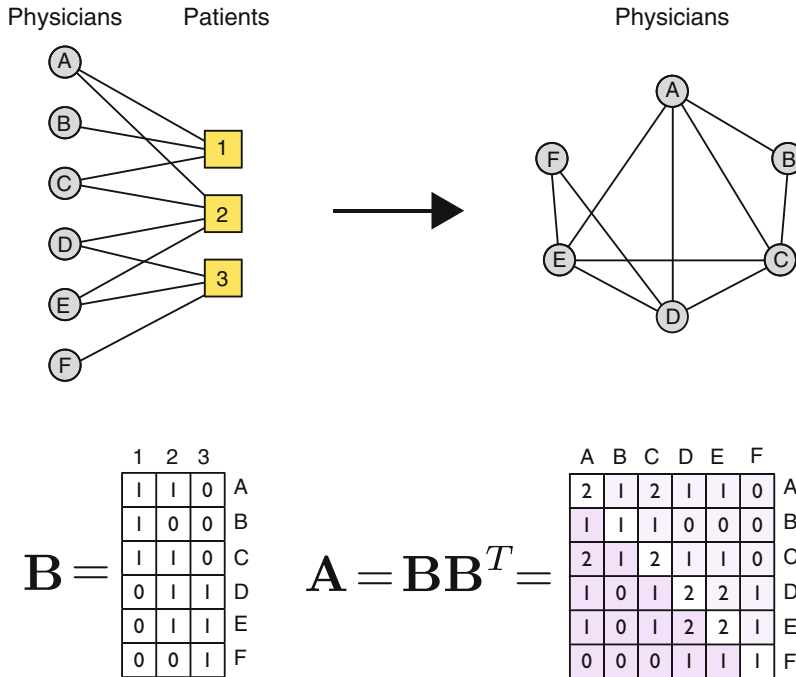
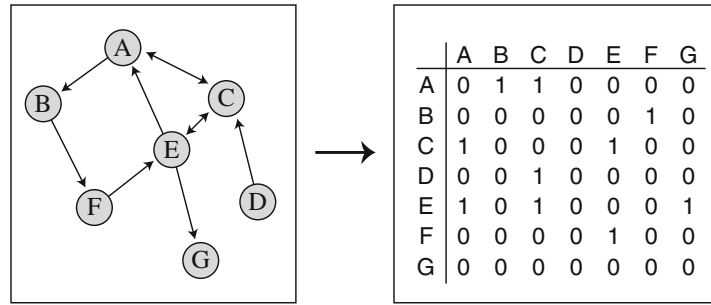
square for one-mode and rectangular for two-mode networks. Elements of the matrix contain the value of the relationship linking the corresponding units or actors, so that element  $ij$  represents the relationship from actor  $i$  to actor  $j$ . With binary ties (1 = tie present, 0 = tie absent), the matrix representation is known as an adjacency matrix. Irrespective of how the network is valued, the diagonal elements of the matrix representing the network equal 0 as self-ties are not permitted. Several network properties can be computed through matrix operations.

In graphical form, units or actors are vertices and nonnull relationships are lines. Nondirected relationships are known as “edges” and directed ones as “arcs”; arrows at the end(s) of arcs denote their directionality. Value-weighted graphs can be constructed by displaying nonnull tie values along arcs or edges, or by letting thinner and thicker lines represent line values. Such graphical imagery is a hallmark of social network analysis (Freeman 2004).

Two-mode (or bipartite) networks may be represented in set-theoretic form as hypergraphs consisting of a set of actors of one type, together with a collection of subsets of the actors defined on the basis of a common actor of the second type (Wasserman and Faust 1994). This representation highlights the multiparty relationships that may exist among those actors of one type that are linked to a given actor of the other type, e.g., the set of all physicians affiliated with a particular clinic or service. In matrix form, element  $ij$  of an affiliation matrix  $A$  indicates that actor  $i$  of the first type is linked to actor  $j$  of the second type. Affiliation networks may usefully be represented as bipartite graphs in which nodes are partitioned into two disjoint subsets and all lines link nodes in different sets.

An induced one-mode network  $A$  may be obtained by multiplying an affiliation matrix  $B$  by its transpose,  $A = BB^T$ ; entry  $ij$  of the outer-product  $BB^T$  gives the number of affiliations shared by a pair of actors of one type (see Fig. 2, which emulates a figure in Landon et al. (2012)). Dually, the inner-product  $B^T B$  yields a one-mode network of shared affiliations among actors of the second type (Breiger 1974). The diagonals of the

**Fig. 1** Graphical and matrix representation of a social network. Digraph (left) and adjacency matrix (right), which is denoted in the text as  $A$ . Note: Self-ties are not relevant in studies involving relationships



**Fig. 2** A schematic illustrating a projection from a two-mode (bipartite) to a one-mode (unipartite) network. For example, Medicare records link each doctor to a number of patients, defining a bipartite network consisting of two types of nodes, doctors and patients. An edge can only exist between different types of nodes (a doctor and a patient), and the network is fully described by the (in this case  $6 \times 3$ ) bipartite adjacency matrix  $B$ . A one-mode

projection of the doctor-patient network is obtained by multiplying the bipartite adjacency matrix  $B$  by its transpose,  $B^T$ , to yield a  $6 \times 6$  symmetric one-mode adjacency matrix  $A$ , whose elements indicate the number of patients the two physicians have in common. The diagonal elements of  $A$  correspond to the number of patients the given physician “shares with themselves” (i.e., the number of patients they care for)

outer and inner matrix products give the degree of the actors (i.e., the number of ties to actors of the other mode).

In health services applications, an investigator is often interested in a one-mode network that is not directly observed but rather is induced from a two-mode network. Such one-mode projection networks are motivated theoretically by a claim

that shared actors from the other mode act as surrogates for ties between the actors. For example, physicians with many patients in common might have heightened opportunities for contact through consultations or sharing of information about those patients, and thus the number of shared patients is a surrogate for the actual extent of interaction between pairs of physicians.



Examples of provider (physician, hospital, health service area) networks obtained as one-mode projections of bipartite networks in health services research are given in (Barnett et al. 2011, 2012a, b; Pham et al. 2009).

An often overlooked feature of bipartite network analysis is the mechanism by which network data is obtained. Networks obtained from one-mode projections have different statistical properties from directly observed one-mode networks. Consider a patient-physician bipartite network and suppose a threshold is applied to the physician one-mode projection such that true social ties are assumed to exist or not according to whether one or more patients are shared. Then a patient that visits three physicians induces ties between all three physicians. The same complete set of ties between the three physicians is also induced by three patients that each visit different pairs of the three physicians. However, the projection does not preserve the distinction (see section “[Bipartite or Two-Mode Networks](#)” for further comment).

---

## Descriptive Measures

### Unipartite or One-Mode Networks

The number of units or actors ( $N$ ) is known as the order of the network. A common network statistic is network density ( $D$ ), defined as the number of ties across the network ( $L$ ) divided by the number of possible ties; for directed networks  $D = L / (N(N - 1))$  and for nondirected networks  $D = L / (2N(N - 1))$ . Thus, density equals the mean value of the binary (1, 0) ties across the network. The same definition can be used for general relational data, in which case the resulting measure is sometimes referred to as *strength*. While results in this chapter are generally presented for binary networks, corresponding measures for weighted networks often exist (Opsahl et al. 2010).

The tendency for relationships to form between people having similar attributes is known as homophily (McPherson et al. 2001). Homophily

involves subgroup-specific network density statistics. With high homophily according to some attribute, networks tend toward segregation by that attribute – the extreme case occurs when the network consists of separate components (i.e., no ties between actors in different components) defined by levels of the attribute. In the other direction, one obtains a bipartite network where all ties are between different types of actors (extreme heterophily).

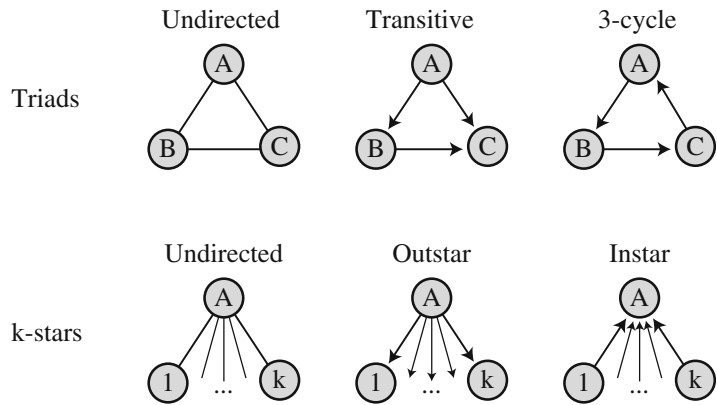
The out- and in-degree for an actor  $i$  are the number of ties from,  $a_{i+} = \sum_{j=1}^N a_{ij}$  (column sum), and to,  $a_{+j} = \sum_{i=1}^N a_{ij}$  (row sum), actor  $i$ . These are also referred to as *expansiveness* and *popularity*, respectively. For example, a positive correlation between out- and in-degree suggests that popular individuals are expansive.

The number of ties (or value of the ties) in a network is given by  $L = N\bar{d}$ , where  $\bar{d}$  denotes the mean degree (or strength) of an individual, implying the density of the network is given by  $D = \bar{d} / (N - 1)$ . This result is not specific to in- or out-degree due to the fact that the total number of inward ties must equal the total number of outward ties, implying mean in-degree equals mean out-degree.

The variance of the degree distribution measures the extent to which tie-density (or connectedness) varies across the network (Snijders 1981). Often actors having higher degree have prominent roles in the network (Freeman 1979). A special type of homophily is the phenomenon where individuals form ties with individuals of similar degree, commonly referred to as *assortative mixing*. In directed networks, assortative mixing can be defined with respect to both out-degree and in-degree (Piraveenan et al. 2010). The opposite scenario to a network with the same degree for all actors is a  $k$ -star – a network configuration with  $k$  relationships are incident to the focal actor (Fig. 3) – in which there are no ties between the other actors.

The length of a path between two actors through the network is defined as the number of ties traversed to get from one actor to the other. The elements of the adjacency matrix multiplied by itself  $k - 1$  times, denoted  $A^k$ , equal the number

**Fig. 3** Triadic and k-star configurations



of paths of length  $k$  between any two actors with the number of  $k$ -cycles (including multiple or repeated loops) on the diagonal. The shortest path between two actors is referred to as the *geodesic distance*.

### Clustering

Certain subnetworks have particular theoretical prominence. The first step-up from the trivial single actor subnetwork, also known as an isolated node, is the network comprising two actors (a “dyad”). The presence and magnitude of a tendency toward symmetry or reciprocity in a directed network can be measured by comparing the number of mutual dyads (ties in both directions) to the number expected under a null model that does not accommodate reciprocity. If the number of mutual dyads is higher than expected, there is a tendency towards reciprocation.

A *triad* is formed by a group of three actors. Figure 3 shows a “transitive triad,” so-named as it exhibits the phenomenon that a “friend of a friend is a friend.” Nonparametric tests for the presence of transitivity or other forms of triadic dependence are based on the distribution of the number of closed and nonclosed triads conditional on the number of null (no ties intact), directed (one tie intact), and mutual dyads (both ties intact) collectively known as the *dyad census*; the degree distribution; and other lower-order effects (e.g., homophily of relevant individual characteristics) in the observed network. Such tests are described in Wasserman and Faust (1994, chapter 14).

### Centrality

Centrality is the most common metric of an actor’s prominence in the network and many distinct measures exist. They are often taken as indicators of an actor’s network-based “structural power.” Such measures are often used as explanatory variables in individual-level regression models (Barnett et al. 2012a).

Different centrality measures are characterized by the aspects of an actor’s position in the network that they reflect. For example, degree-based centrality – the degree of an actor in an undirected network and in- or out-degree in a directed network – reflects an actor’s level of network connectivity or involvement in the network. Betweenness centrality computes the frequency with which an actor is found in an intermediary position along the geodesic paths linking pairs of other actors. Actors with high betweenness centrality have high capacity to broker or control relationships among other actors. A third major centrality measure, closeness centrality, is inversely proportional to the sum of geodesic distances from a given actor to all others. The rationale underlying closeness measures is that actors linked to others via short geodesics have comparatively little need for intermediary units, and hence have relative independence in managing their relationships. Closeness measures are defined only for networks in which all actors are mutually related to one another by paths of finite geodesic distance, i.e., single component networks. Finally, eigenvalue centrality is sensitive to the presence or strength of connections, as well

as those of the actors to which an actor is linked (Bonacich 1987). It assumes that connections to central actors indicate greater prominence than do (similar-strength) connections to peripheral actors. The key component of the measure is the largest eigenvalue of an adjacency or other matrix representation of the network (Bonacich 1987).

Network-level centrality indices (Freeman 1979) are network-level statistics that resemble the degree variance whose values grow larger to the extent that a single actor is involved in all relationships (as in the “star” network shown in Fig. 3).

### Cliques, Components, and Communities

The assignment of actors to groups is an important and growing field within social networks. The rationale for grouping actors is that it may reveal salient social distinctions that are not directly observed. The general statistical principle adhered to is that individuals within a group are more alike than individuals in different groups. Groups are typically formed on the basis of network ties alone, the rationale being that the similarity of individuals’ positions in the network is in-part revealed by the pattern of ties involving them. Thus, actors in densely connected parts of the network are likely to be grouped together. A related concept to a group is a clique, a maximal subset of actors having density 1.0 (i.e., ties exist between all pairs of individuals in a binary network). The larger the clique the stronger the evidence that the collective individuals are in the same group. Grouping algorithms based on maximizing the ratio of within-group to between-group ties are unlikely to split large cliques as doing so creates a lot of between-group ties. However, a clique need not be its own group.

Components of a network are defined by the nonexistence of any paths between the actors in them. Often a network is comprised of one large component and several small components containing few individuals. A more practical way of grouping individuals than by cliques is through  $k$ -connected components (White and Harary 2001), a maximal subset of actors mutually linked to one another by at least  $k$  node-independent paths (i.e., paths that involve disjoint

sets of intermediary actors who also lie within the subgraph). Such a criterion is related to  $k$ -coreness, a measure of the extent to which subgraphs with all internal degrees  $\geq k$  occur (Seidman 1983) in a network.

There are several other ways for grouping the actors in a network. Model-based methods include mixed-membership stochastic block models (Airoldi et al. 2008) and latent-class models in which the group is treated as a categorical individual-level latent variable (Handcock et al. 2007) while nonparametric methods used in network science include *modularity* and its variants. These methods are discussed in section “Network Communities”, where the grouping of actors is referred to as *community detection*.

### Bipartite or Two-Mode Networks

In practice two-mode networks are rarely directly analyzed. If one of the modes instigates ties or is of primary interest, the network involving just those actors is often analyzed as a single-mode network. For example, in a physician-patient referral network, the physicians often instigate ties through patient referrals while patients are chiefly responsible for who they see first. The projection from a two-mode network to a one-mode network links nodes in one mode (e.g., physicians) if they share a node of the other mode (e.g., patients). A weighted network can be formed with the number of shared actors of the other mode (or function thereof) as weights.

In describing networks obtained from a projection of a two-mode network, the usual practice is to use unipartite descriptive measures. However, several layers of information are lost, including the number of actors in the other mode underlying a tie and the degree distribution of the actors in the other mode, from treating a one-mode projection as an actual network. Even if the two-mode network is completely random, ties in a one-mode projection that arise from a single (e.g.,) patient with ties to (e.g.,) three physicians are not separate events. More generally, a patient who visits  $k$ -physicians generates a  $k$ -clique among those physicians and tells us nothing about whether

physician sharing of one patient is correlated with physician sharing of another patient – the question of primary interest in the study of the diffusion of treatment practices. Thus,  $k$ -cliques for  $k > 2$  may be excluded from measures of transitivity in two-mode networks.

Descriptive measures for two-mode networks may be computed that parallel those for one-mode networks (Wasserman and Faust 1994). Centrality measures based on the bipartite network representation are covered in Faust (1997). Borgatti and Everett (1997) review visualization, subgroup detection, and measurement of centrality for two-mode network data. More descriptive measures for two-mode networks have recently been proposed. For example, a two-mode measure of transitivity defined as the ratio of the total number of six cycles (closed paths of six ties through six nodes) in the two-mode network divided by the total number of open five-paths through six nodes (Opsahl 2011). In the context of the patient-physician network, *physician transitivity* exists if physicians A and B sharing a patient and physicians B and C sharing a patient makes it more likely for physicians A and C to share a patient. It is only if the two pairs of physicians have different patients in common that the physician triad may be transitive and only if the third pair share a different patient from the first two that the event can be attributed to transitivity. The involvement of distinct patients makes the physician-physician ties distinct events and thus informative about clustering of physicians (and patients).

In general, the matrix equation  $A = BB^T$  in which a bipartite network adjacency matrix  $B$  is multiplied by its transpose yields a weighted one-mode network (the elements contain the number of shared actors of the other mode). To avoid losing information about the number of actors leading to a tie between primary nodes, weights can be retained or monotonically transformed in the projected network. Weighted analogies of descriptive measures of binary networks can be evaluated on the weighted one-mode projection. For example, the calculation of degree is emulated by summing the weights of the edges involving an

individual, yielding their *strength*. Degree and strength together distinguish between actors with many weak ties and those with a few strong ties. Analogous measures of centrality can also be computed for the weighted one-mode projection (Opsahl et al. 2010). However, whether ties between  $k$  physicians arise through them all treating the same patient, from each pair of physicians sharing a unique patient, or some in-between scenario cannot be determined post-transformation; thus, the projection transformation expends information.

A further strategy is to set weights for the bipartite network prior to forming the projection. For example, in coauthorship networks, the tie connecting an author to a publication might receive a weight of  $1/(N_j - 1)$  where  $N_j$  is the number of authors on paper  $j$  (Newman 2001). (Only papers with at least two authors are used to form such networks.) The rationale is that the greater the number of authors the lower the expected interaction between any pair (a similar logic underlies the example weight matrix described in section “[Network Influence Models](#)”). The sum of the weights across all publications common to two authors is then the basis of their relationship in the author network.

If the events defining the bipartite network occur at different times (e.g., medical claims data often contain time-stamps for each patient-physician encounter), a directed one-mode network may be formed. The value of the A-B and B-A ties in the physician-physician network could be the number of patients who visited A before B and B before A, respectively. In the resulting directed network each physician has a flow to and from each other physician. Subsequent transformation of the flows to binary values yields dyads with states null, directed, and mutual as in a directed unipartite binary network.

Because medical claims and surveys are frequent sources of information about one entity's experience (e.g., a patient) with another entity (e.g., a health plan or physician), bipartite network analysis is an area that promises to have enormous applicability to health services research. Hence, new methods for bipartite network analysis are needed.

## Part II: Statistical Models

We now consider the use of statistical models in social network analysis. Particular emphasis is placed on methods for estimating social influence or peer effects and models for analyzing the network itself, including accounting for social selection through the estimation of effects of homophily.

### Network Influence Models

Reported claims about peer effects of health outcomes such as BMI, smoking, depression, alcohol use, and happiness have recently tantalized the social sciences. In large part, the discussion and associated controversies have arisen from the statistical methods used to estimate peer effects (O'Malley 2013; Christakis and Fowler 2013).

Let  $y_{it}$  and  $\mathbf{x}_{it}$  denote a scalar outcome and a vector of variables, respectively, for individual  $i = 1, \dots, N$  at time  $t = 1, \dots, T$  ( $\mathbf{x}_{it}$  includes 1 as its first element to accommodate an intercept). In this section, the relationship status of individuals  $i$  and  $j$  from the perspective of individual  $i$  (denoted  $a_{ij}$ ) is assumed to be time-invariant. For ease of notation no distinction is made between random variables and realizations of them. The vector  $\mathbf{Y}_t$  and the matrices  $\mathbf{X}_t$  and  $\mathbf{A}$  are the network-wide quantities whose  $i$ th element,  $i$ th row, and  $ij$ th element contain the outcome for individual  $i$ , the vector of covariates for individual  $i$ , and the relationship between individuals  $i$  and  $j$  as perceived by individual  $i$ , respectively. The representation of an example adjacency matrix, denoted  $\mathbf{A}$ , is depicted in Fig. 1.

Regression models for estimating peer effects are primarily concerned with how the distribution of a dependent variable (e.g. a behavior, attitude, or opinion) measured on a focal actor is related to one or more explanatory variables. When behaviors, attitudes, or opinions are formed in part as the result of interpersonal influence, outcomes for different individuals may be statistically dependent. The outcome for one actor will be related to those for the other actors who influence her or him, leading to a complex correlation structure.

In social influence analyses the weight matrix,  $\mathbf{W} = [w_{ij}]$  in Fig. 4, apportions the total influence acting on an individual evenly across the individuals with whom they have a network tie. Typically

1.  $w_{ij} \geq 0$ : nonnegative weights.
2.  $w_{ii} = 0$ : no self-influence.
3.  $\sum_j w_{ij} = 1$ : weights give relative influences (because its row-sums equal 1,  $\mathbf{W}$  is said to be row-stochastic).

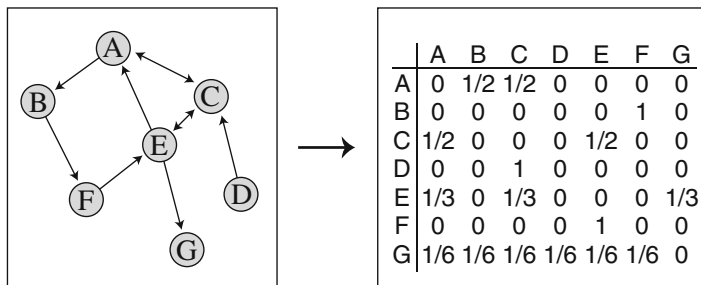
Let  $\bar{y}_{-it} = (\mathbf{W}\mathbf{Y}_t)_i$  denote the influence-weighted average of the outcome  $y$  across the network after excluding (i.e., subtracting) individual  $i$  from the set of individuals to be averaged over. Similarly, let  $\bar{\mathbf{x}}_{-it}^T = (\mathbf{W}\mathbf{X}_t)_i$  denote the vector containing the corresponding influence weighted covariates, often referred to as *contextual variables*.

The most common choice for  $\mathbf{W}$  is the row-stochastic version of  $\mathbf{A}$ . For illustration, suppose that  $\mathbf{A}$  is binary (the elements are 1 and 0). Then the off-diagonal elements on the  $i$ th row of  $\mathbf{W}$  equal  $a_{i+}^{-1}$  if  $a_{i+} > 0$  and  $1/(N-1)$  otherwise (Fig. 4). This framework assumes that an individual's alters are equally influential. In general, influence might only transmit through outgoing ties (e.g., those individuals viewed as friends by the focal actor – a scenario consistent with Fig. 4), or might only transmit through received ties (e.g., individuals who view the focal actor as a friend), or might act in equal or different magnitude in both directions.

Network-related interdependence among the outcomes may be incorporated in two distinct ways. First, an outcome for one actor may depend directly on the lagged outcomes or lagged covariates of the alters to whom she or he is linked. For example, consider the model:

$$y_{it} = \alpha_1 \bar{y}_{-i(t-1)} + \alpha_x^T \bar{\mathbf{x}}_{-i(t-1)} + \beta_1 y_{i(t-1)} + \beta_2^T \mathbf{x}_{i(t-1)} + \varepsilon_{it}, \quad (1)$$

where  $\alpha_1$  is a scalar parameter quantifying the peer effect;  $\alpha_x$  is a  $p$ -dimensional vector of parameters



**Fig. 4** Construction of a network weight matrix  $W$  (right). A directed edge from  $i$  to  $j$  means that node (or individual)  $i$  has a relationship to node  $j$  while element  $ij$  of  $W$  quantifies the extent that individual  $i$  is influenced by individual  $j$ . Although the mathematical form of influence depicted

here assumes that influence only acts in the direction of the edge, influence may in general act in the absence of a tie (e.g., people who consider me as a friend might influence me even if I do not consider them a friend)

of peer effects acting through the  $p$  covariates in  $\mathbf{x}$ ,  $\boldsymbol{\beta} = (\beta_1, \beta_2^T)^T$  is a vector of other regression parameters for the within-individual predictors, and  $E_{it}$  is the independent error assumed to have mean 0 and variance  $\sigma^2$ . The notation used in Eq. 1 is adopted through this section; hence,  $\alpha$  and  $\beta$  denote peer effects and within-individual effects, respectively.

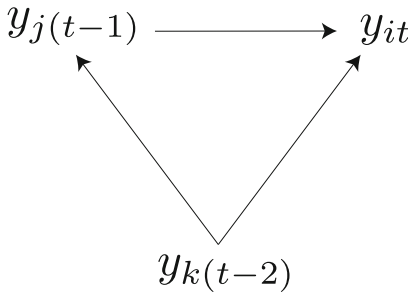
Equation 1 is known as the “linear-in-means model” (Manski 1993) due to the conduit for peer influence being the trait averaged over the alters of each focal actor. The model has a symmetric appearance in that it contains corresponding peer effects for each of the within individual predictors. A common alternative model assumes  $\alpha_x = 0$ ; in other words that peer effects only act through the same variable in the alters as the outcome. Another set of variants arises in the case when there are multiple types of alters with heterogeneous peer effects. Such a situation may be represented in a model by defining distinct influence matrices for each type of peer. Let  $W^{(h)}$  denote the weight matrix formed from the adjacency matrix for the network  $i(t-1)$  comprising only alters of type  $h$  and let  $\bar{y}_{-i(t-1)}^{(h)} = \left( W^{(h)} Y_{t-1} \right)_i$  for  $h = \{1, \dots, H\}$ , where  $H$  is the number of distinct types of alters. Then an extension of the linear-in-means model to accommodate heterogeneous peer effects is:

$$y_{it} = \sum_{h=1}^{(h)} \left( \alpha_1^{(h)} \bar{y}_{-i(t-1)}^{(h)} + \left( \alpha_x^{(h)} \right)^T \bar{\mathbf{x}}_{-i(t-1)} \right) + \beta_1 y_{i(t-1)} + \beta_2^T \mathbf{x}_{i(t-1)} + \varepsilon_{it}. \tag{2}$$

In the special case where  $\alpha_1^1 = \alpha_1^2 = \dots = \alpha_1^{(h)}$  and  $\alpha_x^1 = \alpha_x^2 = \dots = \alpha_x^{(h)}$ , Eq. 2 reduces to Eq. 1. An alternative to Eq. 2 is to fit separate models for each type of peer, which would yield estimates of the overall (or marginal) peer effect for each type of peer as opposed to the independent effect of each type of peer above and beyond that of the other types.

Failing to account for all alters may lead to biased results if the alters are interconnected. Figure 5 presents a simple directed acyclic graph (DAG), which is a device for determining whether or not an effect is identifiable, involving three individuals  $i, j$ , and  $k$ . The nodes represent the variables of interest (a trait measured on each individual such as their BMI) and the arrows represent causal effects (the origin of the arrow is the cause and the tip is the effect). Consider the peer effect of individual  $j$  at  $t-1$  on individual  $i$  at  $t$ . A causal effect is identifiable if it is the only unblocked path between two variables. Because individual  $k$  is a cause of both individual  $j$  and individual  $i$ , the peer effect of  $j$  on  $i$  will be confounded by individual  $k$  unless the analysis conditions on  $y_{k(t-2)}$ .

The scenario depicted in Fig. 5 does not present any major difficulties as long as effects



**Fig. 5** Simplified directed acyclic graph (DAG) illustrating confounding of a peer effect by a third individual. The DAG is simplified because it does not explicitly show the variable  $y_{k(t-1)}$ , which is an intermediary between  $y_{k(t-2)}$  and  $y_{it}$ . (Because the point made here does not depend on  $y_{j(t-2)}$  and  $y_{i(t-1)}$  they are not depicted.) If  $y_{k(t-2)}$  (or  $y_{k(t-1)}$ ) is conditioned on, the path  $y_{j(t-1)} \leftarrow y_{k(t-2)} (\rightarrow y_{k(t-1)}) \rightarrow y_{it}$  is unblocked and therefore confounds  $y_{j(t-1)} \rightarrow y_{it}$ , whose effect is the peer effect of interest. Although the DAG looks like a digraph of a network, a DAG is a different construction

involving individual  $k$  are accounted. However, if individual  $k$  is not known about or is ignored, then the analysis may be exposed to unmeasured confounding. This point has particular relevance to social network analyses as networks are often defined by specifying boundaries or rules for including individuals as opposed to being finite, closed systems (Laumann et al. 1983). In situations where such boundaries break true ties, influential peers may be excluded, potentially leading to biased results.

**Estimation of Contemporaneous Peer Effects**

From a practical standpoint, it may be infeasible to use a model with only lagged predictors such as Eq. 1. For instance, the time points might be so far apart that statistical power is severely compromised. Therefore, it is tempting to use a model with contemporaneous predictors such as:

$$y_{it} = \alpha_0 \bar{y}_{-it} + \alpha_1 \bar{y}_{-i(t-1)} + \alpha_x^T \mathbf{x}_{-it} + \beta_1 y_{i(t-1)} + \beta_2^T \mathbf{x}_{it} + \varepsilon_{it}, \tag{3}$$

where adjusting for  $\bar{y}_{i(t-1)}$  seeks to isolate the peer effect acting since  $t - 1$ . However, because  $\bar{y}_{-it}$  is correlated with the outcome variables of other

observations, OLS will be inconsistent. Therefore, methods are needed to account for endogeneity arising from the correlation between  $\bar{y}_{-it}$  and  $\varepsilon_{it}$  for  $j \neq i$  – in network science parlance the state of  $\bar{y}_{-it}$  is said to be an internal product or consequence of the system as opposed to an external (exogenous) force.

In Christakis and Fowler (2007), the most widely cited of the Christakis-Fowler peer effect papers, the endogeneity problem is resolved using a novel theoretical argument. They purported that it is reasonable to assume in a friendship network that the influence acting on the focal actor (the ego) is greatest for mutual friendships, followed by ego-nominated friendships, followed by alter-nominated friendships, and finally dyads with no friendships. Furthermore, they reasoned that because unmeasured common causes should affect each dyad equally. Because the estimated peer effects were large and positive for mutual friendships but close to 0 for alter and null friendships, consistent with their theory, it was suggested that this constituted strong evidence of a peer effect. Despite the compelling argument, Shalizi and Thomas (2011) revealed that unobserved factors affecting tie-formation (homophily) may confound the relationship and thus lead to biased effects. The estimation of peer effects is a topic of ongoing vigorous debate in the academic and the popular press. Alternative approaches to the theory-based approach of Christakis and Fowler are now described.

A parametric model-based solution to endogenous feedback is to specify a joint distribution for  $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \dots, \varepsilon_{Nt})$ . Then the reduced form of the model satisfies  $\mathbf{Y}_t = \alpha_0 \mathbf{W} \mathbf{Y}_t + \alpha_1 \mathbf{W} \mathbf{Y}_{t-1} + \mathbf{W} \mathbf{X}_{t-1} \boldsymbol{\alpha}_x + \beta_1 \mathbf{Y}_{t-1} + \mathbf{X}_{t-1} \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_t$  for  $\mathbf{Y}_t$  to yield  $\mathbf{Y}_t = (\mathbf{I} - \alpha_0 \mathbf{W})^{-1} \{ \alpha_1 \mathbf{W} \mathbf{Y}_{t-1} + \mathbf{W} \mathbf{X}_{t-1} \boldsymbol{\alpha}_x + \beta_1 \mathbf{Y}_{t-1} + \mathbf{X}_{t-1} \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_t \}$ . The resulting model emulates a spatial autocorrelation model (Anselin 1988). One way of facilitating estimation is by specifying a probability distribution for  $\boldsymbol{\varepsilon}_t$ . However, relying on the correctness of the assumed distribution for identification may make the estimation procedure sensitive to an erroneous assumed distribution.

A semiparametric solution is to find an instrumental variable (IV),  $z_{it}$ , a variable that is related to  $\bar{y}_{-it}$  but conditional on  $\bar{y}_{-it}$  and  $(\bar{y}_{-i(t-1)}, \mathbf{x}_{-it}, y_{i(t-1)}, \mathbf{x}_{it})$  does not cause  $y_{it}$ . If  $\bar{x}_{-it}$  is excluded

from Eq. 3, its elements can potentially be used as Ivs (Fletcher 2008). However, IV methods can be problematic if the instrument is weak or if the assumption that the IV does not directly impact  $y_{it}$  (the exclusion restriction) is violated, an untestable assumption. Thus, in fitting a model with contemporaneous peer effects, one faces a choice between assuming a multivariate distribution holds, relying on the nonexistence of unmeasured confounding variables, or relying on the validity of an IV. None of these assumptions can be evaluated unconditionally on the observed data.

While joint modeling and IV methods provide theoretical solutions to the estimation of contemporaneous peer effects, the notion of causality is philosophically challenged when the cause is not known to occur prior to effect. Therefore, longitudinal data provide an important basis for the identification of causal effects, in particular in negating concerns of reverse causality. If the observation times are far apart the use of lagged alter predictors may, however, substantially reduce the power of an analysis.

### Dyadic Influence Analyses

If the dyads consist of mutually exclusive or isolated pairs of actors there are no interdyad ties and influence only acts within dyads. An example of such a situation occurs when individuals can have exactly one relationship and the relationship is reciprocated, as is the case with spousal dyads. The network influence models of section “[Network Influence Models](#)” reduce to *dyadic influence models* in which the predictors are based on individual alters. For example, the dyadic influence model analogous to Eq. 3 is obtained by replacing the subscript  $-i$  with  $j$ . That is,

$$y_{it} = \alpha_0 \bar{y}_{jt} + \alpha_1 \bar{y}_{j(t-1)} + \alpha_x^T \mathbf{x}_{jt} + \beta_1 y_{i(t-1)} + \mathbf{x}_{it}^T \boldsymbol{\beta}_2 + \varepsilon_{it}. \quad (4)$$

The model in Eq. 4 may be estimated using generalized estimating equations (GEE), avoiding specifying a distribution for  $\varepsilon_{it}$ . However, if any relationships are bidirectional, standard software packages will yield inconsistent estimates of the

peer effects as they do not account for the statistical dependence introduced by individuals who play the dual role of ego and alter at time  $t$  (VanderWeele et al. 2012).

### Frontiers in Social Influence

There has recently been a lot of interest and discussion concerning causal peer effects. Issues that have been discussed include the use of ordinary least squares (OLS) for the estimation of contemporaneous peer effects (Lyons 2011) and the identification of peer effects independent of homophily (Shalizi and Thomas 2011). The discussion has helped elevate social network methodology to the forefront of many disciplines. For example, VanderWeele et al. (2012) show that OLS still provides a valid test of the null hypothesis that the peer effect is zero when the true peer effect is zero. Therefore, OLS can be used to test for peer effects despite the fact that OLS estimates are inconsistent under the alternative hypothesis.

Christakis and Fowler (2007) use tie directionality to account for unmeasured confounding variables under the assumption that their effect on relationship status is the same for all types of relationships. The rationale is that the estimated peer effect in dyads where the relationship is not expected to be conducive to peer influence (“control relationships”) provides a baseline against which to identify the peer effect for other types of relationships. However, this test fails to offer complete protection against unmeasured homophily (Shalizi and Thomas 2011), reflecting the vulnerability of observational data to unmeasured sources of bias. However, sensitivity analyses that evaluate the effect-size needed to overturn the results may be conducted to help support a conclusion by illustrating that the confounding effect must be implausibly large to reverse the finding (VanderWeele 2011).

Instrumental variable (IV) methods have also been used to estimate peer effects. A common source of instruments is alters’ attributes other than the one for which the peer effect is estimated (Fletcher 2008; Fletcher and Lehrer 2009). Potential IVs must predict the attribute of interest in the alter but must not be a cause of the same attribute in other individuals. Attributes that are invisible



such as an individual's genes appear to be ideal candidate genes. For instance, an individual with two risk alleles of an obesity gene is at more risk of increased BMI but conditional on that individual's BMI their obesity genes should not affect the BMI of other individuals. However, if the obesity genes are revealed through another behavior (a phenomenon known as *pleiotropy*) that is associated with BMI then, unless such factors are conditioned on, genes will not be valid IVs.

## Relational Analyses

Sociocentric network studies assemble data on the ties representing the relationship linking a set of individuals, such as all physicians within a medical practice. Models for such data posit that global network properties are the result of phenomena involving subgroups of (most commonly) four or fewer actors (Robins et al. 2005). Examples of such regularities are actor-level tendencies to produce or attract ties (homophily and heterophily), dyadic tendencies toward reciprocity, and triadic tendencies toward closure or transitivity. A relational model, in essence, specifies a set of microlevel rules governing the local structure of a network. In this section, models for cross-sectional relational data are considered first followed by longitudinal counterparts of them.

The simplest models for sociocentric data assume dyadic independence. Under the random model, all ties have equal probability of occurring and the status of one has no impact on the status of another (Erdős and Rényi 1959). More general dyadic models were developed in Holland and Leinhardt (1981) and later were extended in Wang and Wong (1987). Because independence is still assumed between dyads, the information from the data about the model parameters accumulates in the form of a product of the probability densities for the status of the dyadic observation over each dyad:

$$L = \prod_{i < j}^N \text{pr}(a_{ij}, a_{ji} | \boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{x}_{ij}, \mathbf{x}_{ji}), \quad (5)$$

where  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)^T$  and  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_N)^T$  are vectors of actor-specific parameters representing the actors' expansiveness (propensity to send ties) and popularity (propensity to receive ties), respectively, and  $\mathbf{x}_{ij}$  is a vector of covariates relevant to  $a_{ij}$  (this may include covariates specific to either actor and combined traits of both actors). It is important to realize that covariates can be directional; thus,  $\mathbf{x}_{ij}$  need not equal  $\mathbf{x}_{ji}$ . Although the model may include other parameters,  $\boldsymbol{\alpha}$  and  $\boldsymbol{\gamma}$  play an important role in network analysis due to their relationship to the degree distribution of the network and so are explicitly denoted.

When relationship status is binary, the distribution of  $(a_{ij}, a_{ji})$  is a four-component multinomial distribution. The probabilities are typically represented in the form of a generalized logistic regression model (an extension of the logistic regression model to  $\geq 2$  categories) having the form

$$\text{pr}(a_{ij}, a_{ji} | \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \kappa_{ij}^{-1} \exp(\mu_{ij} a_{ij} + \mu_{ji} a_{ji} + \rho_{ij} a_{ij} a_{ji}), \quad (6)$$

where

$$\kappa_{ij} = 1 + \exp(\mu_{ij}) + \exp(\mu_{ji}) + \exp(\mu_{ij} + \mu_{ji} + \rho_{ij}),$$

and  $\mu_{ij}$ ,  $\mu_{ji}$ , and  $\rho_{ij}$  are functions of  $(\alpha_i, \alpha_j, \gamma_i, \gamma_j)$  and  $(\mathbf{x}_{ij}, \mathbf{x}_{ji})$ . The term  $\mu_{ij}$  includes factors associated with the likelihood that  $a_{ij} = 1$  but not necessarily the likelihood that  $a_{ji} = 1$ . In a nondirected network the predictors can be directional and so it is likely that  $\mu_{ij} \neq \mu_{ji}$ . However, the only covariates included in  $\rho_{ij}$  must be non-directional as they affect the likelihood of  $(a_{ij}, a_{ji}) = (1, 1)$ ; the sign of  $\rho_{ij}$  indicates whether a mutual tie is more (if  $\rho_{ij} > 0$ ) or less (if  $\rho_{ij} < 0$ ) likely to occur than predicted by the density terms and so is a measure of *reciprocity* or *mutuality*. Null mutuality is implied by  $\rho_{ij} = 0$ .

In dyadic models, the terms  $\mu_{ij}$ ,  $\mu_{ji}$ , and  $\rho_{ij}$  account for the local network about actors  $i$  and  $j$  through the inclusion of  $(\alpha_i, \alpha_j, \gamma_i, \gamma_j)$ . Furthermore, other effects can be homogeneous across actors or actor-specific. For example, the  $p_1$  model

(Holland and Leinhardt 1981) assumes  $\mu_{ij} = \mu + \alpha_i + \gamma_j$  and  $\rho_{ij} = \rho$ , implying the covariate-free joint probability density function of the network given by

$$p_1(\mathcal{A}) \propto \exp \left\{ \mu s_1(\mathcal{A}) + \sum_i^N \alpha_i s_{2i}(\mathcal{A}) + \sum_j^N s_{3i}(\mathcal{A}) + \rho s_4(\mathcal{A}) \right\},$$

where  $s_1(\mathcal{A}) = \sum_{i=j} a_{ij}$ ,  $s_{2i}(\mathcal{A}) = a_{i+}$ ,  $s_{3j}(\mathcal{A}) = a_{+j}$  and  $s_4(\mathcal{A}) = \sum_{i \neq j} a_{ij} a_{ji}$ . Thus, the  $p_1$  model depends on  $2N + 2$  network statistics and associated parameters. If the  $p_1$  model holds within (ego, alter)-shared values of categorical attributes, a stochastic block model is obtained by allowing block-specific modifications to the density and reciprocity of ties (Fineberg and Wasserman 1981; Holland et al. 1983; Wang and Wong 1987). An extension would allow reciprocity to also vary between blocks. Because the stochastic blockmodel extension of the  $p_1$  model is saturated at the actor-level due to the expansiveness and popularity fixed effects, no assumption is made about differences in the degree-distributions of the actors in different blocks. Stochastic block models are the basis of mixed-membership and other recent statistical approaches for node-partitioning social network data (Goldenberg et al. 2009; Choi et al. 2010; Karrer and Newman 2011). Individuals in the same block of a stochastic block model are often referred to as being structurally equivalent.

### Models of Networks as Single Observations

A criticism of dyadic independence models is that they fail to account for interdependencies between dyads. The  $p^*$  or exponential random graph model (ERGM) generalizes dyadic independence models (Frank and Strauss 1986; Wasserman and Pattison 1996). An ERGM has the form

$$\Pr(\mathcal{A}; \theta) = \kappa(\theta)^{-1} \exp \left( \sum_k \theta_k s_k(\mathcal{A}) \right), \quad (7)$$

where  $\mathcal{A}$  denotes a possible state of the network,  $s_k(\mathcal{A})$  denotes a network statistic evaluated over  $\mathcal{A}$  (e.g., the number of ties, the number of reciprocated ties),  $\kappa(\theta) = \sum_{\mathcal{A} \in \mathcal{A}} \exp(\sum_k \theta_k s_k(\mathcal{A}))$ , and  $\mathcal{A}$  is the set of all  $2^{N(N-1)}$  possible realizations of a directed network. In general, the scale factor  $\kappa(\theta)$  that sums over each distinct network does not factor into a product of analogous terms. As a result, it is computationally infeasible to exactly evaluate the likelihood function of dyadic dependent ERGMs for even moderately sized  $N$  (e.g.,  $N > 20$  is problematic (Hunter and Handcock 2006)). The key feature of the  $p_1$  model that allows the probability of the network to decompose into the product of dyadic-state probabilities is that it only depends on network statistics  $s_k(\mathcal{A})$  that sum individual ties or pairs of ties from the same dyad.

If dyads are independent unless they share an actor, the network is a *Markov Random Graph* (Frank and Strauss 1986). Markov Random Graphs may include terms for density, reciprocity, transitivity and other triadic structures, and  $k$ -stars (equivalent to the degree distribution) – these terms contain sums of the products of no more than three ties. Such terms may be multiplied with actor attribute variables to define interaction effects. (An interaction is the effect of the product of two or more variables, e.g., if males and females have different tendencies to reciprocate ties then gender is said to interact with reciprocity.)

Networks that extend Markov Random Graphs by allowing four-cycles but no fifth- or higher-order terms are *partially conditionally dependent*. In such networks, a sufficient condition for dependence of  $a_{ij}$  and  $a_{kl}$  is that  $a_{ik} = a_{jl} = 1$  or  $a_{il} = a_{jk} = 1$  (Wang et al. 2009). Thus, two edges may be dependent despite not having any actors in common. Partial conditional dependence is the basis of the new parameterizations of network statistics developed by Snijders (2006) that have led to better fitting ERGMs (see below).

Under ERGMs, the conditional likelihood of each tie given the other ties in the network has the logistic form:

$$\Pr(a_{ij} = 1 | \mathbf{A}_{ij}^c) = \left[ 1 + \exp(\boldsymbol{\theta}^T \delta(\mathbf{A}_{ij}^c)) \right]^{-1}, \quad (8)$$

where  $\mathbf{A}_{ij}^c$  is  $\mathbf{A}$  with  $a_{ij}$  excluded,  $\delta(\mathbf{A}_{ij}^c) = S(\mathbf{A}_{ij}^+)$   $- S(\mathbf{A}_{ij}^-)$  is the vector of changes in network statistics that occur if  $a_{ij}$  is 1 rather than 0. Thus, the parameters of an ERGM are interpreted as the change in the log of the odds that the tie is present to not being present conditional on the status of the rest of the network (Snijders 2006). A large positive parameter suggests that more configurations of the type represented in the network statistic appear in the observed network more often than expected by chance, all else equal (Robins et al. 2009).

Due to the factorization of the likelihood function in Eq. 5, likelihood-based estimators of dyadic independence models have desirable statistical properties such as consistency and statistical efficiency. However, if the model for the network includes predictors based on three or more actors, no such factorization occurs and Markov chain Monte Carlo (MCMC) is required to optimize the likelihood function for Eq. 7, which for each observation involves making computations on  $k^N (N-1)^2$  ( $k = 4$  if directed and  $k = 2$  if nondirected) distinct networks. ERGMs have been demonstrated to be estimable on networks with  $N \approx 1600$  (Goodreau 2007), but computational feasibility depends on the terms in the model and the amount of memory available. The ERGM (“Exponential Random Graph Model”) package that is part of the Stat-net suite in R, developed by the Statnet project, estimates ERGMs (Handcock et al. 2010).

Other estimation difficulties include failure of the optimization algorithm to converge and the fitted model producing nonsensical “degenerate” predicted networks. *Degeneracy* arises because for certain specifications of  $s_k(\mathbf{A})$  the network statistics are highly collinear or there is unaccounted effect heterogeneity across the network. As a result, under the fitted model the local neighborhood of networks around the observed network may have probability close to 0 and those networks with positive probability (often the

empty and complete graphs) may be radically different from each other and thus the observed network (Handcock et al. 2003; Robins et al. 2007). Although the average network over repeated draws has similar network statistics to the observed network, the individual networks generated under the fitted model do not bear any resemblance to the observed network.

Because an actor of degree  $m$  contributes  $k$ -stars for  $k \leq m$ ,  $k$ -star configurations are nested within one another and thus are highly correlated. Therefore, when multiple  $k$ -stars are predictors, extensive collinearity results. However, the estimated coefficients of successive  $k$ -star configurations (e.g., 2-star, 3-star, 4-star) tend to decrease in magnitude and have alternating signs, an observation often seen when multiple highly collinear variables are included in a regression model. This observation led to the development of the *alternating  $k$ -star* (Snijders 2006), given by

$$\text{AS}(\lambda) = \sum_{k=2}^{N-1} (-1)^k \frac{S_k}{\lambda^{k-2}} \text{ for } \lambda > 1,$$

where  $S_k$  denotes the number of  $k$ -stars, being used in place of multiple individual  $k$ -star terms in Eq. 7. A positive estimate of the coefficient of  $\text{AS}(\lambda)$  suggests that the degree distribution is skewed towards higher degree nodes while a negative coefficient implies large degrees are unlikely. The value of  $\lambda$  can be specified or estimated from the data (Hunter 2007).

Network statistics for triadic configurations – the triangle (a nondirected closed triad) in nondirected networks and transitive triads, three-cycles, closed three-out stars, closed three-in stars in directed networks – are the most prone to degeneracy. One reason is that heterogeneity in the prevalence of triads across the network leads to heterogeneity in the density of ties across the network (Robins et al. 2009). A model that assumes homogeneous triadic effects across the network is unable to describe networks with regions of high and low density; the generated networks are either dominated by excessive low-density regions or by excessive high-density regions. This observation suggests a hierarchical

modeling strategy where the first step is to use a community detection algorithm (see section “[Network Communities](#)”) to partition the network into blocks of nodes. Then fit an ERGM (or other model) to the subnetwork corresponding to each community, allowing the network statistics to have different effects within each community. The just-described modeling strategy combines methods of network science and social network analysis.

A similar approach has been used to overcome severe computational difficulties that often occur when one or multiple triadic (triangle-type) terms are included in the model. A  $k$ -triangle is a set of  $k$  triangles resting on a common base. For example, if individuals  $i, j$ , and  $k$  are one closed triad and individuals  $i, j$ , and  $l$  are another then the four individuals form a 2-triangle with the edge  $y_{ij}$  common to both. Let  $T_k$  denote the number of  $k$ -triangles in the network. Thus,  $T_1$  denotes the total number of closed triads,  $T_2$  the total number of 2-triangles, and so on. The *alternating  $k$ -triangle statistic*

$$AT(\lambda) = \sum_{k=1}^{N-3} (-1)^k \frac{S_{k+1}}{\lambda^k} \text{ for } \lambda > 1,$$

was developed to perform for triadic structures what  $AS(\lambda)$  performs for  $k$ -stars (Snijders 2006). The presence of  $\lambda$  makes  $AT(\lambda)$  nonlinear in the triangle count, giving lower probability to highly clustered structures. By making the number of actors who share  $k$  partners the core term,  $AT(\lambda)$  can be rewritten as a geometrically weighted edgewise shared partner (GWESP) statistic (Goodreau 2007; Hunter 2007).

The  $AS(\lambda)$  and  $AT(\lambda)$  statistics do not differentiate between outward and inward ties. Recently, directed forms of these statistics have been introduced (Robins et al. 2009). The directed versions of the  $k$ -star are threefold, corresponding to two paths, shared destination node (activity), shared originator node (popularity). The directed versions of the  $k$ -triangle represent transitivity, activity closure, popularity closure, and cyclic-closure.

## Bipartite ERGMs

An alternative approach to modeling a one-mode projection (by construction a nondirected network) from a two-mode network is to directly model the two-mode network. An advantage of direct modeling is that all the information in the data is used. ERGMs or any other model applied to bipartite data need to account for the fact that ties can only form in dyads including one actor from each mode. In a dyadic independence model this is recognized simply by excluding all same mode dyads from the dataset. In general, the denominator  $\kappa(\theta)$  in Eq. 7 only sums over networks in which there are no within mode ties. If the number of actors in the two modes are  $N$  and  $M$ , there are  $2^{NM}$  distinct nondirected networks.

The density and degree distributions may be represented in a bipartite ERGM as in a unipartite ERGM. However, with two modes it may be that two types of each network statistic and other predictor is needed. Representations of homophily in two-mode networks are defined across modes. Likewise, because there are no within-mode ties, statistics that account for closure must also depend only on inter-mode ties.

The smallest closed structure in a bipartite graph is a four-cycle (closed four-path). An example of a four-cycle is the path A–1–C–2–A in Fig. 2; it includes four distinct actors and four edges are traversed to return to the initial actor. A simple measure of closure contrasts the number of closed four-cycles out of all three paths containing four unique actors with the overall density of ties. A simple model for testing whether clustering (closure) is present in a bipartite network includes density, both sets of  $k$ -stars, three-path, and four-cycle statistics as predictors. A significant positive effect of the four-cycle statistic suggests that two actors of degree two in one mode that have one of the actors in the other mode in common are more likely to also have the second actor in common, relative to two randomly selected actors of degree two from the same mode. For example, in a physician-patient network, clustering implies having one patient in common increases the likelihood of having another patient in common. Physicians A and C both have patients 1 and 2 in common, hence they

provide evidence for bipartite closure. However, physicians E and F have patient 3 in common; despite being eligible to exhibit bipartite closure they do not, and hence they provide evidence against bipartite closure.

Analogies of ERGMs and solutions to problematic issues exist for bipartite networks. For example, to avoid problems of high colinearity between the  $k$ -star terms, alternating  $k$ -star statistics can be used in place of them (Wang et al. 2009). Let  $S_D(\mathbf{B})$  denote the number of ties from one mode to the other,  $AS_1(\mathbf{B})$  and  $AS_2(\mathbf{B})$  denote the alternating  $k$ -star statistics for each mode,  $S_{3P}(\mathbf{B})$  denote the number of three-paths, and  $S_{4C}(\mathbf{B})$  denote the number of closed four-cycles for a network  $\mathbf{B}$ . The resulting bipartite ERGM for  $\mathbf{B}$  has the form:

$$\Pr(\mathbf{B}; \theta) = \kappa(\theta)^{-1} \exp(\theta_0 S_D(\mathbf{B}) + \theta_1 AS_1(\mathbf{B}) + \theta_2 AS_2(\mathbf{B}) + \theta_3 S_{3P}(\mathbf{B}) + \theta_4 S_{4C}(\mathbf{B})), \quad (9)$$

where  $\kappa(\theta)$  sums over the  $MN$  possible bipartite graphs. The statistic  $S_{4C}(\mathbf{B})/S_{3P}(\mathbf{B})$  is the proportion of times that two patients each visit the same two physicians out of all the occurrences where two patients both have one visit to one physician and one patient visits the other physician. The coefficient  $\theta_4$  is the effect associated with this lowest-order form of closure in a two-mode sense (but should not be thought of as reciprocity because the network is nondirected).

### Longitudinal ERGMs

The development of relational models has primarily focused on cross-sectional data. However, extensions of ERGMs to longitudinal scenarios have been developed – most often involving a Markov assumption to describe dependence across time. The first longitudinal ERGMs treated tie-formation and tie-dissolution as equitable events in the evolution of the network (Hanneke et al. 2010). A more general formulation treats tie-formation (attractiveness in the context of network science) and tie-duration (the complement of tie-duration referred to as fitness in network science) as separable processes, thereby allowing the same network statistic to impact tie-formation

and tie-dissolution differently (Krivitsky and Handcock 2010).

Like ERGMs for cross-sectional data, longitudinal ERGMs are defined by statistics that count the number of occurrences of substructures in the network. However, in addition to the current state of the network, such statistics may also depend on previous states. Under Markovian dependence, network statistics only depend on the current and the most recent state; for example, the number of ties that remain intact from the preceding observation. The recently released TERGM (“temporal exponential random graph model”) package in the Statnet suite in R estimates ERGMs for discrete temporal (i.e., longitudinal) sociocentric data (Hanneke et al. 2010).

### Actor-Orientated Approaches

An alternative approach for modeling network evolution is the actor-oriented model (Snijders 1996, 2001, 2005). This centers on an objective function that actors seek to maximize and which may be sensitive to multiple network properties, including reciprocity, closure, homophily, or contact with high-degree actors. The model assumes that actors control their outgoing ties and change them in order to increase their satisfaction with the network in one or more respects as quantified by the objective function. It resembles a rationale choice model in which each agent attempts to maximize their own utility function. Estimated parameters indicate whether changes in a given property raise or lower actor satisfaction.

An important distinction of actor-oriented models from ERGMs is that the relevant network statistics in the actor-oriented model are specific to individuals rather than being aggregations across the network. However, like ERGMs, estimation is computationally intensive. The SIENA package in StOCNET (Huisman and Van Duijn 2004, 2005) uses a stochastic approximation algorithm but struggles with networks of appreciable size (e.g., thousands of individuals). Because they only resemble ERGMs in the limiting steady-state case, actor-oriented models may also suffer from degeneracy but the problem is less profound (Goldenberg et al. 2009).

**Joint Models**

A virtue of the actor-oriented modeling framework in SIENA is that an actor’s relationships can be modeled jointly with the social-influence effects of an actor’s peers on their own traits. If the model is correctly specified, it has the potential to account for unmeasured confounding factors that affect both the evolution of relationship status and the values of individuals attributes, yielding unbiased estimates of the effects of observed variables affecting social influence and the evolution of the network. Such a model was developed by Steglich and colleagues (Steglich et al. 2010), but to date work in this area is limited.

**Latent Independence Approaches**

In ERGMs a huge increase in computational complexity occurs between the dyadic-independent and dyadic-dependent models. A second concern about ERGMs is that in general they are not consistent under sampling in the sense that statistical inferences drawn from the network for the sample do not generalize to the full network (Shalizi and Rinaldo 2012). The few ERGMs to exhibit such consistency include the dyadic independent  $p_1$  and stochastic block models. An alternative modeling strategy provides a more graduated transition between independence and dependence scenarios by using random effects to model dyadic dependence and also ensures consistency between the results of analyzing the sample and the population of interest. Random effects are used to account for dyadic independence in the  $p_2$  model (Duijn et al. 2004; Zijlstra et al. 2006) introduced below.

The  $p_2$  model is much like the  $p_1$  model except that the expansiveness  $\alpha_i$  and popularity  $\gamma_i$  parameters are random as opposed to fixed effects.

Typically,  $(\alpha_i, \gamma_i)$  is assumed to be bivariate normal with covariance matrix  $\Sigma_{\alpha\gamma}$ . Therefore, the  $p_2$  model is given by

$$\text{pr}(a_{ij}, a_{ji} | x_{ij}, x_{ji}) = \kappa_{ij}^{-1} \exp(\mu_{ij} a_{ij} + \mu_{ji} a_{ji} + \rho_{ij} a_{ij} a_{ji}), \tag{10}$$

$$\begin{aligned} \text{where } \kappa_{ij} &= 1 + \exp(\mu_{ij}) + \exp(\mu_{ji}) \\ &\quad + \exp(\mu_{ij} + \mu_{ji} + \rho_{ij}), \\ \mu_{ij} &= \mu + \alpha_i + \gamma_j + \beta^T x_{ij}, \\ \rho_{ij} &= \rho + \beta^T x_{2ij}, \end{aligned}$$

and  $(\alpha_i, \gamma_i) \sim \text{Normal}(\mathbf{0}, \Sigma_{\alpha\gamma})$ . Thus,  $\mathbf{x}_{ij} = (\mathbf{x}_{1ij}, \mathbf{x}_{2ij})$  and  $\mathbf{x}_{2ij}$  includes a subset of covariates that are symmetric ( $\mathbf{x}_{2ij} = \mathbf{x}_{2ji}$ ) in reflection of the fact that reciprocity is a symmetric phenomenon. Conditional on  $(\alpha_i, \gamma_i)$  the model implies that the relationship status of one dyad does not depend on that of another. A positive off-diagonal element of  $\Sigma_{\alpha\gamma}$  implies that expansive individuals also tend to be popular.

The  $p_2$  model can be extended to account for more general forms of dyadic dependence than the latent propensity of an individual to send or receive ties. Let each individual have a vector of latent variables, denoted  $z_i$  in the case of individual  $i$ , that together with the same for individual  $j$  affects the value of the relationship between  $i$  and  $j$ . The dependence of tie-status on  $z_i$  is generally represented using a simple mathematical function. The major types of models are latent class models (Nowicki and Snijders 2001; Airoldi et al. 2008), latent distance models (Hoff et al. 2002; Handcock et al. 2007), and latent eigenmodels (Hoff 2005, 2008). These models are characterized by the form of the latent variable

$$\xi(z_i, z_j) = \begin{cases} \lambda z_i, z_j \text{ where } z_i, z_j \in \{1, \dots, K\} \text{ and } \lambda z_i, z_j = \lambda z_j, z_i \\ -|z_i - z_j|^c \text{ where } c > 0 \text{ and } z_i, z_j \text{ have } K \text{ elements} \\ z_i^T U z_j^T \text{ where } z_i \sim N(0, \Sigma_z) \text{ and } U \text{ is a } K - \text{dimensional diagonal matrix} \end{cases} \tag{11}$$

which is included as an additional predictor in  $\mu_{ij}$ . In Eq. 11 the form and interpretation of  $z_i$  changes from denoting a scalar  $\xi(z_i, z_j)$  categorical latent

variable in the latent class model (first row) to a position in a continuously valued multidimensional space in the latent distance and latent eigenmodels

(second and third rows, respectively). The term  $\xi(z_i, z_j)$  can be added to either the  $\mu_{ij}$  or  $\rho_{ij}$  components of the  $p_2$  model to allow higher-order dependence to moderate the effect of density and reciprocity, respectively.

In the latent class specification the array of values of  $\lambda z_i z_j$  form a symmetric  $K \times K$  matrix  $\Lambda$ . A basic specification is  $\lambda z_i z_j = \lambda_0$  if  $z_i = z_j$  (nodes in same partition) and  $\lambda z_i z_j = 0$  if  $z_i \neq z_j$  (nodes in different partitions) (Nowicki and Snijders 2001; Airoldi et al. 2008). Latent class models extend stochastic-block models to allow latent clusters as well as observed clustering variables. This family of models is suited to network data exhibiting structural equivalence, that is, under the model individuals are hypothesized to belong to latent groups such that members of the same group have similar patterns of relationships.

In the latent distance specification the most common values for  $c$  are 1 and 2, corresponding to absolute and cartesian distance, respectively. The distance metric accounts for latent homophily – the effect of unobserved individual characteristics that induce ties between individuals. In this model,  $z_i$  can be interpreted as the position of individual  $i$  in a social space (Hoff et al. 2002). This model accounts for triadic dependence (e.g., transitivity) by requiring that latent distances between individuals obey the triangle inequality. Latent distance models are available in the LatentNet package in R (Krivitsky and Handcock 2008).

The latent eigenmodel is the most general specification and accounts for both structural equivalence and latent homophily. Furthermore, the parameter space of the latent eigenmodel model of dimension  $K$  generalizes that of the latent class model of the same dimension and weakly generalizes the latent distance model of dimension  $K - 1$ . Conversely, the latent distance model of dimension  $K$  does not generalize the one-dimensional latent eigenmodel model (Hoff 2008). The closeness of the latent factors  $U^{1/2} z_i$  and  $U^{1/2} z_j$  quantifies the structural equivalence of actors  $i$  and  $j$  positions in the network; a tie is more likely if  $U^{1/2} z_i$  and  $U^{1/2} z_j$  have a similar direction and magnitude, allowing for more clustering than under Eq. 10. On the other hand, latent homophily is accounted for by the diagonal elements of  $U$ , which can be positive or

negative (allowing for heterophily as well as homophily). The model constrains the extent to which the quadratic forms  $z_i^T U z_j$ ,  $z_i^T U z_k$ , and  $z_j^T U z_k$  constructed from the latent vectors vary from one another. The greater the magnitude of  $\Sigma_z = \text{cov}(z_i)$  the greater the extent to which ties are expected to cluster and form cliques. The latent eigenmodel model is appropriate if a network exhibits clustering due to both structural equivalence and unmeasured homophily.

In Hoff (2005) and (2008) models are specified at the tie level with reciprocity (in directed networks) represented as the within-dyad correlation between two tie-specific latent variables. Modeling reciprocity as a latent process differs from the  $p_2$  model, in which reciprocity is represented as a direct effect (Paul and O'Malley 2013). Therefore, an alternative family of latent variable models for networks is obtained by augmenting the density term in the  $p_2$  model with Eq. 11. An advantage of specifying a joint model at the dyad level is that the resulting (extended- $p_2$ ) model involves  $N(N - 1)$  fewer latent variables, possibly alleviating computational issues such as nonidentifiability of parameters or multiple local optima.

The challenges of estimating models involving latent variables resemble those of factor analysis or other dimension-reduction methods. First, an appropriate value of  $K$  may not be able to be specified from existing knowledge of the network, and estimating  $K$  from the data is not straightforward. Second, computational challenges in estimating the latent variables can make the method difficult to apply to large networks. However, such issues are more easily overcome than degeneracy in ERGMs. Degeneracy is avoided in these models as the model for a dyad determines the distribution of the network. In other words, the factorization of the likelihood into a product of like terms ensures that networks sampled under the model are almost surely in the neighborhood of the observed network, increasingly so as  $N$  increases (i.e., asymptotically). Another contrast with ERGMs is that the model describes a population as opposed to the single observed network. Thus, in latent variable models the data-generating process is modeled whereas ERGMs are specific to the observed network and so have more in common with finite population inference.

Another advantage of conditional independence models over ERGMs is that the same types of models can be applied to valued relational data. Analogous to generalized linear models, the link function and any parametric distributions assumptions that define a conditional independence network model can be tailored to the type of relationship variable (scale, count, ratio, categorical, multivariate). However, a recent adaptation of ERGMs has been proposed for modeling count-valued socio-centric data (Krivitsky 2012).

Offsetting the above advantageous features of conditional independence models is that terms such as  $\xi(z_i, z_j)$  are limited from the hypothesis testing and interpretational standpoint in that they do not distinguish particular forms of social equivalence or latent homophily. For example, the effect of transitivity is not distinguished from that of cyclicity or higher-order clustering, such as tetradic closure. Therefore, the choice of model in practice might depend on the importance of testing specific hypotheses about higher-order effects to obtaining a model whose generative basis allows it to make predictions beyond the data set on which the model was estimated.

### Longitudinal Conditional Independence Models

Longitudinal counterparts of conditional independence models are obtained by introducing terms that account for longitudinal dependence (e.g., past states of the dyad). A simple Markov transition model was developed in O'Malley and Christakis (2011) with tie-formation and tie-dissolution treated as unrelated processes. Conditional on the past state of the dyad and the sender and receiver random effects, the value of each tie is assumed to be statistically independent of that of any other tie. A more general formulation extends the  $p_2$  model, allowing dependence between ties within a dyad (reciprocity), heterogeneous effects in the formation and dissolution of ties, and the inclusion of higher-order effects (e.g., third-order interactions to account for transitivity) as lagged predictors (Paul and O'Malley 2013).

The approach in Paul and O'Malley (2013) is notable for attempting to capture the best of both worlds: it allows localized (actor or dyadic) versions of the higher-order predictors available in ERGMs to be included as predictors, but avoids degeneracy by using their lagged values as opposed to their current values as predictors. Therefore, conditional on the observed and latent predictors, dyads are cross-sectionally independent but longitudinally dependent on prior states of other dyads (in addition to their own past states) in the network. An extension that builds on Paul and O'Malley (2013) is to incorporate the latent class, distance, or eigenfactor terms in Eq. 11 in the model. Such a model was entertained in Westveld and Hoff (2011) but has not yet been developed.

---

## Part III: Network Science

We now switch attention to methods that have been derived and used in the field of network science. In general, network science approaches avoid assumptions about distributions in models. For example, to test whether a network exhibits a certain property, the commonly employed approach is to use a permutation test to develop a null distribution for a statistic that embodies the property in question and then evaluate how extreme the observed value of the statistic is with respect to the null distribution. This technique is the cornerstone of the procedure used to evaluate the degree of separation to which social clustering can be detected in Szabo and Barabasi (2007).

Network science focuses not only on social networks but also covers information networks, transportation networks, biological networks, and many others. Most of the networks studied within network science are non-directed as ties are typically thought of as connections as opposed to measures for which the distinction between instigator and receiver is relevant. Thus, the networks in this section are assumed to be nondirected unless stated otherwise.



## Generative Models of Network Formation

Network science has taken a somewhat different approach to modeling networks than the social sciences or statistics. Essentially all models developed within network science are *generative models*, sometimes also known as forward models, in contrast to probabilistic models such as ERGMs. These models start from a set of simple hypothesized mechanisms, often functioning at the level of individual nodes and ties, and attempt to describe what types of network structures emerge from a repeated application of the proposed mechanisms. Many of the models describe growing networks, where one starts from a small connected seed network consisting of a few connected nodes, and then grows the network by subsequent addition of nodes, usually one at a time. The *attachment rules* specify how exactly an incoming node attaches itself to the existing network.

Generative models are commonly exploratory in nature. If they reproduce the type of structure observed in an empirical network, it is plausible that the proposed mechanisms may underlie network formation in the real world. The main insight to be gained from a generative model is a potential explanation for why a network possesses the type of structure it does. Many of the models are simple in nature, which occasionally leads to analytical tractability, but the main reason for simplicity is the potential to expose clearly the main mechanism(s) driving the phenomenon of interest. It is not uncommon for generative models to possess only two or three parameters, yet occasionally simple generative mechanisms can explain some of the key features surprisingly well. Once a model can explain the main features, it can be fine-tuned by adding more specific or nuanced mechanisms. A few examples of generative models are now described.

### Cumulative Advantage Model

Cumulative advantage refers to phenomena where success seems to breed success, such as in the case of accumulation of further wealth to already wealthy individuals. In networks of scientific

citations, where a node represents a scientific paper, each node has some number of edges pointing to nodes that correspond to cited papers (de Solla Price 1965). In the present context, for example, there would be an edge pointing from the node representing this chapter to the node representing the 1965 *Science* paper of Price. While the out-degree of nodes is fairly uniform, as the length of bibliographies is fairly constrained, the in-degree distribution was found to be fat-tailed with the functional form of a power-law,  $P(k) \sim k^{-\alpha}$  (de Solla Price 1965).

Price later proposed a mathematical model for cumulative advantage processes, “the situation in which success breeds success” (Price 1976). In this model, nodes are added to the network one at a time, and the average out-degree of each node is fixed. The attachment rule in the model specifies that each new paper will cite existing papers with probability proportional to the number of citations they already have. Thus each incoming node will attach itself with some number of directed edges to the existing network, the exact number of ties being drawn from a distribution, and the nodes these new edges are pointing to will be chosen proportional to their in-degree. In this formulation, however, papers with exactly zero citations can never accrue citations. To overcome this problem, one can either consider the original publication as the first citation so that each paper starts with one citation or, alternatively, add a small constant to the number of citations (Price 1976). Either way, the outcome is that the target nodes are chosen in proportion to their in-degree plus this small positive constant. A derivation of the resulting in-degree distribution is given by Newman (2010). Denoting the average out-degree of a node by  $c$  and using  $a$  to denote the small positive constant, the in-degree distribution  $P(k)$  for large values of  $k$  has the power-law form  $P(k) \sim k^{-\alpha}$ , where  $\alpha = 2 + a/c$ .

This simple model (although the derivation of the result is quite involved) is able to reproduce the empirical citation (in-degree) distribution for scientific papers with surprising accuracy given that the model only contains two parameters. It may seem odd that the model does not incorporate any notion of paper quality, which surely should

be an important driver of citations. Here it is important to notice that the model does not make any attempt to predict *which* paper becomes popular (although it can be shown, using the model, that papers published at the inception of a field have a much higher probability to become popular). Instead, the model incorporates the quality of papers implicitly, and indeed the number of citations to a paper is frequently seen as an indicator of its quality. Popular papers are also easily discovered, which further feeds their popularity. The idea of using popularity as a proxy for quality may extend to other areas where resources are scarce, for example, skilled surgeons are in high demand.

### Preferential Attachment Model

The cumulative advantage model of Price (1976) is developed as a modification of the Polya urn model, which is used to model a sampling process where each draw from the urn, corresponding to a collection of different types of objects, changes the composition of the urn and thereby changes the probability of drawing an object of any type in the future. The standard Polya urn model consists of an urn containing some number of black and white balls, drawing a ball at random and then returning it to the urn along with a new ball of the same color (Feller 1966). Independently of Price, Barabasi and Albert introduced a similar model in 1999 (Barabasi and Albert 1999). They examined the degree distributions of an actor collaboration network (two actors are connected if they are cast in the same movie), World Wide Web (two web pages are connected if there is a hyperlink from one page to the other), and power grid (two elements (generators, transformers, substations) are connected if there is a high-voltage transmission line between them), finding that they approximately followed power-law distributions. Although the actor collaboration network and the power grid networks are defined much like a projection from a two-mode to a one-mode network, a subtle difference between them is that direct interaction between the nodes can be assumed. In other words, the nodes can be thought of as directly linked.

Both of the generic network models in existence at the time, the Erdős-Rényi and the Watts-

Strogatz models, operated on a fixed set of  $N$  vertices, and assumed that connections were placed or rewired without any regard to the degrees of the nodes to which they were connected. The model of Barabasi and Albert changed both of these aspects. First, they introduced the notion of network growth, such that at each time step a new node would be added to the network. Second, this new node would connect to the existing network with exactly  $m$  nondirected edges, and the nodes they attached to were chosen in proportion to their degree. The probability for the incoming vertex to connect to vertex  $i$  depends solely on its degree  $k_i$  and is given by

$$\Pi(k_i) = k_i / \sum_j k_j.$$

The model was solved by Barabasi and Albert using rate equations, which are differential equations for the evolution of node degree over time where both degree and time, as an approximation, are treated as if they were continuous variables (Barabasi and Albert 1999; Barabasi et al. 1999). More general solutions were provided by Krapivsky et al. also using rate equations (Krapivsky et al. 2000) and Dorogovtsev et al. using master equations which, like rate equations, are differential equations for the evolution of node degree, but they (correctly) treat degree as a discrete variable while still making the continuous-time approximation for time (Dorogovtsev et al. 2000). In the master equation approach, one writes down an equation for the evolution of the number of nodes of a given degree. Let us use  $N_k(t)$  to denote the number of nodes of degree  $k$  in the network at time  $t$ , where time is identified with network size, i.e., time  $t$  corresponds to the network at the point of its evolution when it consists of  $t$  nodes. (The nodes making up the seed network can be usually ignored in the limit as time increases.) The number  $N_k(t)$  can change in two ways: it can either increase as an incoming node attaches itself to a node of degree  $k - 1$  and thus turn it into a node of degree  $k$ , or it can decrease as an incoming node attaches itself to a node of degree  $k$ , turning into a node of degree  $k + 1$ . The former situation leads to  $N_k(t + 1) = N_k(t) + 1$

and the latter to  $N_k(t+1) = N_k(t) - 1$ . Transitions larger than one, e.g., from  $k$  to  $k+2$  or from  $k$  to  $k-2$  are very unlikely and can be ignored. The value of  $N_m(t)$  increases by one per time step as each incoming node has degree  $m$ , which also means there are no nodes with degree less than  $m$ , and hence the equations used to model the evolution of quantities like  $N_k(t)$  are not valid for  $k < m$ . The resulting degree distribution has the form

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)},$$

which asymptotically converges in distribution to  $P(k) \sim k^{-3}$ .

The preferential attachment model of Barabasi and Albert has attracted a tremendous amount of scientific interest in the past few years, and consequently numerous modifications of the model have been introduced. For example, extensions of the model allow:

- Ties to appear and disappear between any pairs of vertices (the original formulation only considers the addition of ties between the incoming vertex and set of vertices already in existence).
- Vertices to be deleted either uniformly at random or based on their connectivity.
- The attachment probability  $\Pi(k_i)$  to be super-linear or sub-linear in degree, or to consist of several terms.
- Nodal attributes, such as the *attractiveness* (the propensity with which new ties form with the node) or *fitness* (the propensity with which established ties remain intact) of a node, and the attachment probability can incorporate these attributes in addition to degree.
- Edges to assume weights instead of  $\{0, 1\}$  binary values to codify connection strength between any pair of elements.

In the context of physician networks, a preferential attachment model could be used to examine the process of new physicians seeking colleagues to ask for advice upon joining a medical organization, such as a hospital. Under the preferential

attachment hypothesis, new physicians would be more likely to form ties with and thus seek advice from popular established physicians or physicians in the same cohort (e.g., Medical school or residency program).

### Social Network Models

The class of models known as *network evolution models* can be defined via three properties: (i) the models incorporate a set of stochastic attachment rules which determine the evolution of the network structure explicitly on a time-step-by-time-step basis; (ii) the network evolution starts from an empty network consisting of nodes only, or from a small seed network possessing arbitrary structure; and (iii) the models incorporate a stopping criterion, which for growing network models is typically in the form of the network size reaching a predetermined value, and for dynamical (nongrowing) network models the convergence of network statistics to their asymptotic values. Many network evolution models do not reference intrinsic properties or attributes of nodes, and in this sense they are similar to the various implementations of preferential attachment models that do not postulate node-specific fitness or attractiveness.

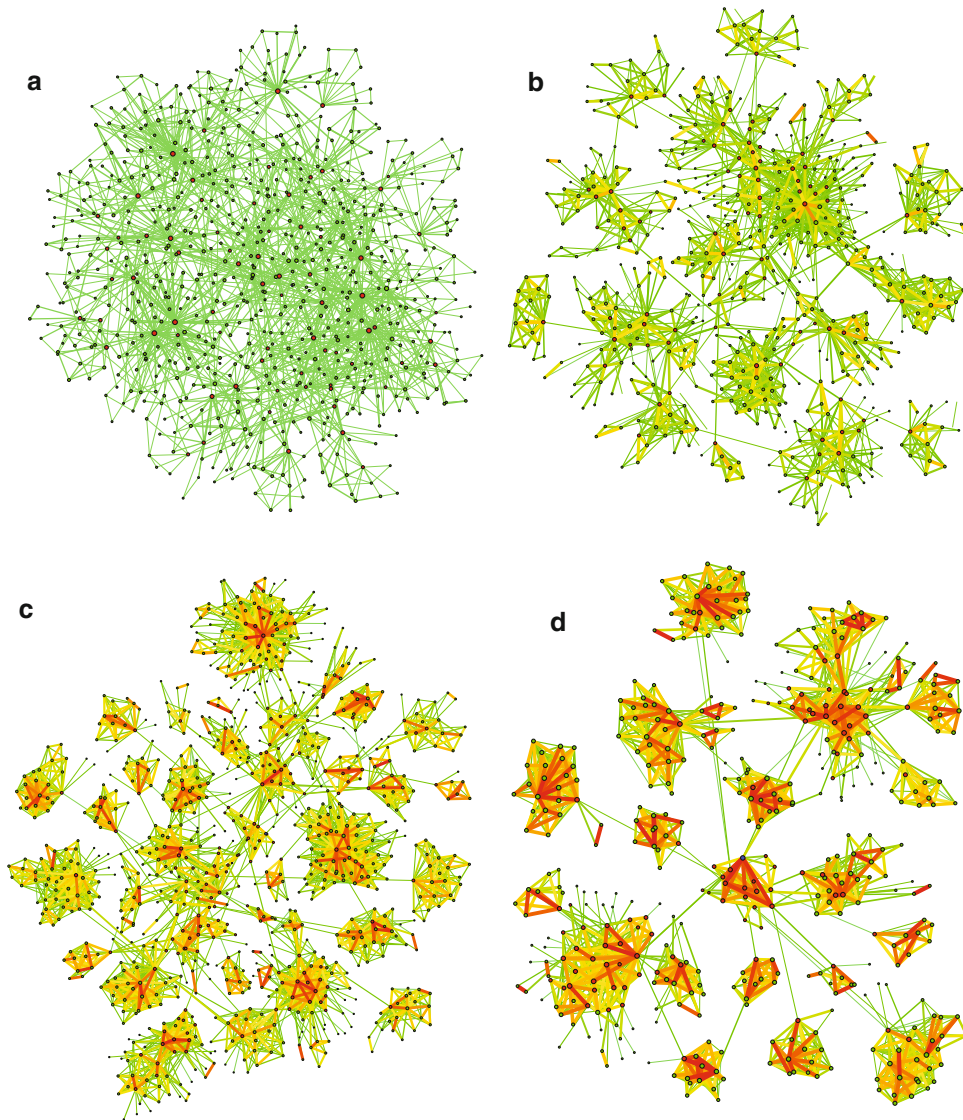
Most network evolution models that are intended to model social networks employ some variants of focal closure and cyclic closure (see, e.g., Kossinets and Watts (2006)). *Focal closure* refers to the formation of ties between individuals based on shared foci, which in a medical context could correspond to a group of doctors who practice in a particular hospital (the focus). The concept of shared foci in network science is analogous to homophily in social network analysis. More broadly, ties could represent any interest or activity that connects otherwise unlinked individuals. In contrast, *cyclic closure* refers to the idea of forming new ties by navigating and leveraging one's existing social ties, a process that results in a cycle in the underlying network. Because the network is nondirected, the term cycle is used interchangeably with closure. This differs from when the network is directional and a cycle is a specific form of closure, with transitivity being another form. *Triadic closure*, which is the

special case of cyclic closure involving just three individuals, refers to the process of getting to know friends of friends, leading to the formation of a closed triad in the nondirected network. Most social networks are expected to (i) have skewed and fat-tailed degree distributions, (ii) be assortatively mixed (high-degree individuals are connected to high-degree individuals), (iii) be highly clustered, and (iv) possess the small-world property (average shortest path lengths are short, or more precisely, scale as  $\log(N)$ ), and (v) exhibit community structure.

The models by Davidsen et al. (2002) and Marsili et al. (2004) exemplify dynamic (non-growing) network evolution models for social networks. Both have a mechanism that starts by selecting a node  $i$  in the network uniformly at random. In the model of Davidsen et al., if node  $i$  has fewer than two connections, it is connected to a randomly chosen node in the network; otherwise two randomly chosen neighbors of node  $i$  are connected together. In the model of Marsili et al., node  $i$  (regardless of its degree) is connected with probability  $\eta$  to a randomly chosen node in the network; then a second-order neighbor of node  $i$ , i.e., a friend's friend, is connected with probability  $\xi$  to node  $i$ . The first mechanism in each model, the random connection, emulates focal closure, because there are no nodal attributes signifying shared interests. The point is that the formation of these connections is not driven by the structure of existing connections but, from the point of view of network structure, is purely random. The second mechanism, the notion of triadic closure, is implemented in slightly different ways across the models. If these mechanisms were applied indefinitely, the result would be a fully connected network. To avoid this outcome, the models also delete ties at a constant rate, which makes it possible for network statistics of interest to reach stationary distributions. In the model of Davidsen et al., tie deletion is accomplished by choosing a node in the network uniformly at random, and then removing all of its ties with some probability; Marsili et al. accomplish the same phenomenon by selecting a tie uniformly at random, and then deleting it with probability  $\lambda$ . Growing network evolution models, such as those by Vázquez

(2003) and Toivonen et al. (2006), do not usually incorporate link deletion, but instead grow the network to a prespecified size, which obviates the need for link deletion.

Marsili et al. use extensive numerical simulations, as well as a master equation approach applied to a mean-field approximation of the model, to explore the impact of varying the probabilities  $\eta$  (global linking),  $\xi$  (neighborhood linking), and  $\lambda$  (link deletion) for average degree and average clustering coefficient. Consider a situation where the value of  $\xi$  (neighborhood linking) is increased while keeping the value of  $\lambda$  (link deletion) fixed. At first, for small values of  $\xi$ , components with more than two nodes are rare, and the network can be said to be in the sparse phase. Upon increasing the value of  $\xi$  up to a specific point, a large connected component emerges, and the value of the average degree suddenly jumps up. This point equals  $\xi_2/\lambda$  and is known as the critical point – it marks the beginning of the dense phase in the phase diagram of the system. As  $\xi$  is increased further, the network becomes more densely connected. Reversing the process by slowly decreasing the value of  $\xi$  identifies a range of values from  $\xi_1 \leq \xi \leq \xi_2$  where the largest connected component remains densely connected and the average degree remains high. Only when the value of  $\xi$  is decreased below a point denoted by  $\xi_1$  does the network “collapse” and reenter the sparse phase. This phenomenon, which demonstrates some of the connections between network science and statistical physics, is typical of first-order or discontinuous phase transitions in statistical physics, and it demonstrates how hysteresis, the effect of the system remembering its past state, can arise in networked systems. Although Markov dependence is a special case of hysteresis, its use is generally restricted to probabilistic models whereas hysteresis is typically aligned with nonlinear models of physical phenomena having a continuous state-space. From the social network point of view this means that the network can remain in a connected phase even if the rate of establishing new connections at the current rate would not be sufficient for getting the network to that phase in the first place. In more practical terms, this



**Fig. 6** Network structures produced by the model of Kumpula et al. by varying the reinforcement parameter as follows: (a)  $\delta = 0$ , (b)  $\delta = 0.1$ , (c)  $\delta = 0.5$ , and (d)  $\delta = 1$ . Figure adapted from Kumpula et al. (2007)

finding implies that it is possible to maintain a highly connected network with a relatively low “effort” (the  $\xi$  parameter in the model) once the network has been established, but that same low level of effort would not be sufficient for establishing the dense phase of network evolution in the first place. (The analogy in social network analysis is that the threshold for forming a (e.g.,) friendship is greater than that needed for it to remain intact.)

The model by Kumpula et al. (2007), which is another dynamical (nongrowing) network evolution model for social networks, implements cyclic closure and focal closure (see Fig. 6) in a manner similar to the models of Davidsen et al. and Marsili et al., but introduces a minor modification.

Unlike the previous models which produce binary networks with  $A_{ij} = \{0, 1\}$ , this model produced weighted networks with  $A_{ij} \geq 0$ . The main modification deals with the triadic closure

step, which here is implemented as a weighted two-step random walk. Starting from a randomly chosen node  $i$ ; this node chooses one of its neighbors  $j$  with probability  $w_{ij}/s_i$ , where  $s_i = \sum_j w_{ij}$  is the strength of node  $i$ , i.e., the sum of the edge weights connecting it to its neighbors. If node  $j$  has neighbors other than  $i$ , such a node  $k$  will be chosen with probability  $w_{jk}/(s_j - w_{ij})$ , where there is a requirement that  $k \neq i$ . The weights  $w_{ij}$  and  $w_{jk}$  on the edges just traversed will be increased by a value  $\delta$ . In addition, if there is a link connecting node  $i$  and node  $k$ , the weight  $w_{ik}$  on that link is similarly increased by  $\delta$ ; otherwise a new link is established between node  $i$  and  $k$  with  $w_{ik} = 1$ . When  $\delta = 0$ , there is no clear community structure present, but as the value of  $\delta$  is increased, very clear nucleation of communities takes place. This phenomenon occurs when  $\delta > 0$  because a type of positive feedback or memory gets imprinted on the network, which reinforces existing connections, and makes future transversal of those connections more likely. This is not unlike the models of cumulative advantage or preferential attachment discussed above, but now applies to individual links as opposed to nodes. If one inspects the community structure produced by the model, most of the strong links appear to be located within communities, whereas links between communities are typically weak. This type of structural organization is compliant with the so-called weak ties hypothesis, formulated in Granovetter (1973), which states, in essence, that the stronger the tie connecting two individuals, the higher the fraction of friends they have in common. Onnela et al. showed that a large-scale social network constructed from the cell phone communication records of millions of people was in remarkable agreement with the hypothesis – only the top 5% of ties in terms of their weight deviated noticeably from the prediction. The networks produced by the model of Kumpula et al. are clearly reminiscent of observed real-world social networks, and the inclusion of the tuning parameter  $\delta$  makes it straightforward to create networks with sparser or denser communities. The downside is that the addition of weights to the model appears to make it analytically intractable.

Nodal attribute models, in stark contrast to network evolution models, specify nodal attributes for each node, which could be scalar or vector valued. The probability of linkage between any two nodes is typically an increasing function of the similarity of the nodal attributes of the two nodes in consideration. This is compatible with the notion of homophily, the tendency for like to attract like. Nodal attribute models can also be interpreted as spatial models, where the idea is that each node has a specific location in a social space. The models by Boguñá et al. (2004) and Wong et al. (2006) serve as interesting examples. Nodal attribute models do not specify attachment rules at the level of the network, and in some sense can be seen as latent variable models for social network formation. These types of models have been studied less in the network science literature than network evolution models.

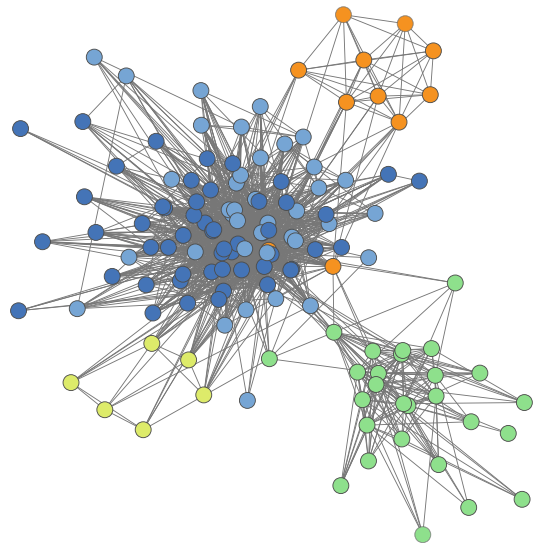
Clearly, nodal attribute models have a strong resemblance to models developed and studied in the social network literature that treat dyads as independent conditional on observed attributes of the individuals, other covariates, and various latent variables (individual-specific random effects in the case of the  $p_2$  model, categorical latent variables in the case of latent class models, continuous latent variables under the latent-space, and latent eigenmodels in section “[Latent Independence Approaches](#)”). Unlike network science, work on such models in the social network literature has been more prominent than work on network evolution. A difference in the approach of some nodal attribute models and social network models is that the former may use specific rules for determining whether a tie is expected, such as a threshold function (in a sense emulating formal decision making), whereas the latter rewards values of parameters that make the model most consistent with the observed network(s).

## Network Communities

Many network characteristics are either microscopic or macroscopic in nature; the value of a microscopic characteristic depends on local network structure only, whereas the value of a

macroscopic characteristic depends on the structure of the entire network. Node degree is an example of a microscopic quantity: the degree of a node depends only on the number of its connections. In contrast, network diameter, the longest of all pairwise shortest paths in the network, can change dramatically by the addition (or removal) of even a very small number of links anywhere in the network. For example, a  $k$ -cycle consists of  $k$  nodes connected by  $k$  links such that a cycle is formed with each node connected to precisely two nodes. The diameter of such a network is  $\lfloor k/2 \rfloor$ , where the floor function  $\lfloor x \rfloor$  maps a real number  $x$  to the largest previous integer, such that for an even  $n$  it follows that  $\lfloor n/2 \rfloor = n/2$ . For large values of  $n$ , adding just a few links quickly brings down the value of network diameter. There is a third, intermediate scale that lies between the microscopic and macroscopic scales which is often known as the *mesoscopic* scale. For example, a  $k$ -clique could justifiably be called a mesoscopic object (especially if  $k$  is large). Another type of mesoscopic structure is that of a network community, which can be loosely defined as a set of nodes that are densely connected to each other but sparsely connected to other nodes in the network (but not to the extent of resulting in distinct components).

There has been considerable interest especially in the physics literature focusing on how to define and detect such communities, and several review papers cover the existing methods (Porter et al. 2009; Fortunato 2010; Newman 2012). The motivation behind many of these efforts is the idea that communities may correspond to functional units in networks, such as unobserved societal structures. The examples range from metabolic circuits within cells (Guimera and Nunes Amaral 2005) to tightly knit groups of individuals in social networks (Newman and Girvan 2004; Traud et al. 2012). The interested reader can consult the review articles on community detection methods (Porter et al. 2009; Fortunato 2010; Newman 2012) for more details. Another application is health care where, for instance, Landon et al. (2012) have deduced communities of physicians based on network ties representing them treating the same



**Fig. 7** Communities in a patient-sharing network of physicians. Each vertex corresponds to a physician, and a pair of physicians are connected with a tie if they share patients. The community assignment of each physician is indicated by the node color. In this case the “green” and “orange” communities are fairly distinct

patients within the same period of time. The clustering of physicians in communities is shown for one particular Hospital Referral Region (a health care market encompassing at least one major city where both cardiovascular surgical procedures and neurosurgery are performed) in the United States (Fig. 7).

One potential application of network science methods for community detection is in the area of health education and disease prevention (e.g., screening). Due to limited resources, it may not be possible to send materials or otherwise directly educate every member of the population. The partition of individuals into groups would facilitate a possibly more efficient approach whereby the communities are first studied to identify key individuals. Then a few key individuals in each community are trained and advised on mechanisms for helping the intervention to diffuse across the community. A general characteristic of interventions where such an approach might be useful are those where intensive training is required to be effective and where delegation of resources through passing on knowledge or advice is possible.

## Modularity Maximization

A number of network community detection methods define communities implicitly via an appropriately chosen quality function. The underlying idea is that a given network can be divided into a large number of partitions, or subsets of nodes, such that each node belongs to one subset, and each such partition  $P$  has a scalar-valued quality measure associated with it, denoted by  $Q(P)$ . In principle one would like to enumerate all possible partitions and compute the value of  $Q$  for each of them, and the network communities would then be identified as the partition (or possibly partitions) with the highest quality. In practice, however, the number of possible partitions is exceedingly large even for relatively small networks, and therefore heuristics are needed to optimize the value of  $Q$ . Community detection methods based on quality function optimization therefore have two distinct components, which are the functional form of the quality function  $Q$ , and the heuristic used for navigating a subset of partitions over which  $Q$  is maximized.

The most commonly used optimization-based approach to community detection is modularity maximization, where modularity is one possible choice for the quality function  $Q$ ; in statistical terminology, modularity maximization would be regarded as a nonparametric procedure due to the fact that no distributional nor functional form assumptions are relied upon. There are many variants of modularity, but here the focus is on the original formulation by Newman and Girvan (Newman and Girvan 2003, 2004; Newman 2006). Modularity can be seen as a measure that characterizes the extent of homophily or assortative mixing by class membership, and one way to derive it is by considering the observed and expected numbers of connections between vertices of given classes, where the class of vertex  $i$  is given by  $c_i$ . The following derivation follows closely that of Newman (2010), although other derivations, based, for example, on dynamic processes, are also available.

We start by considering the observed number of edges between vertices of the same class, which is given by  $\frac{1}{2} \sum_{i,j} A_{ij} \delta(c_i, c_j)$ , where  $\delta(\cdot, \cdot)$  is the

Kronecker delta, and the factor  $1/2$  prevents double-counting vertex pairs. To obtain the expected number of edges between vertices of the same class, cut every edge in half, resulting in two stubs per edge, and then connect these stubs at random. For a network with  $m$  edges, there are a total of  $2m$  such stubs. Consider one of the  $k_i$  stubs connected to vertex  $i$ . This particular stub will be connected at random to vertex  $j$  of degree  $k_j$  with probability  $k_j/2m$ , and since vertex  $i$  has  $k_i$  such stubs, the number of expected edges between vertices  $i$  and  $j$  is  $k_i k_j/2m$ . The expected number of edges falling between vertices of the same class is now  $\frac{1}{2} \sum_{i,j} \frac{k_i k_j}{2m} \delta(c_i, c_j)$ . The difference between the observed and expected number of within class ties is therefore  $\frac{1}{2} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$ . Given that the number of edges varies from one network to the next, it is convenient to deal with the fraction of edges as opposed to the number of edges, which is easily obtained by dividing the expression by  $m$ , resulting in

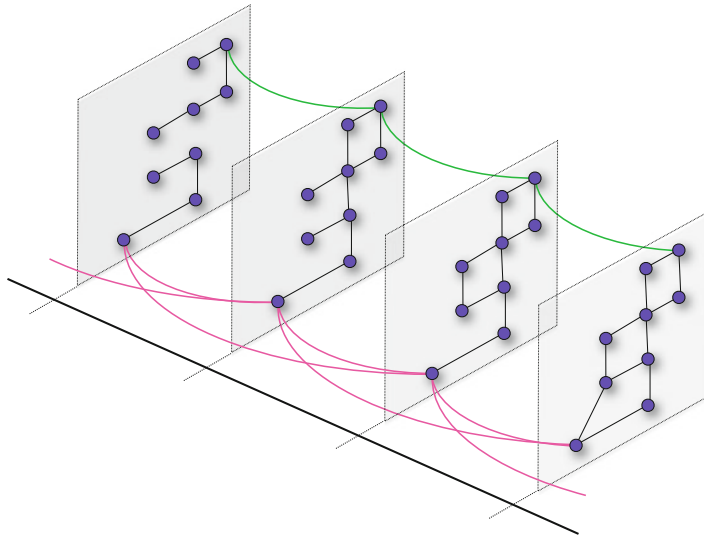
$$Q_M(P) = \frac{1}{2m} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j).$$

The assignment  $\mathcal{P}$  of nodes into classes that maximizes modularity  $Q_M(P)$  is taken as the optimal partition and identifies the assignment of nodes into network communities. Note that modularity can be easily generalized from binary networks to weighted networks, in which case  $k_i$  stands for the strength (sum of all adjacent edge weights) of node  $i$ , and  $m$  is the total weight of the edges in the network.

The expression for modularity has an interesting connection to spin models in statistical physics. In a so-called infinite range  $q$ -state Potts model, each of the  $N$  particles can be in one of  $q$  states called spins, and the interaction energy between particles  $i$  and  $j$  is  $-J_{ij}$  if they are in the same state and zero if they are not in different states. The energy function of the system, known as its Hamiltonian, is given by the sum over all of the pairwise interaction energies in the system

$$H(\{\sigma\}) = - \sum_{i,j} J_{ij} \delta(\sigma_i, \sigma_j),$$





**Fig. 8** Schematic of a multislice network. Each slice represents a network encoded by the adjacency tensor  $A_{ijs}$ , where subscripts  $i$  and  $j$  are used to index the nodes and subscript  $s$  is used to index the slices. Each node is coupled to itself in the other slices, and the structure of this coupling, encoded by the  $C_{jrs}$  tensor, depends on whether the slices correspond to snapshots taken at different times (time-dependent network), to communities detected at different resolution levels (multiscale network), or to a network consisting of multiple types of interactions (multiplex network). For time-dependent and multiscale

networks, the slice-to-slice coupling extends for each node a tie to itself across neighboring slices only as exemplified for the node in the *upper right corner* of the slices; for multiplex networks, the slice-to-slice coupling extends a tie from each node to itself in all the slices as exemplified for the node in the *lower left corner*. Whatever the form of this coupling, it is applied the same way to each node, although for visual clarity the slice-to-slice couplings are shown just for two nodes. Figure adapted from Mucha et al. (2010)

where  $\sigma_l$  indicates the spin of particle  $l$  and  $\{\sigma\}$  denotes the configuration of all  $N$  spins. Finding the minimum energy state (the ground state) of the system corresponds to finding  $\{\sigma\}$  such that  $H(\{\sigma\})$  is minimized. The states of the particles (spins) correspond to community assignments of nodes in the network problem, and minimizing  $H(\{\sigma\})$  is mathematically identical to maximizing modularity  $Q_M(\mathcal{P})$ . In the physical system, depending on the interaction energies, the spins seek to align with other spins (interact ferromagnetically) or they seek to have different orientations (interact antiferromagnetically). In the community detection problem, two nodes seek to be in the same community if they are connected by an edge that is stronger than expected; otherwise they seek to be in different communities. This correspondence between the two problems has enabled the application of computational techniques developed for the study of spin systems

and other physical systems to be applied to modularity optimization and, more broadly, to the optimization of other quality functions. Simulated annealing, greedy algorithms, and spectral methods serve as examples of these methods. More details and references are available in community detection review articles (Porter et al. 2009; Fortunato 2010).

Although there are several extensions of modularity maximization, only one such generalization is described here. Mucha et al. developed a generalized framework of network quality functions that allow the study of community structure of arbitrary multislice networks (see Fig. 8), which are combinations of individual networks coupled through links that connect each node in one slice to the same node in other slices (Mucha et al. 2010). This framework allows studies of community structure in time-dependent, multi-scale, and multiplex networks. Much of the work

in the area of community detection is motivated by the observation that the behavior of dynamical processes on networks is driven or constrained by their community structure. The approach of Mucha et al. is based on a reversal of this logic, and it introduces a dynamical process on the network, and the behavior of the dynamical process is used to identify the (structural) communities. The outcome is a quality function

$$Q_{\text{MS}}(\mathcal{P}) = \frac{1}{2\mu} \sum_{i,j,s,r} \left[ \left( A_{ijs} - \gamma_s \frac{k_{is}k_{js}}{2m_s} \right) \delta_{sr} + C_{jsr} \right] \delta(c_{is}, c_{jr}),$$

where  $A_{ijs}$  encodes the node-to-node couplings within slices and  $C_{jrs}$  encodes the node-to-node couplings across slices that are usually set to a uniform value  $\omega$ ;  $m_s$  is the number (or weight) of ties within slice  $s$  and  $\mu$  is the weight of all ties in the network, both those located within slices and those placed across slices;  $\gamma_s$  is a resolution parameter that controls the scale of community slices separately for each slice  $s$ . The standard modularity quality function uses  $c_i$  to denote the community assignment of node  $i$ , but in the multi-slice context two indices are needed, giving rise to the  $c_{is}$  terms, where the subscript  $i$  is used to index the node in question and the subscript  $s$  to index the slice. The outcome of minimizing  $Q_{\text{MS}}$ , which can be done with the same heuristics as minimization of the standard modularity  $Q_{\text{M}}$ , is a matrix  $\mathbf{C}$  that consists of the community assignments  $c_{is}$  of each node in every slice.

The multislice framework can handle any combination of time-dependent, multiscale, and multiplex networks. For example, the slices in Fig. 8 could correspond, say, to a longitudinal friendship network of a cohort of college students, each slice capturing the offline friendships of the students in each year. If data on the online friendships of the students were also available, corresponding to a different type of friendship, one could then introduce a second stack of four slices encoding those friendships. The four offline slices and the four online slices form a multiplex system, and they would be coupled accordingly. One could further introduce multiple resolution scales, and if one

was interested in examining the community structure of the students at three different scales using, say,  $\gamma_s \in \{0.5, 1, 2\}$ , this would result in a three-fold replication of the  $4 \times 2$  slice array with each of the three layers having a distinct value for  $\gamma_s$ . Taken together, this would lead to a three-dimensional  $4 \times 2 \times 3$  array of slices.

### Clique Percolation

Cliques are (usually small) fully connected subgraphs, and a nondirected  $k$ -clique is a complete subgraph consisting of  $k$  nodes connected with  $k(k-1)/2$  links. In materials science the term percolation refers to the movement of fluid through porous materials. However, in mathematics and statistical physics, the field of percolation theory considers the properties of clusters on regular lattices or random networks, where each edge may be either open or closed, and the clusters correspond to groups of adjacent nodes that are connected by open edges. The system is said to percolate in the limit of infinite system size if the largest component, held together by open edges, occupies a finite fraction of the nodes. The method of  $k$ -clique percolation in Palla et al. (2005) combines cliques and percolation theory, and it relies on the empirical observation that network communities seem to consist of several small cliques that share many of their nodes with other cliques in the same community. In this framework, cliques can be thought of as the building blocks of communities. A  $k$ -clique community is then defined as the union of all adjacent  $k$ -cliques, where two  $k$ -cliques are defined to be adjacent if they share  $k-1$  nodes. One can also think about “rolling” a  $k$ -clique template from any  $k$ -clique in the graph to any adjacent  $k$ -clique by relocating one of its nodes and keeping the other  $k-1$  nodes fixed. A community, defined through the percolation of such a template, then consists of the union of all subgraphs that can be fully explored by rolling a  $k$ -clique template. As  $k$  becomes larger, the notion of a community becomes more stringent, and values of  $k = 3, \dots, 6$  tend to be most appropriate because larger values become unwieldy. The special case of  $k = 2$  reduces to bond (link) percolation and  $k = 1$  reduces to site (node) percolation.

The  $k$ -clique percolation algorithm is an example of a local community-finding method. One obtains a network's global community structure by considering the ensemble of communities obtained by looping over all of its  $k$ -cliques. Some nodes might not belong to any community (because they are never part of any  $k$ -clique), and others can belong to several communities (if they are located at the interface between two or more communities). The nested nature of communities is recovered by considering different values of  $k$ , although  $k$ -clique percolation can be too rigid because focusing on cliques typically causes one to overlook other dense modules that are not quite as tightly connected.

The advantage of  $k$ -clique percolation is that it provides a successful way to consider community overlap. Allowing the detection of network communities that overlap is especially appealing in the social sciences, as people may belong simultaneously to several communities (colleagues, family, friends, etc.). However, the case can be made that it is the underlying interactions that are different, and one should not combine interactions that are of fundamentally different types. In statistics, this is analogous to using composite variables or scales that combine multiple items in (e.g.,) health surveys or questionnaires. If the nature of the interactions is known, the system might be more appropriately described as a multiplex network, where one tie type encodes professional interactions, another tie type corresponds to personal friendships, and a third tie type captures family memberships. The multi-slice framework discussed above is able to accommodate memberships in multiple communities as long as distinct interaction types are encoded with distinct (multiplex) ties.

### Comparison to Social Network Approaches to "Community Detection"

The latent class models in section "[Latent Independence Approaches](#)" partitions the actors in a network into disjoint groups that can be thought of as communities. The clustering process can be thought of as a search for structural equivalence in that individuals are likely to be included in the same community if the network around them is

similar to that of their neighbors. The criteria for judging the efficacy of the partition of nodes into communities is embedded in the statistical model implied for the network and as such is a balance between all of the terms in the model. This contrasts a nonmodel-based objective function such as modularity which focuses on maximizing in some sense the ratio of density of ties within and between communities. To illustrate the difference, consider a  $k$ -star. The greater the value of  $k$ , the greater the discrepancy in the degree of the actors. Therefore, if  $k$ -stars occur frequently, the members of the same  $k$ -star are likely to be included in the same group by the latent class model but, due to the difference in degree, are unlikely to be grouped under modularity maximization. However, an advantage of the network science approach is that results are likely to be more robust to model misspecifications than under the social network approach.

In the future it is possible to imagine a bridging of the two approaches to community detection. For example, a model for the network, or the component of the model involving the key determinants of network ties, could be incorporated in the modularity function in (7.1). Depending on the specification, the result might be a weighted version of modularity in which a higher penalty is incurred if individuals with similar traits – or in structurally equivalent positions with respect to  $k$ -stars, triadic closure or other local network configurations – are included in different communities than if individuals with different traits are in different communities. However, to the best of the author's knowledge, such a procedure is not available.

---

## Part IV: Discussion and Glossary

In this chapter, the dual fields of social networks and network science have been described, with particular focus on sociocentric data. Both fields are growing rapidly in methodological results and the breadth of applications to which they are applied.

In health applications, social network methods for evaluating whether individuals' attributes

spread from person-to-person across a population (social influence) and for modeling relationship or tie status (social selection) have been described. Models of relationship status have not been applied as frequently in health applications, where focus often centers on the patient. However, Keating et al. (2007) is a notable exception. Due to the ever-growing availability of data, the interest in peer effects, and the need to design support mechanisms, the role of social network analysis in health care and medicine is likely to undergo continued growth in the future.

A novel feature of this chapter is the attention given to network science. Although network science is descriptively inclined and thus is removed from mainstream translational medical research seeking to identify causes of medical outcomes, the increasing availability of complex systems data provides an opportunity for network science to play a more prominent role in medical research in the future. For example, Barabasi and others have created a Human Disease Network by connecting all hereditary diseases that share a disease-causing gene (Goh et al. 2007). In other work, they created a Phenotypic Disease Network (PDN) as a map summarizing phenotypic connections between diseases (Hidalgo et al. 2009). These networks provided important insights into the potential common origins of different diseases, whether diseases progress through cellular functions (phenotypes) associated with a single diseased (mutated) gene or with other phenotypes, and whether patients affected by diseases that are connected to many other diseases tend to die sooner than those affected by less connected diseases. Such work has the potential to provide insights into many previously untested hypotheses about disease mechanisms.

For example, they may ultimately be helpful in designing “personalized treatments” based on the network position held by an individual’s combined genetic, proteomic, and phenotypic information. In addition, they may suggest conditions for which treatments found to be effective on another condition might also be tried.

There are several important topics that have not been discussed, notably including network sampling. In gathering network data, adaptive methods such as link-tracing designs are often used to identify individuals more likely to know each other and thus to have formed a relationship with other sampled individuals than in a random-probability design. Link-tracing and other related designs are often used to identify hard-to-reach populations (Thompson and Seber 1996; Thompson and Frank 2000; Thompson 2006). However, the sampling probabilities corresponding to link-tracing designs may be difficult to evaluate (generally requiring the use of simulation), and it may not be obvious how they should be incorporated in the analysis. The development of statistical methods that account for the sample design in the analysis of social network data has lagged behind the designs themselves. However, recently progress has been made on statistical inference for sampled relational network data (Handcock et al. 2010).

In the future it is likely that more bridges will form between the social network and the network science fields with models or methods developed in one field used to solve problems in the other. Furthermore, as these two fields become more entwined, it is likely that they will also become more prominent in the solution to important problems in medicine and health care.

**Acknowledgments** The time and effort of Dr. O'Malley and Dr. Onnela on researching and developing this chapter was supported by NIH/NIA grant P01 AG031093 and Robert Wood Johnson Award #58729. The authors thank Mischa Haider, Brian Neelon, and Bruce E Landon for reviewing an early draft of the manuscript and providing several useful comments and suggestions.

---

## Glossary of Terms

To help readers familiar with social networks understand the network science component of the chapter and conversely for readers familiar with network science to understand the social network component, the following glossary contains a comprehensive list of terms and definitions.

## Terms Used in Social Networks

1. Social network: A collection of actors (referred to as actors) and the (social) relationships or ties linking them.
2. Relationship, Tie: A link or connection between two actors.
3. Dyad: A pair of actors in a network and the relationship(s) between them, two relationships per measure for a directed network, one relationship per measure for an undirected network.
4. Triad: A triple of three actors in the network and the relationships between them.
5. Scale or valued relationship: A nonbinary relationship between two actors (e.g., the level of a trait). We focused on binary relationships in the chapter.
6. Directed network: A network in which the relationship from actor  $i$  to actor  $j$  need not be the same as that from actor  $j$  to actor  $i$ .
7. Nondirected network: A network in which the state of the relationship from actor  $i$  to actor  $j$  equals the state of the relationship from actor  $j$  to actor  $i$ .
8. Sociocentric network data: The complete set of observations on the  $n(n - 1)$  relationships in a directed network, or  $n(n - 1)/2$  relationships in an undirected network, with  $n$  actors.
9. Collaboration network: A network whose ties represent the actors' joint involvement on a task (e.g., work on a paper) or a common experience (e.g., treating the same episode of health care for a patient).
10. Bipartite: Relationships are only permitted between actors of two different types.
11. Unipartite: Relationships are permitted between all types of actors.
12. Social contagion, Social influence, Peer effects: Terms used to describe the phenomenon whereby an actor's trait changes due to their relationship with other actors and the traits of those actors.
13. Mutable trait: A characteristic of an actor that can change state.
14. Social selection: The phenomena whereby the relationship status between two actors depends on their characteristics, as occurs with homophily and heterophily.
15. Homophily: A preference for relationships with actors who have similar characteristics. Popularly referred to as "birds of a feather flock together."
16. Heterophily: A preference for relationships with actors who have different characteristics. Popularly referred to as "opposites attracting."
17. In-degree, Popularity: The number of actors who initiated a tie with the given actor.
18. Out-degree, Expansiveness, Activity: The number of ties the given actor initiates with other actors.
19.  $k$ -star: A subnetwork in which the focal actor has ties to  $k$  other actors.
20.  $k$ -cycle: A subnetwork in which each actor has degree 2 that can be arranged as a ring (i.e., a  $k$ -path through the actors returns to its origin without backtracking. For example, the ties A-B, B-C, and C-A form a three-cycle.
21.  $k$  degrees of separation: Two individuals linked by a  $k$ -path ( $k - 1$  intermediary actors) that are not connected by any path of length  $k - 1$  or less.
22. Density: The overall tendency of ties to form in the network. A descriptive measure is given by the number of ties in the network divided by the total number of possible ties.
23. Reciprocity: The phenomena whereby an actor  $i$  is more likely to have a tie with actor  $j$  if actor  $j$  has a tie with actor  $i$ . Only defined for directed networks.
24. Clustering: The tendency of ties to cluster and form densely connected regions of the network.
25. Closure: The tendency for network configurations to be closed.
26. Transitivity: The tendency for a tie from individual A to individual B to form if ties from individual A to individual C and from individual C to individual B exist. A form of triadic closure commonly stated as "a friend of a friend is a friend." Reduces to general triadic closure in an undirected network.
27. Centrality: A dimensionless measure of an actor's position in the network. Higher values

- indicate more central positions. There are numerous measures of centrality. Four common ones are degree, closeness, betweenness, and eigenvalue centrality. Degree and eigenvalue centrality are extremes in that degree centrality is determined solely from an actor's degree (it is internally focused) while eigenvalue centrality is based on the centrality of the actors connected to the focal actor (it is externally focused).
28. Structural balance: A theory which suggests actors seek balance in their relationships; for example, if A likes B and B likes C then A will endeavor to like C as well to keep the system balanced. Thus, the existence of transitivity is implied by structural balance.
  29. Structural equivalence: The network configuration (arrangement of ties) around one actor is similar to that of another actor. Even though actors may not be connected, they can still be in structurally similar situations.
  30. Structural power: An actor in a dominant position in the network. Such an actor may be one in a strategic position, such as the only bridge between otherwise distinct components.
  31. Network component: A subset of actors having no ties external to themselves.
  32. Graph theory: The mathematical basis under which theoretical results for networks are derived and empirical computations are performed.
  33. Digraph: A graph in which edges can be bidirectional. Unlike social networks, digraphs can contain self-ties. Graphs lie in two-dimensional space.
  34. Hypergraph: A graph in dimension three or higher.
  35. Maximal subset: A set of actors for whom all ties are intact in a binary network (i.e., has density 1.0). If the set contains  $k$  actors, the maximal subset is referred to as a  $k$ -clique.
  36. Scalar, vector, matrix: Terms from linear and abstract algebra. A scalar is a  $1 \times 1$  matrix, a vector is a  $k \times 1$  matrix, and a matrix is  $k \times p$ , where  $k, p > 1$ .
  37. Adjacency matrix: A matrix whose off-diagonal elements contain the value of the relationship from one actor to another. For example, element  $ij$  contains the relationship from actor  $i$  to actor  $j$ . The diagonal elements are zero by definition.
  38. Matrix transpose: The operation whereby element  $ij$  is exchanged with element  $ji$  for all  $i, j$ .
  39. Row stochastic matrix: A matrix whose rows sum to 1 and contain nonnegative elements. Thus, each row represents a probability distribution of a discrete-valued random variable.
  40. Random variable: A variable whose value is not known with certainty. It can relate to an event or time period that is yet to occur, or it can be a quantity whose value is fixed (i.e., has occurred) but is unknown.
  41. Parametric: A term used in statistics to describe a model with a specific functional form (e.g., linear, quadratic, logarithmic, exponential) indexed by unknown parameters or an estimation procedure that relies on specification of the complete distribution of the data.
  42. Nonparametric: A model or estimation procedure that makes no assumption about the specific form of the relationship between key variables (e.g., whether the predictors have linear or additive effects on the outcome) and does not rely upon complete specification of the distribution of the data for estimation.
  43. Outcome, Dependent variable: The variable considered causally dependent on other variables of interest. This will typically be a variable whose value is believed to be caused by other variables.
  44. Independent, Predictor, Explanatory variable, Covariate: A variable believed to be a cause of the outcome.
  45. Contextual variable: A variable evaluated on the neighbors of, or other members of a set containing, the focal actor. For example, the proportion of females in a neighboring county, the proportion of friends with college degrees.
  46. Interaction effect: The extent to which the effect of one variable on the outcome varies across the levels of another variable.
  47. Endogenous variable: A variable (or an effect) that is internal to a system.

- Predictors in a regression model that are correlated with the unobserved error are endogenous; they are determined by an internal as opposed to an external process. By definition outcome variables are endogenous.
48. Exogenous variable: A variable (or an effect) that is external to the system in that its value is not determined by other variables in the system. Predictors that are independent of the error term in a regression model are exogeneous.
  49. Instrumental variable (IV): A variable with a non-null effect on the endogeneous predictor whose causal effect is of interest (the “treatment”) that has no effect on the outcome other than that through its effect on treatment. Often-used sufficient conditions for the latter are that the IV is (i) marginally independent of any unmeasured confounders and (ii) conditionally independent of the outcome given the treatment and any unmeasured confounders. In an IV analysis a set of observed predictors may be conditioned on as long as they are not effects of the treatment and the IV assumptions hold conditional on them. While subject to controversy, IV methods are one of the only methods of estimating the true (causal) effect of an endogeneous predictor on an outcome.
  50. Linear regression model: A model in which the expected value of the outcome (or dependent variable) conditional on one or more predictors (or explanatory variables) is a linear combination of the predictors (an additive sum of the predictors multiplied by their regression coefficients) and an unobserved random error.
  51. Longitudinal model: A model that describes variation in the outcome variable over time as a function of the predictors, which may include prior (i.e., lagged) values of the outcome. Observations are typically only available at specific, but not necessarily equally spaced, times. Longitudinal models make the direction of causality explicit. Therefore, they can distinguish between the association between the predictors and the outcome and the effect of a change in the predictor on the change in the outcome.
  52. Cross-sectional model: A model of the relationship between the values of the predictors and outcomes at a given time. Because one cannot discern the direction of causality, cross-sectional models are more difficult to defend as causal.
  53. Stochastic block model: A conditional dyadic independence model in which the density and reciprocity effects differ between blocks defined by attributes of the actors comprising the network. For example, blocks for gender accommodate different levels of connectedness and reciprocity for men and women.
  54. Logistic regression: A member of the exponential family of models that is specific to binary outcomes. It utilizes a link function that maps expected values of the outcome onto an unrestricted scale to ensure that all predictions from the model are well-defined.
  55. Multinomial distribution: A generalization of the binomial distribution to three or more categories. The sum of the probabilities of each category equals 1.
  56. Exponential random graph model: A model in which the state of the entire network is the dependent variable. Provides a flexible approach to accounting for various forms of dependence in the network. Not amenable to causal modeling.
  57. Degeneracy: An estimation problem encountered with exponential random graph models in which the fitted model might reproduce observed features of the network on average but each actor draw bears no resemblance to the observed network. Often degenerate draws are empty or complete graphs.
  58. Latent distance model: A model in which the status of dyads are independent conditional on the positions of the actors, and thus the distance between them, in a latent social space.
  59. Latent eigenmodel: A model in which the status of dyads are independent conditional on the product of the (weighted) latent positions of the actors in the dyad.

60. Latent variable: An unobserved random variable. Random effects and pure error terms are latent variables.
61. Latent class: An unobserved categorical random variable. Actors with the same value of the variable are considered to be in the same latent class.
62. Factor analysis: A statistical technique used to decompose the correlation (or covariance) matrix of a set of random variables into groups of related items.
63. Generalized estimating equation (GEE): A statistical method that corrects estimation errors for dependent observations without necessarily modeling the form of the dependence or specifying the full distribution of the data.
64. Random effect: A parameter for the effect of a unit (or cluster) that is drawn from a specified probability distribution. Treating the unit effects as random draws from a common probability distribution allows information to be pooled across units for the estimation of each unit-specific parameter.
65. Fixed effect: A parameter in a model that reflects the effect of an actor belonging to a given unit (or cluster). By virtue of modeling the unit effects as unrelated parameters, no information is shared between units and so estimates are based only on information within the unit.
66. Ordinary least squares: A commonly used method for estimating the parameters of a regression model. The objective function is to minimize the squared distance of the fitted model to the observed values of the dependent variable.
67. Maximum likelihood: A method of estimating the parameters of a statistical model that typically embodies parametric assumptions. The procedure is to seek the values of the parameters that maximize the likelihood function of the data.
68. Likelihood function: An expression that quantifies the total information in the data as a function of model parameters.
69. Markov chain Monte Carlo: A numerical procedure used to fit Bayesian statistical models.
70. Steady state: The state-space distribution of a Markov chain describes the long-run proportion of time the random variable being modeled is in each state. Often Markov chains iterate through a transient phase in which the current state of the chain depends less and less on the initial state of the chain. The steady state phase occurs when successive samples have the same distribution (i.e., there is no dependence on the initial state).
71. Colinearity: The correlation between two predictors after conditioning on the other observed predictors (if any). When predictors are colinear, distinguishing their effects is difficult, and the statistical properties of the estimated effects are more sensitive to the validity of the model.
72. Normal distribution: Another name for the Gaussian distribution. Has a bell-shaped probability density function.
73. Covariance matrix: A matrix in which the  $ij$ th element contains the covariance of items  $i$  and  $j$ .
74. Absolute or Geodesic distance: The total distance along the edges of the network from one actor to another.
75. Cartesian distance: The distance between two points on a two-dimension surface or grid. Adheres to Pythagoras Theorem.
76. Count data: Observations made on a variable with the whole numbers (0, 1, 2, ...) as its state space.
77. Statistical inference: The process of establishing the level of certainty of knowledge about unknown parameters (or hypothesis) from data subject to random variation, such as when observations are measured imperfectly with no systematic bias or a sample from a population of interest is used to estimate population parameters.
78. Null model: The model of a network statistic typically represents what would be expected if the feature of interest was nonexistent (effect equal to 0) or outside the range of interest.
79. Permutation test: A statistical test of a null hypothesis against an alternative implemented by randomly reshuffling the labels (i.e., the



subscripts) of the observations. The significance level of the test is evaluated by resampling the observed data 50–100 times and computing the proportion of times that the test is rejected.

## Terms Used in Network Science

1. Network science: The approach developed from 1995 onwards mostly within statistical physics and applied mathematics to study networked systems across many domains (e.g., physical, biological, social, etc). Usually focuses on very large systems; hence, theoretical results derived in the thermodynamic limit are good approximations to real-world systems.
2. Thermodynamic limit: In statistical physics refers to the limit obtained for any quantity of interest as system size  $N$  tends to infinity. Many analytical results within network science are derived in this limit due to analytical tractability.
3. Statistical physics: The branch of physics dealing with many body systems where the particles in the system obey a fix set of rules, such as Newtonian mechanics, quantum mechanics, or any other rule set. As the number of bodies (particles) in a system grows, it becomes increasingly difficult (and less informative) to write down the equations of motion, a set of differential equations that govern the motion of the particles over time, for the system. However, one can describe these systems probabilistically. The word “statistical” is somewhat misleading as there is no statistics in the sense of statistical inference involved; instead everything proceeds from a set of axioms, suggesting that “probabilistic” might be a better term. Statistical physics, also called statistical mechanics, gives a microscopic explanation to the phenomena that thermodynamics explains phenomenologically.
4. Generative model: Most network models within network science belong to this category. Here one specifies the microscopic rules governing, for example, the attachment of new nodes to the existing network structure in models of network growth.
5. Cumulative advantage: A stylized modeling mechanism introduced by Price in 1976 to capture phenomena where “success breeds success.” Price applied the model to study citation patterns where power-law or power-law-like distributions are observed for the distribution of the number of citations and successfully reproduced by the model.
6. Polya urn model: A stylized sampling model in probability theory where the composition of the system, the contents of the urn, changes as a consequence of each draw from the urn.
7. Power law: Refers to the specific functional form  $P(x) \sim x^{-\alpha}$  of the distribution of quantity  $x$ . Also called Pareto distribution. See scale-free network.
8. Preferential attachment: A stylized modeling mechanism introduced by Barabasi and Albert in 1999 where the probability of a new node to attach itself to an existing node  $i$  of degree  $k_i$  is an increasing function of  $k_i$ ; in the case of linear preferential attachment, this probability is directly proportional to  $k_i$ . In short, the higher the degree of a node, the higher the rate at which it acquires new connections (increases its degree).
9. Weak ties hypothesis: A hypothesis developed by sociologist Mark Granovetter in his extremely influential 1973 paper “The strength of weak ties.” The hypothesis, in short, states the following: The stronger the tie connecting persons  $A$  and  $B$ , the higher the fraction of friends they have in common.
10. Modularity: Modularity is a quality-function used in network community detection, where its value is maximized (in principle) over the set of all possible partitions of the network nodes into communities. Standard modularity reads as  $Q = (2m)^{-1} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$  where  $c_i$  is the community assignment of node  $i$  and  $\delta$  is Kronecker delta; other quantities as defined in the text.
11. Rate equations: Rate equations, commonly used to model chemical reactions, are similar

- to master equations but instead of modeling the count of objects (e.g., number of nodes) in a collection of discrete states (e.g., the number of  $k$ -degree nodes  $N_k(t)$  for different values of  $k$ ), they are used to model the evolution of continuous variables, such as average degree, over time.
12. Master equations: Widely used in statistical physics, these differential equations model how the state of the system changes from one time point to the next. For example, if  $N_k(t)$  denotes the number of nodes of degree  $k$ , given the model, one can write down the equation for  $N_k(t+1)$ , i.e., the number of  $k$ -degree nodes at time  $t+1$ .
  13. Fitness or affinity or attractiveness: A node attribute introduced to incorporate heterogeneity in the node population in a growing network model. For example, in a model based on preferential attachment, this could represent the inherent ability of a node to attract new edges, a mechanism that is superimposed on standard preferential attachment.
  14. Community: A group of nodes in a network that are, in some sense, densely connected to other nodes in the community but sparsely connected to nodes outside the community.
  15. Community detection: The set of methods and techniques developed fairly recently for finding communities in a given network (graph). The number of communities is usually not specified a priori but, instead, needs to be determined from data.
  16. Critical point: The value of a control parameter in a statistical mechanical system where the system exhibits critical behavior: previously localized phenomena now become correlated throughout the system which at this point behaves as one single entity.
  17. Phase diagram: A diagram displaying the phase (liquid, gas, etc.) of the system as one or more thermodynamic control parameters (temperature, pressure, etc.) are varied.
  18. Phase transition: Thermodynamic properties of a system are continuous functions of the thermodynamic parameters within a phase; phase transitions (e.g., liquid to gas) happen between phases where thermodynamic functions are discontinuous.
  19. Network diameter: The longest of the shortest pairwise paths in the network, computed for each dyad (node pair).
  20. Hysteresis: The behavior of a system depends not only on its current state but also on its previous state or states.
  21. Quality function: Typically a real-valued function with a high-dimensional domain that specifies the "goodness" of, say, a given network partitioning. For example, given the community assignments of  $N$  nodes, which can be seen as a point in an  $N$ -dimensional hypercube, the standard modularity quality function returns a number indicating how good the given partitioning is.
  22. Dynamic process: Any process that unfolds on a network over time according to a set of prespecified rules, such as epidemic processes, percolation, diffusion, synchronization, etc.
  23. Slice: In the context of multislice community detection, refers to one graph in a collection of many within the same system, where a slice can capture the structure of a network at a given time (time-dependent slice), at a particular resolution level (multiscale slice), or can encode the structure of a network for one tie type when many are present (multiplex slice).
  24. Scale-free network: Network with a power-law (Pareto) degree distribution.
  25. Erdős-Rényi model: Also known as Poisson random graph (after the fact that the degree distribution in the model follows a Poisson distribution), Bernoulli random graph (after the fact that each edge corresponds to an outcome of a Bernoulli process), or the random graph (as the progenitor of all random graphs). Starting with a fixed set of  $N$  nodes, one considers each node pair in turn independently of the other node pairs and connects the nodes with probability  $p$ . Erdős and Rényi first published the model in 1959, although Solomonoff and Rapoport published a similar model earlier in 1951.
  26. Watts-Strogatz model: A now canonical model by Watts and Strogatz that was

introduced in 1998. Starting from a regular lattice structure characterized by high clustering and long paths, the model shows how randomly rewiring only a small fraction of edges (or, alternative, adding a small number of randomly placed edges) leads to a small-world characterized by high clustering and short paths. The model is conceptually appealing, and shows how to interpolate, using just one parameter, from a regular lattice structure in one extreme to an Erdős-Rényi graph in the other.

27. Mean-field approximation: Sometimes called the zero-order approximation, this approximation replaces the value of a random variable by its average, thus ignoring any fluctuations (deviations) from the average that may actually occur. This approach is commonly used in statistical physics.
28. Ensemble: A collection of objects, such as networks, that have been generated with the same set of rules, where each object in the ensemble has a certain probability associated with it. For example, one could consider the ensemble of networks that consists of six nodes and two edges, each begin equiprobable.

---

## References

- Airoldi EM, Fienberg SE, Xing EP. Mixed membership stochastic blockmodels. *J Mach Learn Res.* 2008;9:1981–2014.
- Anselin L. *Spatial econometrics: methods and models.* Dordrecht: Kluwer; 1988.
- Barabasi A-L, Albert R. Emergence of scaling in random networks. *Science.* 1999;286:509–12. <http://www.sciencemag.org/content/286/5439/509.abstract>
- Barabasi A-L, Albert R, Jeong H. Mean-field theory for scale-free random networks. *Phys A Stat Mech Appl.* 1999;272:173–87. <http://www.sciencedirect.com/science/article/pii/S0378437199002915>.
- Barnett ML, Landon BE, O'Malley AJ, Keating NL, Christakis NA. Mapping physician networks with self-reported and administrative data. *Health Serv Res.* 2011;46:1592–609.
- Barnett ML, Christakis NA, O'Malley AJ, Onnela J-P, Keating NL, Landon BE. Physician patient-sharing networks and the cost and intensity of care in US hospitals. *Med Care.* 2012a;50:152–60.
- Barnett ML, Keating NL, Christakis NA, O'Malley AJ, Landon BE. Reasons for referral among primary care and specialist physicians. *J Gen Intern Med.* 2012b;27:506–12.
- Berkman L, Glass T. Social integration, social methods, social support, and health. In: *Social epidemiology.* New York: Oxford University Press; 2000. p. 137–73.
- Boguñá M, Pastor-Satorras R, Díaz-Guilera A, Arenas A. Models of social networks based on social distance attachment. *Phys Rev E.* 2004;70:056122. <https://doi.org/10.1103/PhysRevE.70.056122>.
- Bonacich P. Power and centrality: a family of measures. *Am J Sociol.* 1987;92:1170–82.
- Borgatti S, Everett M. Network analysis of 2-mode data. *Soc Networks.* 1997;19:243–69.
- Breiger R. The duality of persons and groups. *Soc Forces.* 1974;53:181–90.
- Cartwright D, Harrary F. A generalization of Heider's theory. *Psychol Rev.* 1956;63:277–92.
- Centola D. Failure in complex social networks. *Math Sociol.* 2009;33:64–8.
- Choi D, Wolfe P, Airoldi E. Stochastic blockmodels with growing number of classes. Arxiv preprint. 2010; arXiv:1011.4644.
- Christakis N, Fowler J. The spread of obesity in a large social network over 32 years. *N Engl J Med.* 2007;357:370–9.
- Christakis NA, Fowler JH. Social contagion theory: examining dynamic social networks and human behavior. *Stat Med.* 2013;32:556–77.
- Coleman J, Katz E, Menzel H. The diffusion of innovations among physicians. *Sociometry.* 1957;20:253–70.
- Coleman J, Katz E, et al. Medical innovation: a diffusion study. Indianapolis: Bobbs-Merrill; 1966.
- Davidson J, Ebel H, Bornholdt S. Emergence of a small world from local interactions: modeling acquaintance networks. *Phys Rev Lett.* 2002;88:128701. <https://doi.org/10.1103/PhysRevLett.88.128701>.
- Dorogovtsev SN, Mendes JFF, Samukhin AN. Structure of growing networks with preferential linking. *Phys Rev Lett.* 2000;85:4633–6. <https://doi.org/10.1103/PhysRevLett.85.4633>.
- Duijn MV, Snijders TAB, Zijlstra B. P2: a random effects model with covariates for directed graphs. *Statistica Neerlandica.* 2004;58:234–54.
- Erdős P, Rényi A. Random graphs. *Publ Math.* 1959;6:290–7.
- Faust K. Centrality in affiliation networks. *Soc Networks.* 1997;19:157–91.
- Feller W. *An introduction to probability theory and its applications, vol. 2.* New York: Wiley; 1966.
- Festinger L. The analysis of sociograms using matrix algebra. *Hum Relat.* 1949;2:153–8.
- Fineberg S, Wasserman S. Categorical data analysis of single sociometric relations. In: *Sociological methodology.* New Jersey: Jossey-Bass; 1981. p. 156–92.
- Fletcher JM. Social interactions and smoking: evidence using multiple student cohorts, instrumental variables, and school fixed effects. *Health Econ.* 2008;19:466–84.

- Fletcher JM, Lehrer SF. The effect of adolescent health on educational outcomes: causal evidence using genetic lotteries between siblings. *Canadian labor market and skills researcher network, working paper no. 32*. 2009.
- Fortunato S. Community detection in graphs. *Phys Reports*. 2010;486:75–174.
- Frank O, Strauss D. Markov graphs. *J Am Stat Assoc*. 1986;81:832–42.
- Freeman L. Centrality in social networks, I. Conceptual clarification. *Soc Networks*. 1979;1:215–39.
- Freeman L. The development of social network analysis: a study in the sociology of science. Vancouver: Empirical Press; 2004.
- Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabasi A-L. The human disease network. *Proc Natl Acad Sci*. 2007;104:8685–90. <http://www.pnas.org/content/104/21/8685.abstract>
- Goldenberg A, Zheng AX, Fineberg SE, Airoidi EM. A survey of statistical network models. *Found Trends Mach Learn*. 2009;2:129–233.
- Goodreau S. Advances in exponential random graph ( $p^*$ ) models applied to a large social network. *Soc Networks*. 2007;29:231–48.
- Granovetter MS. The strength of weak ties. *Am J Sociol*. 1973;78:1360–80.
- Guimera R, Nunes Amaral LA. Functional cartography of complex metabolic networks. *Nature*. 2005;433:895–900.
- Haines V, Hurlbert J. Network range and health. *J Health Soc Behav*. 1992;33:254–66.
- Handcock MS, Robins GL, Snijders TAB, Moody J, Besag J. Assessing degeneracy in statistical models of social networks. *J Am Stat Assoc*. 2003;76:33–50.
- Handcock M, Raftery A, Tantrum J. Model-based clustering for social networks. *J Roy Stat Soc A*. 2007;170:301–54.
- Handcock MS, Hunter DR, Butts CT, Goodreau SM, Krivitsky PN, Morris M. *ergm*: A package to fit, simulate and diagnose exponential-family models for networks. <http://CRAN.R-project.org/package=ergm>. Version 2.2-6. 2010. Project home page at <http://statnetproject.org>
- Hanneke S, Fu W, Xing EP. Discrete temporal models of social networks. *Electron J Stat*. 2010;4:585–605.
- Harary F. On the notion of balance of a signed graph. *Mich Math J*. 1953;2:143–6.
- Harary F. The number of linear, directed rooted and connected graphs. *Trans Am Math Soc*. 1955;78:445–63.
- Heider F. Attitudes and cognitive orientation. *J Psychol*. 1946;21:107–12.
- Hidalgo CA, Blumm N, Barabasi A-L, Christakis NA. A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol*. 2009;5:e1000353. <https://doi.org/10.1371/journal.pcbi.1000353>.
- Hoff PD. Bilinear mixed effects models for dyadic data. *J Am Stat Assoc*. 2005;100:286–95.
- Hoff P. Modeling homophily and stochastic equivalence in symmetric relational data. In: *Advances in neural information processing systems*, vol. 20. Cambridge, MA: MIT Press; 2008. p. 657–64.
- Hoff PD, Raftery AE, Handcock MS. Latent space models for social networks analysis. *J Am Stat Assoc*. 2002;97:1090–8.
- Holland P, Leinhardt S. An exponential family of probability-distributions for directed-graph. *J Am Stat Assoc*. 1981;76:33–50.
- Holland P, Laskey K, Leinhardt S. Stochastic blockmodels: some first steps. *Soc Networks*. 1983;5:109–37.
- House J, Kahn R. Measures and concepts of social support. In: *Social support and health*. Orlando: Academic; 1985. p. 83–108.
- Huisman M, Van Duijn M. Software for statistical analysis of social networks. In: *The Sixth International Conference on Logic and Methodology*; Amsterdam: 2004.
- Huisman M, Van Duijn M. Software for social networks analysis. In: *Models and methods in social network analysis*. Cambridge: Cambridge University Press; 2005.
- Hunter D. Curved exponential family models for social networks. *Soc Networks*. 2007;29:216–30.
- Hunter DR, Handcock MS. Inference in curved exponential family models for networks. *J Comput Graph Stat*. 2006;15:565–83.
- Iwashyna TJ, Chang VW, Zhang JX, Christakis AN. Physician social networks and variation in prostate cancer treatment in three cities. *Health Serv Res*. 2002;37:1531–51.
- Karrer B, Newman MEJ. Stochastic blockmodels and community structure in networks. *Phys Rev E*. 2011;83:016107. <https://doi.org/10.1103/PhysRevE.83.016107>.
- Katz L. On the matrix analysis of Sociometric data. *Sociometry*. 1947;10:233–41.
- Katz L. A new status index derived from sociometric analysis. *Psychometrika*. 1953;18:39–43.
- Katz L, Powell JH. Measurement of the tendency toward reciprocation of choice. *Sociometry*. 1955;18:659–65.
- Keating NL, Ayanian JZ, Cleary PD, et al. Factors affecting influential discussions among physicians: a social network analysis of a primary care practice. *J Gen Intern Med*. 2007;22:794–8.
- Klov Dahl A. Social networks and the spread of infectious diseases. *Soc Sci Med*. 1985;21:1203–16.
- Kossinets G, Watts DJ. Empirical analysis of an evolving social network. *Science*. 2006;311:88–90. <http://www.sciencemag.org/content/311/5757/88.abstract>
- Krapivsky PL, Redner S, Leyvraz F. Connectivity of growing random networks. *Phys Rev Lett*. 2000;85:4629–32. <https://doi.org/10.1103/PhysRevLett.85.4629>.
- Krivitsky PN. Exponential-family random graph models for valued networks. 2012. arXiv preprint, 1101.1359v2 [stat.ME] 19 Jan 2012.
- Krivitsky PN, Handcock MS. Fitting position latent cluster models for social networks with latentnet. *J Stat Softw*. 2008;24. <http://statnetproject.org>
- Krivitsky PN, Handcock MS. A separable model for dynamic networks. 2010. arXiv preprint, 1011.1937v1 [stat.ME].
- Kumpula JM, Onnela J-P, Saramäki J, Kaski K, Kertész J. Emergence of communities in weighted networks.

- Phys Rev Lett. 2007;99:228701. <https://doi.org/10.1103/PhysRevLett.99.228701>.
- Landon BE, Keating NL, Barnett ML, Onnela JP, Paul S, O'Malley AJ, Keegan T, Christakis NA. Variation in patient-sharing networks of physicians across the United States. *JAMA*. 2012;308:265–73.
- Laumann E, Marsden P, Prenskey D. The boundary specification problem in network analysis. In: Burt R, Minor M, editors. *Applied network analysis: a methodological introduction*. Beverly Hills: Sage; 1983. p. 18–34.
- Lorrain F, White H. Structural equivalence of individuals in social networks. *J Math Sociol*. 1971;1:49–80.
- Lyons R. The spread of evidence-poor medicine via flawed social-network analyses. *Stat Polit Policy*. 2011;2:1–26.
- Manski CA. Identification of endogenous social effects: the reflection problem. *Rev Econ Stud*. 1993;60:531–42.
- Marsden P. Network methods in social epidemiology. In: *Methods in social epidemiology*. New York: Jossey-Bass; 2006. p. 267–86.
- Marsden PV, Friedkin NE. Network studies of social influence. *Sociol Methods Res*. 1993;22:127–51.
- Marsili M, Vega-Redondo F, Slanina F. The rise and fall of a networked society: a formal model. *Proc Natl Acad Sci USA*. 2004;101:1439–42.
- McPherson ML, Smith-Lovin C, et al. Birds of a feather: homophily in social networks. *Annu Rev Sociol*. 2001;27:415–44.
- Moreno JL. Who shall survive? Nervous and mental disease processing. The University of Michigan, Ann Arbor; 1934.
- Mucha PJ, Richardson T, Macon K, Porter MA, Onnela J-P. Community structure in time-dependent, multiscale, and multiplex networks. *Science*. 2010;328:876–8. <http://www.sciencemag.org/content/328/5980/876.abstract>
- Newcomb TM. An approach to the study of communicative acts. *Psychol Rev*. 1953;60:393–404.
- Newman ME. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys Rev*. 2001;64:016132.
- Newman MEJ. Modularity and community structure in networks. *Proc Natl Acad Sci*. 2006;103:8577–82.
- Newman M. *Networks: an introduction*. New York: Oxford University Press; 2010.
- Newman MEJ. Communities, modules and large-scale structure in networks. *Nat Phys*. 2012;8:25–31.
- Newman MEJ, Girvan M. Mixing patterns and community structure in networks. In: Pastor-Satorras R, Rubi J, Diaz-Guilera A, editors. *Statistical mechanics of complex networks*. Berlin: Springer; 2003.
- Newman MEJ, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E*. 2004;69:026113. <https://doi.org/10.1103/PhysRevE.69.026113>.
- Nowicki K, Snijders TAB. Estimation and prediction for stochastic blockstructures. *J Am Stat Assoc*. 2001;96:1077–87.
- O'Malley AJ. The analysis of social network data: an exciting frontier for statisticians. *Stat Med*. 2013;32:539–55.
- O'Malley AJ, Christakis NA. Longitudinal analysis of large social networks: estimating the effect of health traits on changes in friendship ties. *Stat Med*. 2011;30:950–64.
- O'Malley AJ, Marsden PV. The analysis of social networks. *Health Serv Outcome Res Methodol*. 2008;8:222–69.
- O'Malley AJ, Arbesman S, Steiger DM, Fowler JH, Christakis NA. Egocentric social network structure, health, and pro-social behaviors in a National Panel Study of Americans. *PLoS One*. 2012;7:e36250. <https://doi.org/10.1371/journal.pone.0036250>.
- Opsahl T. Triadic closure in two-mode networks: redefining the global and local clustering coefficients. *Soc Networks*. 2011;34. <https://doi.org/10.1016/j.socnet.2011.07.001>.
- Opsahl T, Agneessens F, Skvoretz J. Node centrality in weighted networks: generalizing degree and shortest paths. *Soc Networks*. 2010;32:245–51.
- Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*. 2005;435:814–8. <https://doi.org/10.1038/nature03607>.
- Paul S, O'Malley AJ. Hierarchical longitudinal models of relationships in social networks. *J R Stat Soc Ser C Appl Stat*. 2013;62:705–22.
- Pham HH, O'Malley AS, Bach PB, Saiontz-Martinez C, Schrag D. Primary care physicians' links to other physicians through Medicare patients: the scope of care coordination. *Ann Intern Med*. 2009;150:236–42.
- Piraveenan M, Prokopenko M, Zomaya AY. Assortative mixing in directed biological networks. *IEEE Trans Comput Biol Bioinform*. 2010;9:66–78. To appear.
- Pollack CE, Weissman G, Bekelman J, Liao K, Armstrong K. Physician social networks and variation in prostate cancer treatment in three cities. *Health Serv Res*. 2012;47:380–403.
- Porter MA, Onnela J-P, Mucha PJ. Communities in networks. *Not Am Math Soc*. 2009;56(1082–1097):1164–6.
- Price DDS. A general theory of bibliometric and other cumulative advantage processes. *J Am Soc Inf Sci*. 1976;27:292–306. <https://doi.org/10.1002/asi.4630270505>.
- Robins G, Pattison P, Woolcock J. Small and other worlds: global network structures from local processes. *Am J Sociol*. 2005;110:894–936.
- Robins GL, Snijders TAB, Wang P, Handcock MS, Pattison PE. Recent developments in exponential random graph ( $p^*$ ) models for social networks. *Soc Networks*. 2007;29:192–215.
- Robins GL, Pattison PE, Wang P. Closure, connectivity and degree distributions: exponential random graph ( $p^*$ ) models for directed social networks. *Soc Networks*. 2009;31:105–7.
- Rubin D. Bayesian inference for causal effects: the role of randomization. *Ann Stat*. 1978;6:34–58.
- Seidman SB. Network structure and minimum degree. *Soc Networks*. 1983;5:269–87.
- Shalizi RR, Rinaldo A. Consistency under sampling of exponential random graph models. 2012. arXiv preprint. arXiv:1111.3054v3

- Shalizi CR, Thomas AC. Homophily and contagion are generically confounded in observational social network studies. *Sociol Methods Res.* 2011;40:211–39.
- Simmel G. *The sociology of Georg Simmel*. New York: The Free Press; 1908.
- Snijders T. The degree variance: an index of graph heterogeneity. *Soc Networks.* 1981;3:163–74.
- Snijders T. Stochastic actor-oriented models for network change. *J Math Sociol.* 1996;21:149–72.
- Snijders TAB. The statistical evaluation of social network dynamics. In: *Sociological methodology*. Oxford, UK: Basil Blackwell; 2001. p. 361–95.
- Snijders TAB. Models for longitudinal social network data. In: *Models and methods in social network analysis*. Cambridge: Cambridge University Press; 2005. p. 215–47.
- Snijders TAB. Statistical methods for network dynamics. In: Luchini SR et al., editors. *Proceedings of the XLIII Scientific Meeting, Italian Statistical Society*, Basil Blackwell, Ltd; 2006. p. 281–96
- de Solla Price DJ. Networks of scientific papers. *Science.* 1965;149:510–5. <http://www.sciencemag.org/content/149/3683/510.short>.
- Steglich C, Snijders TAB, Pearson M. Dynamic networks and behavior: separating selection from influence. *Sociol Methodol.* 2010;40:329–93.
- Szabo G, Barabasi AL. Network effects in service usage. 2007. Arxiv preprint. <http://lanl.arxiv.org/abs/physics/0611177>
- Thompson S. Adaptive web sampling. *Biometrics.* 2006;62:1224–34.
- Thompson S, Frank O. Mode-based estimation with link-tracing sampling designs. *Survey Methodol.* 2000;26:87–98.
- Thompson S, Seber GAF. *Adaptive sampling*. New York: Wiley; 1996.
- Toivonen R, Onnela J-P, Saramäki J, Hyvönen J, Kaski K. A model for social networks. *Phys A Stat Mech Appl.* 2006;371:851–60. <http://www.sciencedirect.com/science/article/pii/S0378437106003931>
- Traud AL, Mucha PJ, Porter MA. Social structure of Facebook networks. *Phys A Stat Mech Appl.* 2012;391:4165–80. <http://www.sciencedirect.com/science/article/pii/S0378437111009186>
- VanderWeele TJ. Sensitivity analysis for contagion effects in social networks. *Sociol Methods Res.* 2011;40:240–55.
- VanderWeele TJ, Ogburn EL, Tchetgen Tchetgen EJ. Why and when “Flawed” social network analyses still yield valid tests of no contagion. *Stat Polit Policy.* 2012;3:1050. <https://doi.org/10.1515/2151-7509.1050>.
- Vázquez A. Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations. *Phys Rev E.* 2003;67:056104. <https://doi.org/10.1103/PhysRevE.67.056104>.
- Wang W, Wong G. Stochastic Blockmodels for directed graphs. *J Am Stat Assoc.* 1987;82:8–19.
- Wang P, Sharpe K, Robins GL, Pattison PE. Exponential random graph ( $p^*$ ) models for affiliation networks. *Soc Networks.* 2009;31:12–25.
- Wasserman SS, Faust K. *Social network analysis: methods and applications*. Cambridge: Cambridge University Press; 1994.
- Wasserman S, Pattison P. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and  $p^*$ . *Psychometrika.* 1996;61:401–25.
- Westveld AH, Hoff PD. A mixed effect model for longitudinal relational and network data, with applications to international trade and conflict. *Ann Appl Stat.* 2011;5:843–72.
- White D, Harary F. The cohesiveness of blocks in social networks: node connectivity and conditional density. *Sociol Methodol.* 2001;31:305–59.
- Wong LH, Pattison P, Robins G. A spatial model for social networks. *Phys A Stat Mech Appl.* 2006;360:99–120. <http://www.sciencedirect.com/science/article/pii/S0378437105004334>
- Zijlstra BJH, Duijn MV, Snijders TAB. The multilevel P2 model: a random effects model for the analysis of multiple social networks. *Methodology.* 2006;2:42–7.



# Survey Methods in Health Services Research

# 27

Steven B. Cohen

## Contents

<b>Introduction</b> .....	662
<b>Designing National Health-Care Surveys to Inform Health Policy and Health Services Research</b> .....	663
Types of Health and Health-Care Surveys .....	663
Objectives and Content .....	664
<b>Survey Design Framework</b> .....	666
Cross-Sectional and Longitudinal Survey Designs .....	666
Use of Complex Nationally Representative Survey Designs .....	667
Sample Size Determination .....	668
<b>Controlling for Sampling Error and Bias in Survey Estimates</b> .....	669
Sample Size Targets and Precision Requirements .....	669
Building Survey Response Rates .....	671
Survey Procedures to Facilitate Respondent Cooperation .....	672
<b>Estimation of Health-Care Parameters</b> .....	672
Development of Sampling Weights .....	672
Adjustments for Unit Nonresponse .....	673
Adjustments for Survey Attrition .....	674
Post-stratification Adjustments .....	675
<b>Variance Estimation Considerations</b> .....	676
<b>Integrated Survey Designs: Analytical Enhancements Achieved through the Linkage of Surveys and Administrative and Secondary Data</b> .....	676
An Example of Survey Integration: The Medical Expenditure Panel Survey .....	678
Advantages of Integrated Survey Designs .....	679
Linked Provider Data on Expenditures Improves the Accuracy of National Medical Expenditure Estimates in the MEPS .....	680
Integrated Design Expands Capacity for Longitudinal Analyses .....	680
Integrated Design of MEPS Facilitates Examination of Response Error .....	681
Constraints .....	681

---

S. B. Cohen (✉)  
 Division of Statistical and Data Sciences,  
 RTI International, Washington, DC, USA  
 e-mail: [scohen@rti.org](mailto:scohen@rti.org)

© This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2019  
 A. Levy et al. (eds.), *Health Services Evaluation*, Health Services Research,  
[https://doi.org/10.1007/978-1-4939-8715-3\\_38](https://doi.org/10.1007/978-1-4939-8715-3_38)

661

<b>Policy-Relevant Examples from the Medical Expenditure Panel Survey (MEPS)</b> .....	682
Design of the MEPS to Inform Health Policy and Health Services Research .....	682
<b>Issues on Measuring and Estimating Health Insurance Coverage in Surveys</b> .....	684
Testing for the Impact of Survey Attrition on Health Insurance Coverage Estimates in the MEPS .....	684
The Utility of Prediction Models to Oversample the Long Term Uninsured .....	686
<b>Summary</b> .....	693
<b>References</b> .....	694

### Abstract

Health-care surveys serve as a critical source of essential information on trends in health-care costs, coverage, access, and health-care quality. The findings derived from these surveys often facilitate the development, implementation and evaluation of policies and practices addressing health care and health behaviors at the national level. This chapter serves to illustrate several survey methods that enhance the performance and utility of health services research efforts. Attention has been given to the topics of sample and survey designs, nonresponse and attrition, estimation, precision, sample size determination, and analytical techniques to control for survey design complexities in analysis. Several of the topics that are featured in this chapter are further connected by their substantive focus on the measurement of trends in health-care costs, coverage, access, and health-care utilization. In addition to highlighting underlying survey operations, estimates, and outputs, the topics that have been covered also serve to identify potential enhancements that facilitate improvements in design, data collection, estimation strategies, and ultimately analytical capacity for health services research efforts.

### Introduction

There is a growing demand for timely, high-quality, and precise estimates of health-care parameters at the national and subnational levels and associated

readily accessible data resources to inform health-care policy and practice. Existing sentinel health-care databases that provide nationally representative population based data on measures of health-care access, cost, use, health insurance coverage, health status, and health-care quality provide the necessary foundation to support descriptive and behavioral analyses of the US health-care system. Such studies help inform assessments of the availability and costs of private health insurance in the employment-related and non-group markets, the population enrolled in public health insurance coverage and those without health-care coverage, and the role of health status in health-care use, expenditures, household decision making, and health insurance and employment choices. Health services research efforts provide essential insights into the drivers of trends in health-care expenditures and service utilization; serve to estimate the impact of changes in financing, coverage, and reimbursement policy; and help determine who benefits and who bears the cost of a change in policy. Government and nongovernmental entities rely upon these data and research efforts to evaluate health reform policies, the effect of tax code changes on health expenditures and tax revenue, and proposed changes in government health programs such as Medicare.

In this chapter, attention is given to key survey methods that enhance the conduct of health services research efforts. To ensure their utility and integrity, it is essential that health and health-care surveys are designed according to high-quality, effective, and efficient statistical and methodological practices



and optimal sample designs. This also necessitates that subsequent applications of estimation strategies to the survey data, as well as analytical techniques and interpretations of resultant research findings, are guided by well-grounded statistical theory. The chapter also features important sample design considerations, with coverage given to topics that include frame development, sample size specifications, precision requirements, and sample selection scheme. Adhering to a total survey error framework, challenges that characterize health services research efforts are identified, and the interdependence between the analysts, the health-care survey designers, and the statisticians is reinforced. In this context, the methods that are discussed are illustrated with examples from national health-care survey efforts, though the techniques are also applicable to sub-national or population subgroup specific target populations.

---

### **Designing National Health-Care Surveys to Inform Health Policy and Health Services Research**

Surveys are a critical source of information for the development, implementation, and evaluation of policies and practices addressing health and health care. When properly designed, surveys can provide accurate, unbiased, and generalizable information on population characteristics, risk factors, health status, health-care access, utilization and insurance coverage, and the health-care system itself. To be most useful, surveys must be designed according to sound statistical and methodological principles. Health surveys are data collection efforts designed to acquire information on the nation's health and health-care characteristics. Several general, though by no means, exhaustive uses of health and health-care survey data include identification of public health problems; program planning and evaluation; health education and health promotion; epidemiological, biomedical, and health services research; measurement of the extent and impact of illness; and the measurement of the use of health-care services, related medical expenditures, and sources of payment for care.

Generally, surveys are operationalized by the selection of a representative sample of the population or universe of interest, referred to as the target population, and the acquisition of information from the sample units obtained in a structured manner through administration of a well-developed questionnaire. The universe of interest is often a population but can be any identifiable group of individual units such as health-care providers or events such as health-care visits. If the sample is selected as a probability sample, in which a frame exists for sample enumeration and every unit selected from the frame has a known probability of selection for the sample, the findings from the sample are generalizable to the population. This is a powerful attribute and enhances the integrity of the data collected. Surveys can have relatively simple or extremely complex designs, but the basic principles of sample design and data collection methodology remain the same. The complexity of the survey often reflects the complexity of the subject under study. As health and health care encompass a wide range of phenomena and relate directly and indirectly to many other domains, it is necessary to develop a range of health surveys to respond to differing needs for information. Each of these surveys is often based on complex designs and sophisticated data collection mechanisms.

### **Types of Health and Health-Care Surveys**

There are three main types of health surveys: population-based surveys that obtain information directly from the subject (or a suitable proxy), surveys that obtain information about entities such as health-care providers, and surveys that are based on administrative records. Population-based surveys are used when it is essential to describe the characteristics of a defined population. Often the population of interest is the general US population and specific subpopulations as defined by such characteristics as age, sex, race/ethnicity, or socioeconomic status. However, a population may also be defined by occupation or

any other well-defined characteristics. For example, physicians could be the ultimate sample units of a population-based survey if the information sought related directly to the physician and his practice characteristics. Population-based surveys are most frequently adopted when the sample unit is the best source for providing the required information. When a well-developed sample frame is available, samples for population-based surveys can be selected from lists of all eligible subjects. For example, surveys of Medicare or health plan beneficiaries can be drawn from a list of enrollees. When such lists are not available, other methods such as area-based probability samples are used. Whatever method is adopted, it is important that the sample is selected as a probability sample and may be evaluated for potential coverage and response biases. Once a sample is selected, different data collection modes can be used to collect the necessary information including mail, phone, Web based, and in person. The nature of the content and the sample will affect the mode chosen. Of critical importance is the survey instrument itself. The questionnaire or other data collection instruments need to be developed so that accurate and valid information is obtained. The identification of the respondent is also an important step in the process. In general, obtaining information directly from the survey subject provides more reliable and valid health data (Madans and Cohen 2005).

Many population-based health surveys obtain health information directly from the subject (or an appropriate proxy) either through in-person or telephone interviews or mail questionnaires. Supplemental information in the form of medical records is often added to the information obtained from subjects to enhance completeness and quality. To obtain objective standardized information on health characteristics including undiagnosed conditions, surveys rely on direct examination of populations. These surveys are extremely complex and expensive to undertake but are of added value to accurately describe the health status of the population, particularly for those subpopulations who lack medical care.

In contrast, surveys of the components of the health-care system provide information on the

structure, capacity, and functioning of that system. These components range from private physicians' offices to hospitals, nursing homes, and home health-care agencies. To fully understand the system, it is necessary to cover all components. In order to select representative samples of these components, it is necessary to have sampling frames, equivalent to the list frames mentioned above, that identify each member of each type of health-care provider. Surveys of providers can provide information on different aspects of the health-care system. Questions can be targeted at describing the number of components in a sector as well as their organizational, legal, or financial characteristics. Information can be obtained on the individual provider or on the interactions among related providers or between providers and patients. Interactions with patients can focus on the delivery of care or on how care is paid for.

Sometimes the most accurate source of information comes from an administrative record that was generated as part of the routine operation of a system. This clearly would be the case when the objective of the research is the system producing the records. For example, utilization of Medicare services is most easily obtained from Medicare administrative records. The entire census of records is usually available for these purposes, but often samples of records are taken when the entire universe is not needed. When possible, information from administrative records is often sought as a way to improve the accuracy of information available from the subject in population-based surveys.

## **Objectives and Content**

Health surveys used for policy and program development can be either focused on a particular health or health-care issue or can be multipurpose in nature. The latter surveys tend to be conducted by public entities and are designed to provide ongoing, descriptive information on a range of topics and tend to be based on larger samples. While the information from these surveys can track changes in the population, they are less effective in obtaining detailed information on a

particular subject or in evaluating the success of the survey. Such information is more appropriately obtained from focused surveys. In order to allow for comprehensive studies of the current health-care system, information is needed on the population's access to health care, their utilization of and expenditures for health-care services, and their health insurance coverage. In a similar vein, an evaluation of the system requires an understanding of the patterns and trends in the use of health-care services and their associated costs and sources of payment. To effectively address these issues, researchers and policymakers need accurate nationally representative data to better permit an understanding of how individual characteristics, behavioral factors, financial and institutional arrangements affect health-care utilization and expenditures in a rapidly changing health-care market. Health surveys are often designed to acquire this information at both the national and subnational levels and for policy-relevant population subgroups of interest (Madans and Cohen 2005).

### **Access to Care**

The population's access to health-care services is an important factor that may influence patterns of health-care utilization and associated health outcomes. Measures of access to care have also been used as indicators to assess the quality of the nation's health-care delivery system. In addition to facilitating determinations of the availability of a usual source of care for the provision of necessary medical care, access to care measures serve to identify barriers to care, which include shortages of health-care providers, financial restrictions, limitations in proximity to services, and constraints associated with waiting times. Population-based national health-care surveys such as the Medical Expenditure Panel Survey (MEPS), cosponsored by the Agency for Healthcare Research and Quality (AHRQ) and the Centers for Disease Control and Prevention's National Center for Health Statistics (CDC/NCHS), collect information on several dimensions of access to health care in America. The survey was designed to yield estimates of the proportion of the population lacking a usual source of care as well as the types and characteristics of providers used by those who do have a usual source

of care. Measures of satisfaction with the usual source of health care are also collected in the survey through in-person interviews, in addition to information on experiencing difficulty or delay in obtaining health care or not receiving needed health-care services.

In addition to national population estimates of access to care derived from this survey, the analytical objectives include a capacity to permit specific comparisons of these measures by age, race/ethnicity, sex, perceived health status, health insurance coverage, and place of residence. These analyses permit the identification of potential disparities in access to care, with particular attention given to individuals with low incomes, persons with disabilities or chronic illness, minorities, women and children, elderly, rural, and inner-city populations. Evaluation of the effects of changes in the US health-care system on access to care for these populations will remain a critical issue for policymakers in the next few years.

### **Use of Health-Care Services**

An understanding of the patterns and trends in the use of health-care services is essential to facilitate evaluations of the current health-care system, in addition to informing proposals for modification. Assessments of the degree of equity in the distribution of health-care services and the identification of health-care disparities require an examination of health-care use across vulnerable population subgroups and how it has changed over time. These investigations are essential to discern how service utilization varies according to the characteristics of the population, their health plans, and their providers and to identify other behavioral and institutional factors associated with disparities in service use.

An examination of the variations in the use of health-care services also helps determine the adequacy of access to care across the population. Underutilization of health-care services may be attributable to limitations in access to care as a consequence of the lack of adequate health insurance, financial resources, or limited availability of services in certain areas. Detailed comparisons of patterns of use by subpopulations presumed to require more care (e.g., the elderly, those in poor

health, or the terminally ill) relative to their less vulnerable counterparts help discern whether those most in need of care are receiving it.

The utilization measures that are required for these analyses typically consist of counts of the number of visits or events for specific health-care services that occur in a given calendar year. More specifically, health-care services include office-based visits, ambulatory hospital-based visits, inpatient hospital stays, dental visits, home health visits, and prescribed medicine purchases. This information is acquired through population-based surveys, surveys of providers, and surveys based on administrative records. Health-care surveys are designed to acquire this information at both the national and subnational levels and for policy-relevant population subgroups of interest. The visible national data collection efforts that acquire this type of health-care utilization information include the MEPS and the National Health Interview Survey (population based), the National Ambulatory Medical Care Survey (provider based), and the Medicare Current Beneficiary Survey (primarily population based and supplemented with administrative records).

### **Cost of Medical Care and Coverage**

Health-care expenditures represent one-sixth of the US gross domestic product, exhibit a rate of growth that exceeds other sectors of the economy, and constitute one of the largest components of the Federal and states' budgets. Although the rate of growth in health-care costs slowed in the mid-1990s, it began to rise again shortly afterward, fueled primarily by increasing costs for hospital care and prescription medications. To effectively address the issue of rising costs, researchers and policymakers need accurate nationally representative data to better permit an understanding of how individual characteristics, behavioral factors, financial incentives, and institutional arrangements affect health-care utilization and expenditures in a rapidly changing health-care market.

The continuing rise in the number of persons without private health insurance has made access to health insurance coverage a critical public policy issue. Informed public policy requires precise

estimates of the size and composition of the insured and uninsured populations, as well as information on how demographic characteristics, economic factors, and health status affect health plan eligibility and decisions to enroll in health insurance plans. The demand for accurate and reliable information on the population's health-care expenditures, insurance coverage, and sources of payment is met by health-care surveys such as the Medical Expenditure Panel Survey (MEPS) cosponsored by the Agency for Healthcare Research and Quality (AHRQ) and NCHS.

---

## **Survey Design Framework**

Once the underlying survey objectives are articulated, greater specificity is required in order to finalize the underlying survey design. With several competing analytic objectives under consideration, priorities need to be established which will serve to guide the necessary precision specifications for the core study estimates of the target population parameters. A final set of survey objectives is then developed that provides details of the core population domains of interest and the required levels of precision for domain estimates. Underlying study hypotheses to be tested also need to be well specified. The precision requirements for the survey estimates will then be subject to further evaluation and subject to re-specification based on cost constraints.

### **Cross-Sectional and Longitudinal Survey Designs**

National health-care sample surveys are generally characterized by cross-sectional or longitudinal designs. The cross-sectional surveys are designed to provide a snapshot of population characteristics that relate to a fixed point or interval in time. Alternatively, longitudinal surveys collect data on more than one occasion from the sample members of the population of analytical interest in order to measure change and to obtain data for time periods too long to recall accurately in a single interview. Longitudinal observations are essential for characterizing variations in the

population attributes that are sensitive to changes in time.

Longitudinal survey designs are primarily adopted to provide the necessary information to assess changes in the behavior of the population over a specific time period. Often referred to as panel designs, they have the capacity to permit measurement of seasonal and annual variations in population characteristics and behavior. These longitudinal designs are essential to permit the acquisition of the necessary data that will support analyses that measure the impact of changes in health status over time for individuals with specific conditions with respect to their use of health-care services and related expenditures. Well-specified sample size requirements for these surveys that are achieved will also permit comparable studies for different economic groups or special populations of interest, such as the poor, elderly, veterans, the uninsured, or racial/ethnic groups. This type of survey design also allows for the development of economic models designed to produce national and regional estimates of the impact of changes in financing, coverage, and reimbursement policy over time, as well as estimates of who benefits and who bears the cost of such changes in policy.

Longitudinal designs are particularly attractive and well suited for studies that examine the extent of changes in health insurance coverage over time as well as the persistence of catastrophic medical expenditures over time. A cross-sectional survey design can provide accurate national survey estimates of the percent of the population with private coverage, public coverage, or the uninsured at a fixed point in time. Alternatively, the most accurate population estimates of the percent of population ever uninsured in a given year or without coverage for an entire year's duration come from data collection efforts that have adopted a longitudinal survey design.

### **Use of Complex Nationally Representative Survey Designs**

Many of the large national health-care surveys are characterized by a complex design structure with

several stages of sampling. Cluster sampling is also a common feature of these national samples that consider area samples. In these multistage sample designs, the first stage of sampling requires the development of a sampling frame in which the land mass of the nation is partitioned into primary sampling units (PSUs) defined as counties or groups of contiguous counties. The eligible set of units are then stratified based on available geographic and socio-demographic information, and a first-stage sample of these primary units is then selected. This process of subsampling areas continues until sample segments consisting of 100–200 housing units are identified and subsampled. The final stage of sampling is often characterized by the selection of a representative sample of housing units, which are then interviewed to obtain the essential survey information on which subsequent health services research will be based.

This type of sample design has the following attractions. The specification of the sampling frame is both cost-effective and less labor intensive, where the list frames of target population members need to only be constructed for the sampled areas. In addition, the interviewing activity is restricted to the sample areas, achieving efficiencies in travel time and cost for in-person interviewing. In contrast, these efficiencies are achieved at the expense of a loss in precision of survey estimates based on the specified sample size relative to the precision that would be achieved based upon a simple random sample selection scheme. The increased variance in survey estimates in a multistage sample design relative to simple random sampling is the result of the greater likelihood of geographically clustered units to have more homogeneous responses. This within cluster homogeneity is measured by the intra-cluster correlation coefficient which measures the correlation between units from the same cluster. Overall, the differential in the variance of a survey estimate of a population mean  $\bar{y}$  based on a complex multistage sample design  $\text{Var}_{\text{Design}}(\bar{y})$  with disproportionate sampling relative to a simple random sample  $\text{Var}_{\text{SRS}}(\bar{y})$  is specified as the design effect.

$$\text{Design effect} = \text{Var}_{\text{Design}}(\bar{y}) / \text{Var}_{\text{srs}}(\bar{y})$$

In addition, the effective sample size for a design that departs from simple random sampling assumptions is specified as the underlying sample size,  $n$ , divided by the design effect.

$$\begin{aligned} &\text{Effective sample size} \\ &= n / [\text{Var}_{\text{Design}}(\bar{y}) / \text{Var}_{\text{srs}}(\bar{y})]. \end{aligned}$$

### Sample Size Determination

Stratification is used in sample designs to improve the precision of survey estimates and also provide greater control of the sample distribution. For less complex designs, when a fixed set of strata ( $h = 1, 2, \dots, H$ ) are defined and data collection costs for surveying units from each distinct stratum and the associated variance estimates for a core criterion variable have been determined, optimum sample allocation strategies have been developed. The values of the samples sizes for each stratum,  $n(h)$ , may be selected to minimize the variance  $\text{Var}(\bar{y})$  of the survey estimate of the criterion variable  $y$ , expressed as a mean, for a fixed cost ( $C$ ) or to minimize the cost for a specified level of precision  $V(\bar{y})$ . Considering a cost function of the form

$$\begin{aligned} &\text{Data Collection Cost } \{C\} \\ &= C(o) + \sum_h^H C(h)n(h) \end{aligned}$$

where  $C(o)$  represents an overhead cost and  $C(h)$  is the data collection cost per unit.

The variance of the estimated mean of a criterion variable will be minimized when  $n(h)$  is proportional to  $N(h)S(h)/\sqrt{c(h)}$ , where  $N(h)$  is the population in stratum  $h$  and  $S(h)$  is the standard deviation for the criterion variable.

When cost is fixed, the overall sample size specification to minimize the variance of survey estimate  $y$  when considering stratified sampling is

$$n = \left\{ \frac{(C - C(o)) \sum_h^H N(h)S(h) / \sqrt{c(h)}}{\sum_h^H N(h) \times S(h) / \sqrt{c(h)}} \right\}$$

Alternatively, when the precision level  $V$  is fixed, the overall sample size specification to minimize cost under stratified sampling assumptions is

$$\begin{aligned} n = & \left[ \sum_h^H W(h)S(h)\sqrt{c(h)} \right] \\ & \times / \left\{ V + (1/N) \sum_h^H W(h)S^2(h) \right\}, \end{aligned}$$

where  $W(h) = N(h)/N$  and  $N = \sum_h^H N(h)$  (Cochran 1977).

In practice, few health-care surveys are conducted with the primary objective of optimizing the design based on a single parameter estimate. When the design specifications require attention to competing precision specifications for a variety of survey estimates, the optimization process becomes much more complex. Often, sample size optimization for multiple variance constraints does not have a closed form solution. Conventional approaches under these circumstances rely on iterative approaches to sample size determination that provide an optimal solution when convergence criteria are satisfied (Chromy 1981).

For national health-care surveys, the precision requirements may be articulated by specifying the amount of error that may be tolerated in the survey estimates. To illustrate this process, assume some margin of error,  $d$ , in the estimated survey mean of a criterion variable of interest  $y$  from the survey has been established, and there is a small risk ( $\alpha$ ) that the sponsors are willing to incur that the actual error is larger than  $d$ . This can be expressed as  $\text{Pr}(|\bar{y} - \bar{Y}| \geq d) = \alpha$ .

For large samples,  $n$  is approximated by (Design effect)  $[z^2 S^2/d^2]$ , where  $z$  is the cutoff point on a standardized normal distribution that cuts off an area  $\alpha$  at the tails and  $S^2$  is the variance of  $y$ .

Another way to determine the sample size is to specify the relative standard error (RSE) required for the resultant survey estimate. The RSE of a survey estimate is defined as the ratio of the standard error of the survey estimate  $\text{SE}(\bar{y})$  divided by the estimate  $\bar{y}$  or  $\text{RSE}(\bar{y}) = \text{SE}(\bar{y})/\bar{y}$

$$\begin{aligned} &\text{Since the } \text{RSE}(\bar{y}) = S/\bar{y}\sqrt{n}, \text{ then } n \\ &= S^2 / (\bar{y}^2 \text{RSE}^2(\bar{y})). \end{aligned}$$

For example, if one was attempting to obtain an estimate of the proportion of the population under age 65 uninsured in a given year,  $p$ , with a

RSE = .05 and a survey design effect = 1.6, and a prior estimate of  $p$  at .15, the necessary sample size would be  $n = 3627$ :

$$\text{Here, } n = 1.6(.15)(.85) / \left[ (.15)^2 (.05)^2 \right] = 3627$$

and

$$\text{RSE}(p) = (1.6 \times (.15 \times .85))^{1/2} / [3627^{1/2} (.15)] = .05$$

## Controlling for Sampling Error and Bias in Survey Estimates

Survey estimates of population parameters are subject to two core sources of error, variable error and bias. Variable error is the random component of survey error attributable to sampling error and variable measurement errors. The bias component captures systematic errors in the survey estimates attributable to the estimation procedure, survey nonresponse, and other sources of nonsampling errors inherent in the measurement process. Consequently, the total error associated with using the sampling estimate  $\bar{y}$  to estimate a population parameter  $\bar{Y}$  is defined as the difference between the estimator and the parameter or  $\{\bar{y} - \bar{Y}\}$ . Using this framework, the total error may be decomposed into the following two terms:

$$\bar{y} - \bar{Y} = \{\bar{y} - E(\bar{y})\} + \{E(\bar{y}) - \bar{Y}\},$$

where  $E(\bar{y})$  is the expected value of the statistic  $\bar{y}$  over repeated sample selections. The first term represents variable errors associated with the sampling and measurement process. The second term represents the bias of the estimator, quantified as the differential between the expected value of the sample statistic over repeated samples and the true value of the population parameter. Based on this specification of the total error associated with the survey estimate of a population parameter, the accuracy of the survey estimate can be assessed by deriving its mean square error. More specifically, the mean square error of an estimator  $\bar{y}$  is defined as the expected value of the squared total error,

$$\begin{aligned} \text{MSE}(\bar{y}) &= E(\bar{y} - \bar{Y})^2 \\ &= E(\bar{y} - E(\bar{y}))^2 + (E(\bar{y}) - \bar{Y})^2 \\ &= \text{Var}(\bar{y}) + \text{Bias}^2 \end{aligned}$$

A desired objective of national health-care survey designs to inform health services research efforts, and all surveys in general, is the minimization of the mean square error of survey estimates. This requires attention to controlling the allowable level of error attributable to each of these distinct sources of error, the error due to sampling and the bias in survey estimates. A well-designed survey requires careful attention given to the data collection protocol, the design of the survey questionnaire, and the operationalization of data collection strategies to minimize survey nonresponse. With respect to sources of survey nonresponse, this would include unit nonresponse, survey attrition in panel or longitudinal surveys, and item nonresponse. Once the data collection phase of the survey is completed, the variable component of the error is usually fixed. Scrutiny should then be given to the identification of the drivers of survey nonresponse and the sources of systematic measurement error in response profiles. Consideration should also be given to the implementation of nonresponse adjustment strategies, imputation to correct for item nonresponse, and logical editing procedures to alleviate response error inconsistencies in the survey responses.

## Sample Size Targets and Precision Requirements

To illustrate the sample size targets for a national health-care survey and associated precision targets, the following example is provided (Cohen 2000). Often, an overall precision requirement for the national health-care survey is specified as the achievement of an average design effect specification for survey estimates of the policy-relevant population subgroups (e.g., average design effect = 1.6). Precision requirements for the survey are then presented in terms of relative

standard errors for the following survey estimates (Table 1):

- A 20% population estimate at the person level for each specified domain (e.g., a percent population estimate such as the rate of uninsured for the population under age 65)
- Mean estimates of the following measures of health-care utilization and expenditures at the

person level (precision requirement specified as an average relative standard error):

- Total health expenditures
- Utilization and expenditure estimates for inpatient hospital stays
- Utilization and expenditure estimates for ambulatory physician visits
- Utilization and expenditure estimates for dental visits
- Utilization and expenditure estimates for prescribed medicines

**Table 1** Targeted average relative standard errors (RSEs) for subpopulation of analytic interest in the 1997 Medical Expenditure Panel Survey Household Component

Subpopulation	Average RSE for a population estimate of 20% (e.g., percent uninsured)	Average RSE for mean use and expenditure estimates
Persons with family income less than 200% of poverty level	.020	.035
Persons ages 18–64 predicted to incur high medical expenditures	.040	.070
Persons 65 years and over	.042	.070
Adults (18 and over) with functional impairments measured in terms of ADLs <sup>a</sup>	.080	.135
Adults (18 and over) with other impairments measured in terms of IADLs <sup>b</sup>	.080	.135
Children (under age 18) with activity limitations	.080	.135
Overall population	.015	.023

**Source:** Center for Financing, Access, and Cost Trends, Agency for Healthcare Research and Quality: Medical Expenditure Panel Survey Household Component, 1997

<sup>a</sup>Need help in one or more activities of daily living (ADLs), such as bathing and dressing

<sup>b</sup>Need help in one or more instrumental activities of daily living (IADLs), such as shopping or paying bills

To meet these requirements, the survey must include a minimum number of persons in each domain of interest. The sample sizes necessary to satisfy these precision requirements for the survey estimates are then derived, adjusting for survey nonresponse targets and assumptions regarding the survey’s sample design and estimated design effects. The necessary sample sizes required to meet the precision targets for survey estimates presented in Table 1 are specified in the following table (Table 2; Cohen 2000).

**Table 2** Targeted sample yields at the end of three core data collection rounds for 1997 for subpopulations of analytic interest: 1997 Medical Expenditure Panel Survey Household Component

Subpopulation	Targeted sample yield
Person with family income less than 200% of poverty level	15,000
Persons ages 18–64 predicted to incur high medical expenditure	4000
Persons 65 years and over	3700
Adults (18 and over) with functional impairments measured in terms of ADLs <sup>a</sup>	1000
Adults (18 and over) with other impairments measured in terms of IADLs <sup>b</sup>	1000
Children (under age 18) with activity limitations	1000
Overall population	34,000

**Source:** Center for Financing, Access, and Cost Trends, Agency for Healthcare Research and Quality: Medical Expenditure Panel Survey Household Component, 1997

<sup>a</sup>Need help in one or more activities of daily living (ADLs), such as bathing and dressing

<sup>b</sup>Need help in one or more instrumental activities of daily living (IADLs), such as shopping or paying bills



The current MEPS sample consists of approximately 14,000 households and 32,000 individuals and includes oversampling of African-Americans, Hispanics, Asians, and low-income households. With respect to desired levels of precision for survey estimates, a relative standard error (RSE) specification of less than or equal to 10% is recommended for survey estimates that characterize policy-relevant population subgroups which include racial and ethnic minorities (RSE ( $Y$ ) = standard error ( $Y$ ) divided by the estimate  $Y$ ).

### Building Survey Response Rates

In national household health-care surveys, significant amounts of resources are allocated to obtain the participation of households that constitute the last 5–10% of the overall survey response rate. A substantial number of households that respond toward the end of the survey field period are characterized by an initial refusal to participate. When the specified response rates are in jeopardy of not being met, concerted use of nonresponse conversion techniques are employed in tandem with occasional extensions of the length of the field period. Applications of these “ninth inning” field force engagements to achieve target survey response rates are not cost neutral and often result in significant increments to data collection costs. The primary objective of this approach is to enhance overall longitudinal survey response rates and achieve a reduction in survey error attributable to nonresponse. It has also been noted that reluctant respondents occasionally differ from the more cooperative survey respondents on sociodemographic characteristics, which may translate to significant differences in the core analytic measures obtained from the survey (Stinchcombe et al. 1981; Cohen et al. 2000; Lynn 2009). These differences are a key reason for continuing to spend resources following them. Alternatively, findings from the European Social Survey and a number of state-level health-related surveys in the USA suggest there are few statistically significant differences between the sample obtained before and after refusal conversion

(Stoop et al. 2010). Reluctant respondents are also more likely to attrite over the course of the survey. Within fixed survey budget constraints, these costly late-stage call-back interviews impact on overall data quality, timeliness of data release, and overall sample size specifications.

Several studies have demonstrated the utility of subsampling nonrespondents in a survey to help minimize nonresponse bias and achieve efficiencies in data collection efforts. Many of these applications are modeled after the technique proposed by Hansen and Hurwitz (1946) to select a subsample from the nonrespondents to get an estimate for the subpopulation represented by the nonrespondents (Vartivarian et al. 2006). Variants of the procedure include application of double sampling for ratio and regression estimation with a subsampling of the nonrespondents.

The subsampling of nonrespondents is considered in order to limit survey costs while maintaining a nationally representative sample. In this vein, the National Survey of Family Growth has implemented a multiphase design which employs the subsampling of nonrespondents. These approaches are increasingly attractive to survey designers because they allow for methods to control the costs at the end of a data collection period while addressing concerns about nonresponse rates and errors. For many national in-person household surveys, large costs are incurred for travel to sample segments to interview a small set of sample units, usually those extremely difficult to contact in prior visits or repeatedly displaying some reluctance to respond to the survey. By restricting these expensive visits to a sample of the nonrespondents at the end of the study, a more cost-effective method concentrates remaining resources on increasing response rates. Additional examples of this approach are found in the General Social Survey, the National Comorbidity Survey Replication, the National Survey of America’s Families, and the National Survey of Recent College Graduates. Related efforts focused on subsampling callbacks to improve survey efficiency have yielded mixed results, with trivial savings achieved in applications to the National Comorbidity Survey, contrasted with more cost-effective results attained in the

American Community Survey. Adaptive survey designs have also been considered as a related framework to improve the efficiency of survey data collection through the application of more tailored data collection treatments for different households identified using paradata. A special case of an adaptive survey design is the responsive survey design, where alternative treatments or data collection strategies are identified (Groves and Heeringa 2006).

### Survey Procedures to Facilitate Respondent Cooperation

In national survey efforts such as the MEPS, it is essential to achieve as high a response rate as feasible in order to reduce the potential error due to nonresponse bias that may impact on resultant survey estimates. The “tool chest” of methods to maximize survey participation and maintain cooperation across the multiple rounds of data collection is quite extensive and expensive to administer. The interviewers are often selected from the data collection organization’s pool of experienced staff. Location, previous interviewing experience, work samples, and language fluency are some of the key criteria used for selecting the interviewers and the supervisors. Due to the oversampling of Hispanics in several of these national surveys, a portion of the interviewers must be fluent in both English and Spanish. New household interviewers also receive intensive project-specific training and general interviewing techniques.

The households that are selected to participate are traditionally sent a notification letter which explains how they were selected for the survey along with a brief description of the survey. Field staff then call the household when a number is provided or attainable to further introduce the project and make an appointment to conduct the interview. Intensive follow-up efforts often are made by interviewers to contact persons not at home, to follow-up broken appointments, and to convert refusals. The interviewers are provided with a variety of materials designed to explain the importance of the study and establish its legitimacy. Interviewers are generally required to

record all contacts (in-person, telephone, by mail) that are made with the household and whether they were successful or not. Where appropriate, the conversion attempt may involve reassigning work to a more experienced interviewer.

## Estimation of Health-Care Parameters

### Development of Sampling Weights

Probability sampling is utilized in health-care surveys to permit the analysis of data from the sample to make inferences about the target population of interest. In order to derive unbiased national estimates of population parameters, the selection probability for each sampling unit must be incorporated into the estimation strategy. This is achieved through the introduction of sampling weights, which adjust for the differential probabilities of selection of the respective sampled units in the health-care survey. In this context, stratified, multistage, area probability samples allow for approximately unbiased estimation of health-care parameters at the national level, contingent on the application of sampling weights that reflect the sampled unit probabilities of selection into the sample. The sampling weight is defined as the inverse or reciprocal of a sample unit’s selection probability into the sample. For multistage sample designs, the weights will be specified as the inverse of the product of each sample unit’s stage-specific selection probability.

In a four-stage sample design typical of national household health-care surveys, the initial sampling weight for the  $k$ -th person in the  $j$ -th housing unit in the  $i$ -th sample segment in the  $h$ -th primary sampling unit (generally a county or group of contiguous counties) selection probability is specified as

$$W_{hijk} = [P_h P_{i|h} P_{j|hi} P_{k|hij}]^{-1} = [P_{hijk}]^{-1}$$

where  $P_{hijk}$  is the selection probability for the  $k$ -th person in the  $j$ -th housing unit in the  $i$ -th sample segment in the  $h$ -th primary sampling unit;  $P_h$  is the first-stage selection probability of

selecting the  $h$ -th primary sampling unit;  $P_{jh}$  is the second-stage conditional probability of selecting the  $i$ -th segment, given the  $h$ -th primary sample sampling unit is selected;  $P_{j|hi}$  is the third-stage conditional probability of selecting the  $j$ -th housing unit, given the  $i$ -th segment in the  $h$ -th primary sample sampling unit is selected; and  $P_{k|hij}$  is the final-stage conditional probability of selecting the  $k$ -th individual, given the  $j$ -th housing unit in the  $i$ -th segment in the  $h$ -th primary sample sampling unit is selected.

Generally,  $P_{k|hij} = 1$ , as all members of a sampled household are selected to participate in the survey with certainty. These sampling weights may be interpreted as inflation factors to represent the number of units in the target population associated with the respective sample unit.

### Adjustments for Unit Nonresponse

Once the data collection effort is concluded, care must be taken to further adjust the survey unit sampling weights to correct for survey nonresponse. In general, the greater the difference among subgroups in response rates and the analytic characteristic(s) of interest, the greater is the need to adjust survey weights for nonresponse. In practice, weighting class nonresponse adjustments are implemented under the assumption that nonresponding sampling units have responded in a manner similar to that of respondents with similar sociodemographic and economic characteristics within the same adjustment class. Properly designed, a weighting class nonresponse adjustment strategy can result in reduced nonresponse bias. The technique requires that the sample be partitioned into mutually exclusive and exhaustive classes, with classification information available for both responding and nonresponding units that are correlated with response propensity and the core criterion variables of the study (Cox and Cohen 1985).

In national health-care surveys, analyses are conducted of characteristics associated with differential nonresponse. These analyses help identify the most important measures to use in developing a nonresponse adjustment to the

survey sampling weights to correct for potential nonresponse bias, most often applied at the housing-unit level. To facilitate these analyses, the demographic, socioeconomic, health-related, and interview-specific profiles of respondents and nonrespondents are examined, based on available data for both groups (Groves et al. 2009). Based on the results of these analyses, weighting classes are specified to adjust for housing unit nonresponse. For illustrative purposes, consider weighting classes defined by cross-classifications of the following measures from the Medical Expenditure Panel Survey (Cohen et al. 1999):

- Family income of primary reporting unit (less than \$10,000; \$10,000–19,999; \$20,000–34,999; \$35,000 or more; unknown)
- Size of dwelling unit (one, two, three, four, five, or more)
- MSA size (MSA, population 500,000 or more; MSA, population less than 500,000; non-MSA)
- Region (Northeast, Midwest, South, West)
- Employment classification of reference person (government, private sector, not in labor force/never worked/worked without pay, unknown or under 18 years of age)
- DU-level personal help measure (units with at least one member unable to perform personal care activities or other routine needs, remaining units with person 70 and over, remaining units with no limitations)
- Propensity to cooperate, based on providing phone number during NHIS (phone number provided, phone present but no number provided, no phone, unknown)
- Age of reference person (under 25, 25–34, 35–44, 45–64, 65 and over)
- Race/ethnicity of reference person (Hispanic, black non-Hispanic, other)
- Sex of reference person
- Marital status (married, spouse present, other)

Overall, C cells were identified based on cross-classifications of these measures, with cell collapsing often specified according to a hierarchy determined by significance level to insure adequate sample representation of the cell. Following

this approach, the nonresponse adjustment for the  $c$ -th weighting class takes the form

$$B(c) = \frac{\sum_{iec} E(i) DUPSWT(i)}{\sum_{iec} R(i)} \times DUPSWT(i)$$

where  $DUPSWT(i)$  is the initial housing unit weight for the  $i$ -th sample housing unit, which reflects the reciprocal of the housing unit's overall selection probability for the sample survey,  $E(i) = 1$  for all survey housing units selected for interviews,  $E(i) = 0$  otherwise,  $R(i) = 1$  for all selected housing units responding in the survey,  $R(i) = 0$  otherwise, and  $iec$  represents eligible housing units classified in weighting class  $c$ .

Consequently, the estimation weight adjusted for the respective survey's housing unit nonresponse,  $WGTHU1(i)$ , for the  $i$ -th housing unit associated with class  $c$ , takes the form  $WGTHU1(i) = B(c) \times DUPSWT(i)$ . Generally, survey participation is an all or none decision for the entire household, so surveys that interview all members of sampled households will assume this nonresponse adjusted household sampling weight,  $WGTSP1(i) = WGTHU1(i)$ . Alternatively, when there is differential nonresponse within households, an additional weighting class adjustment should be implemented to correct for this additional level of person-level nonresponse in the survey. Based on detailed studies of unit nonresponse in national health-care surveys, studies have revealed survey nonrespondents were more likely to consist of smaller households, reside in metropolitan areas, and have higher incomes.

### Adjustments for Survey Attrition

Some of the large annual national health-care surveys also are characterized by a longitudinal design. The data collected in these ongoing longitudinal surveys may be designed to permit studies of the determinants of health insurance coverage and the use of health services and expenditures over time and to identify changes in the provision of health care in relation to social and demographic factors such as employment or income, the health status and satisfaction with

health care of individuals and families, and the health needs of specific population groups such as the elderly and children. In longitudinal survey designs with multiple rounds of data collection, the overall survey response rate is a multiplicative function of the round-specific response rates. In addition to adjusting for survey nonresponse at the first round of a longitudinal survey with multiple rounds of data collection, additional adjustments to the estimation weights are necessary to help mitigate the potential influence of survey attrition on bias in estimates. When the rate of partial response is modest, it is often preferable to treat the partial respondents as complete nonrespondents. In this case, an additional weighting class adjustment to the survey estimation weight to control for survey attrition is appropriate. For example, if a survey required three rounds of data collection to obtain calendar year information for the population, the first-round person-level estimation weights would be adjusted for survey attrition in the following manner:

$$WGTSP2(i) = F(c) \times WGTSP1(i)$$

for the  $i$ -th person associated with class  $c$ ,

where the nonresponse adjustment for the  $c$ -th weighting class takes the form

$$F(c) = \frac{\sum_{iec} E(i) WGTSP1(i)}{\sum_{iec} R(i)} \times WGTSP1(i)$$

and

$WGTSP1(i)$  is the round 1 nonresponse adjusted person-level weight for the  $i$ -th round 1 respondent;  $E(i) = 1$  for all round 1 respondents with positive values of  $WGTSP1(i)$ ;  $E(i) = 0$  otherwise;  $R(i) = 1$  for all persons with  $E(i) = 1$  who responded for their entire period of eligibility in the calendar year covered by the survey over all three data collection Rounds;  $R(i) = 0$  otherwise; and  $iec$  represents all full- and part-year respondents classified in weighting class  $c$ .

Often, a logistic regression analysis is used to identify the most important measures to include in specifying a nonresponse adjustment to the

estimation weights in a longitudinal survey to correct for part-year response at the person level. To illustrate the identification of weighting class cells,  $c$ , consider cross-classifications of the following measures as of the initial round of the MEPS survey:

- Round 1 interview classification (no initial refusal, initial refusal)
- Size of MEPS family (one, two, three, four, five, or more)
- MSA (MSA, non-MSA)
- Age (under 20, 20–29, 30–44, 45–64, 65 and over)
- Marital status of reference person (married, widowed, divorced, separated, never married).

According to prior studies of survey attrition in a large-scale national longitudinal health-care survey, participants who initially refused to respond in the survey were more likely to drop out of the survey in subsequent rounds, in addition to those residing in metropolitan areas. Furthermore, survey attrition was positively correlated with residing in a household with a large number of members, being elderly, never being married, and being without health insurance coverage.

Once adjustments for unit nonresponse and survey attrition have been implemented, attention must also be given to strategies to correct for item nonresponse. Imputation techniques are then considered to complete the data profiles for survey items to facilitate the derivation of survey estimates. Models are developed to identify the best predictors of the criterion variables affected by item nonresponse. The predictors are then used to inform imputation strategies that serve to substitute a value for the missing data. Standard variance estimation procedures applied to data sets that have implemented imputation strategies often underestimate the component of variance due to imputation. Consideration of multiple imputation techniques, where each missing value is replaced by a set of plausible values, provides a framework to adjust the variances of survey estimates for imputation and also help minimize the bias in survey estimates attributable to item nonresponse (Rubin 1987).

## Post-stratification Adjustments

To further improve upon the precision of the survey estimates obtained from a health-care survey, poststratification or stratification after sample selection is often employed to complement the initial stratification imposed at the selection stage. The methodology assumes the availability of population control totals for the measures used for poststratification or consideration of estimates of population control totals from a large national population-based survey with high levels of precision in survey estimates. Additional gains from poststratification arise as a consequence of enhanced corrections for survey nonresponse and undercoverage. To illustrate the application of a poststratification adjustment to the survey estimation weights via a weighting class adjustment, the following procedure can be used:

$$WGTS3(i) = G(c) \times WGTS2(i)$$

for the  $i$ -th person associated with class  $c$ ,

where the poststratification adjustment for the  $c$ -th weighting class takes the form

$$G(c) = POPTOT(c) / \sum_{iec} WGTS2(i)$$

where  $WGTS2(i)$  is the first-round nonresponse and attrition adjusted person-level weight for the  $i$ -th complete respondent and  $iec$  represents all full- and part-year respondents classified in weighting class  $c$ , and the weighting class  $c$  is defined by cross-classifications of population control totals  $POPTOT(c)$  obtained from the Current Population Survey for the given year for the following measures: Census region (Northeast, Midwest, South, and West), MSA status (MSA, non-MSA), race/ethnicity (Hispanic, Black but non-Hispanic, Asian, and other), sex, age, and poverty status. In a complementary manner, poststratification can also be implemented in through iterative marginal adjustments cycling through the respective population control totals for each measure, also known as “raking.”

Once all these adjustments are made to improve the accuracy of survey estimates, differences in survey estimates derived from alternative

survey sources may still occur. Several factors can contribute to differences in estimates of health-care parameters across surveys. These factors include survey content and questionnaire design, definitions of the criterion measures, survey design and methods, and post-data collection processing such as editing, imputation, and estimation techniques. Survey design features such as length of recall period, sample design, and response rates affect the accuracy and precision of survey estimates of coverage. Alternative methodologies for editing the survey data, imputation procedures, and adjustments for survey nonresponse can also affect the final survey estimates that are generated. In addition, estimates within and across surveys differ depending on the duration of the time period that the survey estimates cover.

---

### Variance Estimation Considerations

To obtain accurate estimates from complex survey data, for either descriptive statistics or more sophisticated analyses based on multivariate models, the survey design complexities need to be taken into account. This is achieved by applying the survey estimation weights to produce the survey estimates and using an appropriate technique to derive standard errors associated with the weighted estimates. Several methods for estimating standard errors for estimates from complex surveys have been developed, including the Taylor series linearization method, balanced repeated replication, and the jackknife method.

The national health-care survey public use files generally include variables to obtain weighted estimates and to implement a Taylor series approach to estimate standard errors for weighted survey estimates. These variables, which jointly reflect the underlying survey design, include the estimation weight, sampling strata, and primary sampling unit (PSU) (Korn and Graubard 1999). The documentation and codebook for the public use files should contain these survey design variables. For example, the documentation should include the person weight, stratum, and PSU variables.

Statistical software packages that are commonly used to estimate standard errors from complex multistage designs using the Taylor series linearization method include SAS<sup>®</sup> (version 8.2 or higher), SUDAAN<sup>®</sup>, Stata<sup>®</sup>, and SPSS<sup>®</sup> (version 12.0 or higher). The software packages vary with respect to the specific types of estimates and models that can be produced accounting for the complex survey design and the treatment of missing data. For complete information on the capabilities of each package, analysts need to refer to the appropriate software user documentation manuals. The websites for SAS, SUDAAN, Stata, and SPSS are <http://www.sas.com>, <http://www.rti.org>, <http://www.stata.com>, and <http://www.spss.com>, respectively. The R language also has a package for complex survey analysis. Information on this package can be found in the June 2003 R News newsletter available on the R website at <http://www.r-project.org>.

Standard errors for these national survey estimates are most accurate when the analytic file contains all of the sample persons (e.g., those with positive values for the person weight variable) and the appropriate syntax is used to analyze population subgroups. The table above provides examples of basic programming code for SAS, SUDAAN, Stata, and SPSS to generate estimates from MEPS person-level files for the survey variable that measures annual health-care expenditures, *totexp* (Table 3).

---

### Integrated Survey Designs: Analytical Enhancements Achieved through the Linkage of Surveys and Administrative and Secondary Data

The analytical capacity, quality, and data content of household-specific health and health-care surveys are visibly enhanced through integrated designs that feature one-to-one data linkages between surveys, administrative and secondary data, and future connectivity to electronic health records. The data linkages include direct matches to additional health and socioeconomic measures acquired for the same set of sample units from

**Table 3** Example of software codes for analysis of complex survey data

SAS	SUDAAN	Stata	SPSS
proc surveymeans; stratum varstr;cluster varpsu;weight perwt; var totexp;	proc descript filetype=sas design=wr;nest varstr varpsu;weight perwt;var totexp;	svyset [pweight=perwt], strata(varstr) psu (varpsu) svymean totexp	csplan analysis/plan file='filename'/ planvars analysis weight=perwt/ design strata=varstr cluster=varpsu/ estimator type=wr.csdescriptives/ plan file='filename'/summary variables=totexp/mean/statistics se.

Source: [http://www.meps.ahrq.gov/mepsweb/survey\\_comp/standard\\_errors.jsp](http://www.meps.ahrq.gov/mepsweb/survey_comp/standard_errors.jsp)

other sources of survey specific or administrative data, in addition to linkage to existing secondary data sources at higher levels of aggregation (both geographic and organizational). One of the more pervasive uses of existing large-scale national surveys or administrative data bases is to serve as a sampling frame to facilitate a cost-efficient identification of an eligible survey population for purposes of sample selection, such as the consideration of the NHIS to serve as a sampling frame for the MEPS and Medicare administrative records to serve as a sampling frame for a survey of Medicare beneficiaries. Health surveys that are so linked to related surveys and/or administrative records from their inception benefit by this capacity for data supplementation that permits enhanced and more extensive analyses that are beyond the more constrained scope of the core health survey. In addition, the use of related surveys or administrative data as sampling frames for health-care surveys often permits enhanced longitudinal analyses when the host sampling frame and the core survey represent successive time intervals and share comparable data elements.

The large majority of the nationally representative population-based health surveys sponsored by the Department of Health and Human Services have also benefited by a capacity to link the survey data to county-level data on health service resources and health manpower statistics available on the Area Resources File. More specifically, the ARF is a county-specific health resources information system containing information on health facilities, health professions, measures of resource scarcity, health status, economic activity, health training programs, and socioeconomic and environmental characteristics. Geographic codes and descriptors are provided to enable linkage to health surveys to expand

analyses conducted by planners, policymakers, researchers, and other professionals examining the nation's health-care delivery system and in factors that may impact health status and health care in the USA. Comparable enhancements to health surveys for supplementation of economic indicators are achievable through linkage of survey data to the socioeconomic indicators made available by the Bureau of the Census through the County and City Data Book and public use files from the decennial census.

Other examples of improved data quality, content, and analytic capacity include linkages between individuals in household-specific health-care surveys with the medical providers and facilities that treat them and with the employers that are the source of their health insurance coverage benefits. In terms of data quality, household reported medical conditions can be evaluated for accuracy relative to medical specific records on medical conditions for the same patient and specific health events. With respect to health-care expenditures collected from household respondents for their reported health-care events, available linked medical provider level data is a more accurate source of information. The availability of such supplemental data on use and expenditures allows for the conduct of methodological studies to evaluate the accuracy of household reported data and informs adjustment strategies to household data in the absence of provider-specific data to reduce bias attributable to response error. To the extent these linkages to provider and employer records include data for time periods beyond the scope of the household surveys; the linkage between survey and administrative data also permit enhanced longitudinal analyses (Cohen et al. 2005).

## **An Example of Survey Integration: The Medical Expenditure Panel Survey**

One of the core health-care surveys in the USA, the Medical Expenditure Panel Survey (MEPS), is characterized by an integrated survey design. Since its inception, the primary analytical focus of the MEPS has been directed to the topics of health-care access, coverage, cost, and use. Over the past several years, the MEPS data have supported a highly visible set of descriptive and behavioral analyses of the US health-care system (Cohen et al. 2009). These include studies of the population's access to, use of, and expenditures and sources of payment for health care, the availability and costs of private health insurance in the employment-related and non-group markets, the population enrolled in public health insurance coverage and those without health-care coverage, and the role of health status in health-care use, expenditures, household decision making, and health insurance and employment choices. As a consequence of its breadth, the data have informed the nation's economic models and their projections of health-care expenditures and utilization. The level of the cost and coverage detail collected in the MEPS has enabled public and private sector economic models to develop national and regional estimates of the impact of changes in financing, coverage, and reimbursement policy, as well as estimates of who benefits and who bears the cost of a change in policy. The MEPS consists of a family of three interrelated surveys: the Household Component (HC), the Medical Provider Component (MPC), and the Insurance Component (IC). The survey is sponsored by the Agency for Healthcare Research and Quality (AHRQ).

The MEPS Household Component was designed to provide annual national estimates of the health-care use, medical expenditures, sources of payment, and insurance coverage for the US civilian noninstitutionalized population. In addition to collecting data to yield annual estimates for a variety of measures related to health-care use and expenditures, MEPS also provides estimates of measures related to health status, demographic characteristics, employment, and access to health care. Estimates

can be provided for individuals, families, and population subgroups of interest. The data collected in this ongoing longitudinal study also permit studies of the determinants of the use of services and expenditures and changes in the provision of health care in relation to social and demographic factors such as employment or income, the health status and satisfaction with health care of individuals and families, and the health needs of specific population groups such as the elderly and children.

The set of households selected for the Household Component is a subsample of those participating in the National Health Interview Survey (NHIS), an ongoing annual household survey of approximately 40,000 households (100,000 individuals) conducted by the National Center for Health Statistics and Centers for Disease Control and Prevention, to obtain national estimates of health-care utilization, health conditions, health status, insurance coverage, and access (Botman et al. 2000). In addition to the cost savings achieved by eliminating the need to independently list and screen households, selecting a subsample of NHIS participants has resulted in an enhancement in analytical capacity of the resultant survey data. The use of the NHIS data in concert with the data collected for the MEPS provides an additional capacity for longitudinal analyses not otherwise available. Furthermore, the large number and dispersion of the primary sampling units (~200) in MEPS has resulted in improvements in precision over prior expenditure survey designs.

The survey consists of an overlapping panel design in which any given sample panel is interviewed a total of five times in person over 30 months to yield annual use and expenditure data for two calendar years. These rounds of interviewing are spaced about 5–6 months apart. The interview is administered through a computer-assisted personal interview mode of data collection and takes place with a family respondent who reports for him/herself and for other family members. Currently, the MEPS sample consists of 14,000 families and 32,000 individuals and reflects an oversample of the following policy-relevant population subgroups:



Hispanics, blacks, and Asians. Data from two panels are combined to produce estimates for each calendar year.

The MEPS Medical Provider Component is a survey of the medical providers, facilities, and pharmacies that provided care or services to sample persons. The primary objective is to collect detailed data on the expenditures and sources of payment for the medical services provided to individuals sampled for the MEPS. Such data are essential to improve the accuracy of the national medical expenditure estimates derived from the MEPS, since household respondents are not always the most reliable source of information on medical expenditures. The data also serve as a primary imputation source of medical expenditure data to correct for the item nonresponse on this measure by the MEPS household sample participants.

The MEPS Insurance Component was designed to produce national and state-level estimates of the cost of employer-sponsored coverage. National, regional, and state-level estimates can be made of the amount, types, and costs of job-related health insurance. Interviews are conducted annually via mail with 30,000 establishments to obtain national and state-specific estimates of the availability of health insurance at the workplace, the type of coverage provided by employers, and the associated costs of coverage.

### **Advantages of Integrated Survey Designs**

The original MEPS sample design called for an independent screening interview to identify a nationally representative sample and facilitate oversampling of policy-relevant population subgroups. Detailed information was to be obtained on sociodemographic, economic, and health status measures to support an oversample of the following policy-relevant groups:

- Adults (18 years and older) with functional impairments
- Children with limitations of activity

- Individuals 18–64 years who were predicted to incur high medical expenditures
- Individuals predicted to have family income less than 200% of the poverty level

Detailed probabilistic models were to be used to target the oversample of individuals likely to incur high levels of expenditures in addition to those with family incomes less than 200% of the poverty level. Data collection and training costs associated with this independent screening interview were projected to exceed several million dollars. As part of the DHHS Survey Integration Plan, this separate screening interview was eliminated. Instead, NHIS was specified as the sampling frame for MEPS. In addition to the cost savings achieved by substituting NHIS as the MEPS sample frame, the design modification resulted in enhanced analytic capacity of the resultant survey data. The use of the NHIS data in concert with the MEPS data provides an additional capacity for longitudinal analyses not available in the original design. Furthermore, the greater number and dispersion of the sample primary sampling units that comprise the MEPS national sample resulted in improvements in precision over the original design specifications. These features are in clear contrast to new frame construction and/or independent screening interviews that characterize unlinked survey design efforts.

The integrated survey design model also provides additional features with respect to improving data collection strategies tied to the core survey to better ensure that target response rates are achieved. When the core survey is linked to a larger host survey, the survey operations and field staff that are armed with detailed record of calls data from the host survey will be better poised to commit and target necessary nonresponse conversion techniques to those cases that included reluctant or hard to reach respondents in the prior data collection effort.

*Capacity to Reduce Bias Attributable to Survey Nonresponse* As a consequence of the complex design of the MEPS HC, the MEPS sample data must be appropriately weighted to obtain

approximately unbiased national estimates for the US civilian noninstitutionalized population. The MEPS estimation weights are built from the estimation weights developed for the NHIS. The use of a sampling weight that has already incorporated the selection probabilities of the sample design and appropriate nonresponse and post-stratification adjustments is an added feature of the integrated survey design. Since survey nonresponse is potentially a significant source of bias in survey estimates, the MEPS dwelling unit sampling weights included an adjustment to help reduce its potential for bias. In general, the greater the difference among subgroups in response rates and the analytic characteristic(s) of interest, the greater is the need to adjust survey weights for nonresponse. In the absence of an integrated survey design, the nonresponse adjustment strategy adopted for the MEPS would be constrained to sociodemographic and economic information that were available at the geographic level (e.g., county, state, division, and region), rather than the detailed information available for each household participant in the NHIS sample selected for the MEPS. This is typical of standard household surveys which use aggregate data at the geographic level to inform the nonresponse adjustments (e.g., per capita income for the county based on secondary data available from the Census, physicians per 1000 populations and other health manpower statistics at the county-level available from the Area Resources File). In the absence of an integrated survey design for the MEPS, none of the household-specific information that were factors in the nonresponse adjustments would be available, other than the measures of MSA size and region. Clearly the MEPS linkage to the NHIS enhances the capacity of the specification of more direct nonresponse adjustments to better correct for survey nonresponse.

Another survey that benefits by this integrated design model is the Medicare Current Beneficiary Survey (MCBS) sponsored by the Centers for Medicare and Medicaid Services. The MCBS is a continuous, multipurpose survey of a nationally representative sample of aged, disabled, and institutionalized Medicare beneficiaries. It provides a

comprehensive source of information on the health status, health-care use and expenditures, health insurance coverage, and socioeconomic and demographic characteristics of the entire spectrum of Medicare beneficiaries. Rather than being linked to a larger survey, the sample for MCBS is drawn from administrative records in CMS's Medicare enrollment file. The Medicare enrollment files also provide mailing addresses for the sample. Medicare administrative files provide not only the sample frame but also service, diagnosis, and charge details for covered events, month-by-month information on enrollment status, payments for Medicaid buy-ins and HMO membership, and data for nonrespondents to the interview.

### **Linked Provider Data on Expenditures Improves the Accuracy of National Medical Expenditure Estimates in the MEPS**

The MEPS Medical Provider Component (MPC) was primarily designed to reduce the bias associated with national medical expenditure estimates derived from household reported data. The estimation strategy that has been considered to support the data replacement strategy is comprehensive in nature, making full use of MPC data to correct for missing and poor-quality household reported expenditure data. In addition, it provides the basis for a recalibration of household reported data, if significant reporting differentials are observed in expenditure data between households and medical providers.

### **Integrated Design Expands Capacity for Longitudinal Analyses**

The MEPS survey integration with the National Health Interview Survey (NHIS) permits an enhanced capacity for longitudinal analyses of trends in health-care utilization, coverage, access, and health status. The parallel structures of the two surveys make their integration for longitudinal

analyses easier to accomplish. To facilitate the conduct of longitudinal cohort analyses using the NHIS and MEPS data in tandem, NHIS/MEPS linkage files have been developed. These NHIS/MEPS linkage files allow users to link persons in the MEPS public use files to the records of the same persons in the previous NHIS public use files. Examples of enhanced longitudinal analyses based on the NHIS-MEPS linked files include studies of the long-term uninsured and the conduct of episodes of illness studies over an extended time interval.

### **Integrated Design of MEPS Facilitates Examination of Response Error**

In addition to serving as the primary source for the expenditures in the MEPS, the design of the Medical Provider Component provides data that could potentially facilitate adjustments to household reported utilization data that correct for reporting errors (both underreporting and overreporting (telescoping errors)), under the assumption that the medical provider reports are the “gold standard.” Within a given event type, the number of reported events can be aggregated up to the person-provider pair level. The distribution of the difference in utilization counts between the medical provider and household reports can then be examined. For each event type at the person-provider level ( $ij$ ), a difference measure,  $DIFF_{ij}$ , may be computed, where:

$$DIFF_{ij} = MPSCOUNT_{ij} - HHSCOUNT_{ij}$$

$MPSCOUNT_{ij}$  = the number of events for the person-provider pair reported in provider survey

$HHSCOUNT_{ij}$  = the number of events for the person-provider pair reported in household survey

The use of MPC data to develop adjustment factors that recalibrate or correct household reported data to reflect utilization counts based on MPC data offers a capacity to inform a utilization adjustment to correct for potential response

error associated with household reports. While the development of adjustment factors that correct for both underreporting and overreporting of health-care utilization by household respondents is permissible, which would allow for household event counts to be either scaled down or up, based on reported or imputed MPS information, an alternative approach would be to limit the adjustment to correct the outlier cases (the poorest household reporters of utilization).

### **Constraints**

It is important to note that several of the desired features of an integrated survey design are the sources of its most prominent limitations. As a consequence of acquiring more information on survey respondents through data augmentation and data linkages over time, these analytical enhancements also increase the potential for disclosure of confidential information. To guard against this, it is necessary to impose greater restrictions on the release of data to the public. The sponsorship and operation of a data center to ensure that confidential data is in a secure environment while permitting more detailed analyses to be conducted with the nonpublicly available data offers a compromise between greater data access and achieving confidentiality protection of data. However, this investment in the development and operation of a secure data center requires additional funds that may compete with sample size enhancements or planned research efforts.

An integrated survey design also requires greater coordination across data sources and organizations. There are often competing demands on the host sample frames that may limit the full benefits of an integrated design from being realized. Furthermore, the enhanced longitudinal data that comes with an integrated survey design will often be characterized by more frequent survey contacts and rounds of data collection which will impact the overall survey response rate. When properly designed and coordinated, as implemented for the MEPS, the integrated survey design remains an attractive model for consideration and adoption.

## **Policy-Relevant Examples from the Medical Expenditure Panel Survey (MEPS)**

### **Design of the MEPS to Inform Health Policy and Health Services Research**

The MEPS research program, broadly defined to encompass data collection, data development, research, and the translation of research into practice, is directly tied to the strategic goal of identifying strategies to improve access, foster appropriate use, and reduce unnecessary expenditures. Few other surveys provide the foundation for estimating the impact of changes on different economic groups or special populations of interest, such as the poor, elderly, veterans, uninsured, and racial/ethnic groups. The public sector relies upon the MEPS research findings to evaluate health reform policies, the effect of tax code changes on health expenditures and tax revenue, and proposed changes in government health programs such as Medicare. In the private sector, these data are also used to develop economic projections.

The Medical Expenditure Panel Survey (MEPS), initiated in 1996, is designed as a continuous ongoing survey to permit annual estimates of health-care utilization, expenditures, insurance coverage, and sources of payment for the US civilian noninstitutionalized population. Over the past several years, the MEPS data and associated research findings have quickly become a linchpin for the nation's economic models and their projections of health-care expenditures and utilization. This combination of breadth and depth of the data enables public and private sector analysts to develop economic models designed to produce national and regional estimates of the impact of changes in financing, coverage, and reimbursement policy, as well as estimates of who benefits and who bears the cost of a change in policy. Since 1977, AHRQ's expenditure surveys have been an important and unique resource for public and private sector decision makers. The survey is unique in the level of detail of information obtained on the health-care services used by Americans at the household level and their associated expenditures

(for families and individuals); the cost, scope, and breadth of private health insurance coverage held by and available to the US population; and the specific services purchased through out-of-pocket and/or third-party payments.

The MEPS data support a wealth of basic descriptive and behavioral analyses of the US health-care system. These include studies of the population's access to, use of, and expenditures and sources of payment for health care, the availability and costs of private health insurance in the employment-related and non-group markets, the population enrolled in public health insurance coverage and those without health-care coverage, and the role of health status in health-care use, expenditures, household decision making, and health insurance and employment choices (Cohen et al. 2009; Cohen 2003).

Efforts to address inequities in the availability of private health insurance and to control health insurance premiums and medical care costs must necessarily focus on the employment-related health insurance market. Historically, the analyses of data from the MEPS family of surveys have figured prominently in this arena. As is evidenced in the recent Institute of Medicine (IOM) Report on "Health Insurance is a Family Matter," the report notes that "the most comprehensive data on who uses what health-care service and how much is paid for those services comes from the Medical Expenditure Panel Survey". MEPS-related analyses are prominently used to inform components of this IOM report focused on issues of insurance coverage and cost.

MEPS-derived estimates of the health insurance status of the US civilian noninstitutionalized population are critical to policymakers and others concerned with access to medical care and the cost and quality of that care. Health insurance helps people get timely access to medical care and protects them against the risk of expensive and unanticipated medical events. When estimating the size of the uninsured population, it is critical to consider the distinction between those uninsured for short periods of time and those who are long-term uninsured across several years in duration. Compared to people with health-care coverage, uninsured people are less likely to visit a doctor,

have a usual source of medical care, receive preventive services, or have a recommended test or prescription filled. Consequently, individuals that experience extended periods of being uninsured are particularly at risk for restrictions in access to care and exposure to serious illness and significant financial jeopardy. Since many individuals undergo transitions in the acquisition and loss of health insurance coverage over time, an important consideration is the length of duration of spells of uninsurance and the capacity of this lack of coverage to lead to less efficient use of health-care services and facilities. In this regard, MEPS research efforts have demonstrated that individuals who experience short spells of being uninsured differ significantly from those who have been uninsured for more than a year on several dimensions which include access to employer-sponsored coverage, their attitudes and preferences regarding the need for coverage, and their sensitivity to the cost of acquiring coverage. In addition to providing cross-sectional estimates of health insurance coverage each year, the MEPS has the added analytical capacity to identify individuals with gaps in coverage over time as well as the duration of the spells of being uninsured for up to 4 years.

In addition to measuring actual out-of-pocket financial burdens for health care, MEPS provides the only nationally representative data that can be used to measure the extent of underinsurance in the USA. Underinsurance is defined as being at risk of spending more than a certain amount of family income on out-of-pocket expenses in the event of a catastrophic medical illness. Estimates of the underinsured require linked information on families health insurance benefits, family income, and risk of experiencing catastrophic medical events.

With health-care absorbing increasing amounts of the nation's resources, the question of how to implement health system design innovations that encourage the provision of high-quality and efficient health-care delivery is a sentinel concern of both private and public payers. To effectively address this issue, researchers and policymakers have benefited from MEPS research findings to better understand how individual characteristics, behavioral factors, financial incentives, and

institutional arrangements affect health-care expenditures in a rapidly changing health-care market. Research findings for the MEPS have also served to provide health-care decision makers with a better understanding of the highly concentrated nature of health-care expenditures and the persistence of these high expenditures over time. MEPS studies that examine the persistence of high levels of expenditures over time have been essential to help discern the factors most likely to drive health-care spending and the characteristics of the individuals who incur them.

Recently, greater attention and prioritization have been given to data collection procedures, predictive modeling, and estimation strategies that help improve the precision and quality of the survey estimates that characterize this policy-relevant population subgroup of individuals with high levels of medical expenditures. Research findings from MEPS also provide clear evidence of the utility and appropriateness of probabilistic models as prediction tools for identifying individuals likely to incur high levels of medical expenditures in future years. To the extent that this policy-relevant subset of the population is amenable to successful prediction through the application of well-developed models, the methodology continues to find several venues for application. Prominent examples of applications ripe for implementation include adoption of oversampling strategies for national health-care surveys and the identification of individuals whose health status improvements through disease management programs could most significantly result in potential reductions in overall future year health-care expenditures.

Given the growing attention focused on achieving a better understanding of the impact of rising prescribed medicine costs on health and the consumption of health services, it is also important to note the utility of the MEPS to inform studies examining the association between the use of newer medicines and morbidity, mortality, and health spending. Using this data resource, researchers have been able to determine the direction of the association between the use of newer drugs and all other types of nondrug medical spending. Attention has also focused on studies that identify inappropriate medication use, which

is a major patient safety concern and has significant consequences with respect to health-care costs. With its wealth of data on health conditions, prescribed medication utilization and expenditures, and associated therapeutic drug classifications, the MEPS data have also been helpful to researchers attempting to identify potentially inappropriate medication use in the community.

## **Issues on Measuring and Estimating Health Insurance Coverage in Surveys**

### **Testing for the Impact of Survey Attrition on Health Insurance Coverage Estimates in the MEPS**

The following study illustrates a test to assess the quality of the nonresponse adjustments employed in the MEPS to adjust for potential nonresponse bias attributable to survey attrition. The overlapping panel design of the MEPS survey is particularly well suited to inform these studies. This comparison of the stability of national estimates of health insurance coverage, subject to varying levels of survey attrition, made use of a model-based analysis that included additional controls for other predispositional factors. More specifically, a multivariate analysis was conducted to discern the influence of survey attrition on predicting the likelihood of being uninsured after controlling for sociodemographic and economic factors associated with this coverage measure. Building on previous research efforts that have identified salient factors associated with the presence or absence of health insurance coverage, a logistic regression model was developed to consist of the subset of significant predictors that distinguished the uninsured from those with either public or private coverage (Cohen et al. 2006b; Cohen 2003).

Using data from the 2002 MEPS for individuals between the ages of 18 and 64, the following factors were determined to be significant correlates in distinguishing between individuals likely to be uninsured for the entire calendar year from their counterparts with some coverage: age, gender, race/ethnicity, living in MSA, marital status,

health status, presence of limitation in activity, level of education, poverty status, born in USA, and total health-care expenditures (Cohen 2003). Once these measures were controlled for in the logistic regression model, it was possible to determine whether an individual's classification with respect to MEPS panel (year 1 of panel 7 vs. year 2 of panel 6), which varied significantly in terms of level of survey attrition, influenced the prediction of the likelihood of being uninsured in a calendar year. Under the assumption that the two distinct MEPS panels that are combined to produce annual survey estimates were characterized by the same survey response rates, one would not expect to observe a significant panel effect. Given the higher level of nonresponse across MEPS panels, where the older panel is affected by greater levels of survey attrition, a test for a MEPS panel effect affords the opportunity to assess the influence of unadjusted components of survey attrition on health insurance coverage estimates in a modeling context. The results of the logistic regression analysis reveal no significant effect for MEPS panel classification in distinguishing the full-year uninsured individuals from their insured counterparts (Table 4), when testing at an alpha level of .05. These results serve to further reinforce the efficacy of the estimation strategies adopted in the MEPS to correct for the impact of survey attrition on health insurance coverage estimates and related model-based studies.

### **Analyses Based on NHIS to MEPS Linkage**

In addition to the within MEPS studies, the linkage of the MEPS to the NHIS permits a related set of analyses to be conducted to discern the impact of survey attrition on national estimates. The design permits appending to the MEPS sample the data profiles collected in the NHIS for the prior year. Using the NHIS data in concert with the restricted sample of MEPS respondents permits the derivation of national estimates for the prior year based on a NHIS subsample characterized by a lower response rate. Using this design feature, the national estimates derived from the MEPS sample, affected by survey attrition, may be compared to the national estimates obtained from the full NHIS, prior to its linkage to MEPS.

**Table 4** Logistic regression analysis of the uninsured, testing for panel effects, US civilian noninstitutional population, ages 18–64, 2002

Contrast	Degrees of freedom	Wald F	P-value Wald F
Overall model	27	137.91	0.0000
Model minus intercept	26	72.69	0.0000
Panel classification	1	1.08	0.2989
Sex	1	63.48	0.0000
Race/ethnicity	3	18.60	0.0000
Health status	4	3.52	0.0082
Limitation in activity	1	14.71	0.0002
Marital status	4	20.59	0.0000
Highest year of education	4	18.01	0.0000
Poverty status	4	62.87	0.0000
USBORN	1	83.88	0.0000
MSA status	1	4.02	0.0462
Income	1	34.34	0.0000
Total health-care expenditures	1	34.91	0.0000

–2 \* Normalized Log-Likelihood Full Model: 14,862.53

Pseudo Model R-Square: 0.175588

**Source:** Center for Financing, Access, and Cost Trends, AHRQ, Household Component of the Medical Expenditure Panel Survey, 2002

In the National Health Interview Survey, three distinct measures of health insurance coverage are collected as part of the annual survey. These measures determine insurance coverage status at the time of the interview, whether there was a period of being uninsured during the 12-month time frame preceding the interview, and the likelihood of being uninsured for durations that exceed a year from the time of the interview. Each year, CDC's National Center for Health Statistics releases national estimates of the uninsured based on these measures, determining the percent of the population uninsured at the time of the interview, uninsured for at least part of the past year, and uninsured for more than a year.

The cross-sectional nature of the NHIS design, and its status as the initial baseline interview for the MEPS, helps facilitate the achievement of a survey response rate that has often exceeded 90%. Given the nationally representative nature of the subsample of the NHIS used for the MEPS each year, one may produce national estimates of health insurance coverage using the NHIS measures for the reserved MEPS subsample (prior to the conduct of MEPS interviews) that are convergent with the estimates obtained from the full sample NHIS. Alternatively, national estimates

based on the same NHIS measures from the linked full-year 2002 MEPS survey will be characterized by a response rate subject to three additional rounds of interviewing and associated sample attrition. A comparison of the health insurance estimates, based on the NHIS variables derived from the sample restricted to MEPS with the full sample NHIS national estimates, permits another assessment of the impact of survey attrition on the resultant health insurance coverage estimates.

Table 5 provides a summary of the national health insurance estimates derived from the NHIS for calendar year 2001. In addition to including the overall estimates of health insurance coverage for the nation, and for the population under age 65, the table includes further breakdowns distinguished by age groups <18 and 18–64. National estimates of these NHIS measures from the MEPS are derived from the MEPS full-year responding sample linked to the prior year NHIS. Based on the full sample 2001 NHIS, the national estimates of being uninsured by specific time periods for the entire US civilian noninstitutional population were 14.2% at the time of the interview, 17.8% for at least part of the past year, and 8.8% for being uninsured for more than a year since the time of the interview.

**Table 5** Comparison of 2001 national estimates of the uninsured derived from the 2001 NHIS and the 2002 MEPS

Percent of uninsured individuals, civilian noninstitutionalized population (standard error)						
Age group	2001 estimates derived from the NHIS			2001 NHIS estimates based on 2002 MEPS		
	Uninsured at time of interview	Uninsured for at least part of the past year	Uninsured for more than a year	Uninsured at time of interview	Uninsured for at least part of the past year	Uninsured for more than a year
All ages	14.2 (0.23)	17.8 (0.26)	8.8 (0.17)	13.9 (0.52)	17.9 (0.58)	8.9 (0.45)
Under 65 years	15.9 (0.25)	20.0 (0.29)	9.9 (0.19)	15.6 (0.57)	20.0 (0.64)	10.0 (0.50)
18–64 years	18.0 (0.26)	22.0 (0.28)	11.6 (0.21)	17.6 (0.62)	22.2 (0.70)	11.7 (0.55)
Under 18 years	10.9 (0.34)	15.1 (0.41)	6.0 (0.24)	10.7 (0.76)	14.9 (0.89)	5.8 (0.60)

**Sources:** Center for Financing, Access, and Cost Trends, AHRQ, Household Component of the Medical Expenditure Panel Survey, 2002

National Center for Health Statistics, CDC, National Health Interview Survey, 2001

Restricting the sample to the full-year MEPS respondents for the subsequent year, the corresponding NHIS-specific national estimates of the uninsured were 13.9% at the time of the interview, 17.9% for at least part of the past year, and 8.9% for being uninsured for more than a year since the time of the interview. As can be observed from a review of the comparisons of the MEPS and NHIS-generated estimates of the uninsured, no significant difference in estimates are evident, when testing at an alpha level of .05. A comparison of the NHIS-derived and the MEPS-derived coverage estimates for the population under age 65 and for age groupings <18 and 18–64 revealed similar levels of convergence. Once again, the results present no evidence of nonresponse bias attributable to survey attrition affecting the national coverage estimates when subject to more restrictive response rate requirements in MEPS.

### The Utility of Prediction Models to Oversample the Long Term Uninsured

Estimates of the health insurance status of the US civilian population are critical to policymakers and others concerned with access to medical care and the cost and quality of that care. Health

insurance helps people get timely access to medical care and protects them against the risk of expensive and unanticipated medical events. When estimating the size of the uninsured population, it is important to consider the distinction between those uninsured for short periods of time and those who are uninsured for several years. Given the risk of exposure to high out-of-pocket medical expenditures faced by the long-term uninsured and associated economic and health-related consequences, this population subgroup is of particular relevance to health policy considerations. Consequently, a prediction model that can accurately identify the long-term uninsured is an important analytical tool. These models have particular relevance as statistical tools to facilitate efficient sampling strategies that permit the selection of an oversample of individuals likely to be uninsured for long periods in the future. This discussion provides a summary of the development of prediction models to identify the long-term uninsured adults under age 65 and includes an evaluation of its potential utility as an oversampling strategy for use in the Medical Expenditure Panel Survey, a core national longitudinal medical care expenditure survey with comprehensive data on health insurance status. This type of modeling effort also enhances the ability to discern the causes of being uninsured and the characteristics of the individuals who are without



coverage. This feature also applies to prediction models that can accurately identify those individuals with transitions in coverage or with no gaps in coverage over a given time interval.

To improve the precision of survey estimates that characterize policy-relevant population subgroups in a cost-efficient manner, oversampling strategies are traditionally included as a core survey design component and implemented in the sample selection phase. When the characteristics of the population that are targeted for an oversample are static in nature, and the sampling frame that will be utilized contains the essential data to facilitate accurate identification of the respective target subpopulation, the underlying conditions permit a straightforward application of disproportionate sampling techniques. Alternatively, when the characteristic of the population targeted for an oversample is subject to transitions over time, the oversampling strategy is subject to much greater uncertainties in terms of achieving the desired sample size enhancements. The greater the departure from a static characteristic, the more challenging the effort and the less certain the outcome. Other obstacles that further limit the successful application of oversampling strategies relate to the level of availability of the key measures essential for the identification of the targeted population subgroup. Consequently, when attention is directed to an effort that attempts to increase the sample yield in a survey of individuals likely to be long-term uninsured in the future, the operation is subject to both constraints at its inception: (1) the focus on a characteristic that is subject to change and (2) a restricted set of available predictor measures available on a sampling frame.

### **Analytical Framework: Model Development**

Given the analytical and substantive importance of those individuals that are without health insurance coverage for extended periods of time (in a given year or longer period duration), the development and specification of accurate models to predict the future likelihood of the occurrence of this event are highly desirable. At the outset, the specification of a clear definition of what

constitutes the long-term uninsured is critical. For this study, the ultimate objective was to develop the best model to predict the set of adults under the age of 65 who are without any health insurance coverage for two consecutive calendar years (Cohen and Yu 2009). With these parameters set, a logistic model specification was considered as most relevant for predicting the set of adults under age 65 most likely to be continuously uninsured for two consecutive calendar years. The longitudinal design of the MEPS, with two consecutive years of data on health-care coverage, use, and expenditures, was ideally suited to permit model development and evaluation.

The logistic model under consideration classified individuals without coverage for two consecutive calendar years as  $Y = 1$ , with all other individuals classified as  $Y = 0$ . Alternative definitions of the long-term uninsured such as lacking coverage for more than a year, being continuously uninsured for more than 2 years, are likewise viable. All the predispositional variables included as potential predictors were based on an individual's data profile prior to the 2-year period of interest. This modeling effort for predicting future health insurance coverage status builds off related efforts that were likewise limited to consideration of the immediate prior year's predispositional characteristics.

Several studies using MEPS data have identified factors associated with distinguishing individuals most likely to be characterized as the long-term uninsured (Selden and Hudson 2006; Short and Graefe 2003). Given the rare classification of children under the ages of 18 to be long-term uninsured (only 2% of children were continuously uninsured over the period 2002–2005), the modeling effort was further restricted to adults between the ages of 18 and 64. The precursor information characterizes an individual's status at a baseline period, which is defined as in the year prior to the 2-year period of analytical focus and interest. In developing the prediction model, a core set of potential predispositional measures were identified that were applicable to health insurance take-up models and readily available from a screener interview. These included age, gender, race/ethnicity, health status, limitations

in ability to work, marital status, education level (as measured by highest year of education completed), region, MSA status, presence of hospitalization, nativity in USA, family size, poverty status, and health insurance coverage status at time of screening (prior status). More specifically, the measure of prior coverage distinguished whether the individual was covered in the prior year or, if not, the period of time without coverage (<6 months, 6 months to <1 year, 1 year to <3 years, 3 years, or more years).

As part of this study, three alternative prediction models were fit to the longitudinal data from the MEPS, the 2004–2005 panel linked to the 2003 NHIS which served as both the MEPS sampling frame and screening interview. In this setting, Model 1 makes use of the full set of potential predictors that are available from the National Health Interview Survey for purposes of facilitating an oversample of individuals predicted to be long-term uninsured in the MEPS. To assess the performance of the fully specified model relative to a model based on a more restricted set of measures, two additional models were considered for comparative purposes. The second model that is considered (Model 2) is restricted to a single measure of one's insurance status at baseline, further distinguished by length of time without coverage for those uninsured at baseline. From a survey operations perspective, the straightforward application and limited data requirements of this model have particular appeal. Alternatively, the third model (Model 3) replicates the set of measures considered for Model 1 with the exclusion of the insurance status measure at baseline.

### **Likelihood of Being in the Continuously Uninsured in 2004–2005, Based on 2003 Profiles**

In the final logistic regression model developed for predicting adults between the ages of 18–64 likely to be continuously uninsured for two subsequent years, baseline health insurance status, race/ethnicity, marital status, education level (as measured by highest year of education completed), nativity in USA, income, and gender were determined to be significant predictors when testing at a .05 level of significance

among the potential set of explanatory measures under consideration (Table 6). The standard errors of all the survey estimates derived from the MEPS in this study and associated test statistics have been adjusted for the impact of clustering due to the multistage survey design and unequal weighting.

Individuals with the longest durations of prior spells without coverage were significantly more likely to be continuously uninsured over the subsequent 2-year period. Hispanics, males, and individuals born outside of the USA were also more likely to be continuously uninsured in the future. Furthermore, low-income individuals, those with less than 12 years of education, residence in the South or Midwest, and those who were never married in 2003 were associated with a greater likelihood of being classified as long-term uninsured for the period 2004–2005. Finally, a likelihood ratio test for the goodness of fit for this model rejected the null hypothesis that the model's coefficients were jointly equal to zero and the pseudo- $R^2$  for the model is 0.228 and it had the lowest Akaike information criterion (AIC = 4572.3). A receiver operating characteristic (ROC) analysis was also performed for each model, examining the area under the curve (AUC). The selected model also exhibited the highest AUC (.880).

Remarkably, the second model under consideration (Model 2) that was restricted to a single measure of one's insurance status, further distinguished by length of time without coverage for those uninsured at baseline, exhibited a relatively comparable goodness of fit and a pseudo- $R^2$  of 0.195 (Table 7). Alternatively, the third model (Model 3) which replicated the set of measures considered for Model 1 with the exclusion of the insurance status measure at baseline exhibited less powerful goodness of fit and the lowest pseudo- $R^2$  of 0.130.

### **Determination of the Cutoff Threshold in Predicted Probability to Facilitate Oversampling**

Once these predictive models to identify the likelihood of being continuously uninsured have been developed, additional analyses are necessary to

**Table 6** Logistic regression model to identify individuals aged 18–64 likely to be continuously uninsured in 2004–2005, based on 2003 profiles (2004–2005 MEPS, 2003, NHIS)

Independent variables and effects	Beta coeff.	SE beta	T-test B = 0	P-value T-test B = 0
Intercept	-2.30224	0.24615	-9.35282	0.00000
Sex				
Female	-0.52614	0.08397	-6.26584	0.00000
Race/ethnicity recode				
Hispanic	0.53787	0.14531	3.70152	0.00027
Non-Hispanic Black	-0.05864	0.17903	-0.32753	0.74355
Non-Hispanic Others	-0.36421	0.30747	-1.18455	0.23737
Region				
Midwest	0.54791	0.18237	3.00437	0.00294
South	0.86221	0.17992	4.79216	0.00000
West	0.37245	0.20332	1.83179	0.06822
MARITL				
Married/DK	-0.71323	0.11881	-6.00321	0.00000
Widowed/divorced/separated	-0.33652	0.14520	-2.31754	0.02132
Living w/partner	-0.32202	0.17812	-1.80787	0.07188
EDUCYR				
12 years/GED	-0.19730	0.11301	-1.74583	0.08212
Some college/DK	-0.37671	0.12214	-3.08426	0.00228
BA/BS degree	-0.74324	0.19376	-3.83588	0.00016
Adv degree	-0.91486	0.22327	-4.09753	0.00006
USBORN				
No/DK	0.58786	0.13694	4.29267	0.00003
INCOME				
\$20K–\$75K	-0.38947	0.10632	-3.66309	0.00031
\$75K+	-0.72273	0.21587	-3.34795	0.00095
How long since last had health coverage covered				
<= 6 months	1.72635	0.17993	9.59470	0.00000
6 months–1 year	2.00910	0.19350	10.38289	0.00000
1–3 years	2.32315	0.14945	15.54488	0.00000
3 years+/DK	2.93583	0.11909	24.65196	0.00000
Sample size:				8888
Pseudo-R <sup>2</sup> :				0.228
-2 *Normalized Log-Likelihood Full Model:				4528.34
Approximate chi-square (-2 * Log-L Ratio):				2298.75
Degrees of freedom:				21

**Source:** 2004–2005 Medical Expenditure Panel Survey, Center for Financing, Access and Cost Trends, Agency for Healthcare Research and Quality; 2003 NHIS, NCHS, CDC

identify the appropriate cutoff threshold in predicted probability for screening purposes to facilitate an oversample of this target population. To accomplish this determination of an operational cutoff point for each model, the predicted probabilities of being identified as continuously uninsured were determined for each sample individual based on the underlying model

specification and rank ordered from highest probability to lowest. The predicted probability of being uninsured for two consecutive years in the future ( $P = \text{Exp}(y)/(1 + \exp(y))$ ) was derived from a transformation of an individual’s predicted log odds ( $y$ ) based on the respective prediction model under consideration. Based on MEPS longitudinal data, 12.9% of the US civilian

**Table 7** Logistic regression model to identify individuals ages 18–64 likely to be continuously uninsured in 2004–2005, based on 2003 coverage profiles (2004–2005 MEPS, 2003, NHIS)

Independent variables and effects	Beta coeff.	SE beta	T-test B = 0	P-value T-test B = 0
Intercept	−2.98189	0.07370	−40.45741	0.00000
How long since last had health coverage covered				
<= 6 months	1.99909	0.16837	11.87347	0.00000
6 months–1 year	2.33658	0.17966	13.00526	0.00000
1–3 years	2.69189	0.13720	19.61994	0.00000
3 years+/DK	3.48061	0.11050	31.49911	0.00000
Sample size:				8888
Pseudo-R <sup>2</sup> :				0.195
−2 * Normalized Log-Likelihood Full Model:				4894.40
Approximate chi-square (−2 * Log-L Ratio):				1932.69
Degrees of freedom:				4

**Source:** 2004–2005 Medical Expenditure Panel Survey, Center for Financing, Access and Cost Trends, Agency for Healthcare Research and Quality; 2003 NHIS, NCHS, CDC

noninstitutionalized population between the ages 18 and 64 were continuously uninsured for the period 2004–2005. Consequently, initial cutoff point for prediction classification was established by determining the value of the predicted probability above which the sum of the estimation weights associated with the MEPS sample participants represented the top 12.9% of the distribution of the ranked prediction probabilities of being long-term uninsured. As a consequence of the disproportionate sampling scheme adopted in the MEPS to facilitate oversampling of policy-relevant population subgroups, and additional adjustments to the estimation weights to adjust for nonresponse and poststratification, it was necessary to determine the cutoff point based on an estimated population-based distribution of predicted probabilities to insure greater applicability of the approach beyond the MEPS setting. This cutoff translated to a predicted probability of 0.355 (or log odds of  $-0.598$ ) based on the fully specified model (Model 1). Similarly, when considering the model that was restricted to a single measure of one's insurance status at baseline (Model 2), the cutoff translated to a predicted probability of 0.268 (or log odds of  $-1.006$ ). Alternatively, for the model which excludes the insurance status measure at baseline (Model 3), the cutoff translated to a predicted probability of 0.428 (or log odds of  $-0.290$ ).

With respect to those who were long-term uninsured, 59.5% were correctly identified by the model, based on the initial cutoff rule applied to the Model 1 predicted likelihood (model sensitivity; Table 8). In addition, of those with some coverage over the 2-year period, 94.0% were correctly identified (model specificity), based on their predicted likelihood relative to the cutoff threshold. When examining predictive capacity, 59.5% of individuals predicted to be long-term uninsured were correctly classified by the model. It was also observed that when considering higher values for the threshold cutoff (top 10%, top 5%), the potential predictive capacity of the model in identifying the long-term uninsured increased (Table 9). Using the top 5% as the threshold, the percent of those predicted to be long-term uninsured rose to 73.7%. However, this gain in model predictive capacity was at the expense of potential sample yield, given the greater restriction on the resultant eligible sample that fell above the threshold. When simultaneously considering model performance on predictive capacity, sensitivity, and specificity, while efficiently achieving accurate targeted yields from oversampling subject to fixed overall sample size constraints, adoption of the initial cutoff rule was the preferred approach. By establishing a cutoff rule in this manner, one has the capacity to implement a sample selection scheme permitting the oversampling of the long-term uninsured in “real time,” via a screening

**Table 8** Examination of the sensitivity, specificity, and predictive capacity of alternative cutoff values – model 1

Likelihood of lower pred.prob. of long-term uninsured	Logit cutoff	Predicted probability cutoff	Sensitivity	Specificity	True positive	False negative
0.8000	-1.8167	0.1398	0.7350	0.8789	0.4730	0.0427
0.8100	-1.6711	0.1583	0.7194	0.8882	0.4876	0.0446
0.8200	-1.5156	0.1801	0.7042	0.8975	0.5037	0.0465
0.8300	-1.3715	0.2024	0.6905	0.9068	0.5227	0.0480
0.8400	-1.1730	0.2363	0.6723	0.9157	0.5411	0.0502
0.8500	-0.9506	0.2788	0.6479	0.9236	0.5561	0.0534
0.8600	-0.7902	0.3121	0.6275	0.9320	0.5771	0.0558
0.8700	-0.6228	0.3491	0.5988	0.9392	0.5929	0.0594
<b>0.8714</b>	<b>-0.5975</b>	<b>0.3549</b>	<b>0.5951</b>	<b>0.9401</b>	<b>0.5948</b>	<b>0.0599</b>
0.8800	-0.4734	0.3838	0.5694	0.9463	0.6106	0.0630
0.8900	-0.3303	0.4182	0.5335	0.9526	0.6244	0.0675
0.9000	-0.1804	0.4550	0.4924	0.9579	0.6337	0.0726
0.9100	-0.0699	0.4825	0.4531	0.9636	0.6481	0.0774
0.9200	0.0788	0.5197	0.4180	0.9698	0.6719	0.0815
0.9300	0.2582	0.5642	0.3765	0.9751	0.6912	0.0864
0.9400	0.4694	0.6152	0.3317	0.9801	0.7112	0.0916
0.9500	0.6293	0.6523	0.2869	0.9849	0.7371	0.0967

**Source:** 2004–2005 Medical Expenditure Panel Survey, Center for Financing, Access and Cost Trends, Agency for Healthcare Research and Quality; 2003 NHIS, NCHS, CDC

**Table 9** Required sample size of adults 18–64 to yield a sample of 1760 individuals continuously without health insurance coverage over 2 years (50% increase)

	No model-based oversample	Model-based oversample: model 1 – fully specified model	Model-based oversample: model 2 – single baseline coverage measure	Model-based oversample: model 3 – excludes baseline coverage measure
Required overall sample size	15,000	11,058	11,028	11,512
Oversampling rate	N.A.	1.80	1.75	2.58
Model prediction rate – % correct predictions	N.A.	55.5%	57.1%	38.8%

Assumes base sample size of 10,000 individuals aged 18–64 in a MEPS panel responding for their entire 2-year period of eligibility in the survey

interview that collects the necessary input information required for the prediction model under consideration.

**Examination of the Sensitivity, Specificity, and Predictive Capacity of Alternative Probabilistic Models**

Once a parsimonious model was identified, which consisted of the subset of predictors that were all significant at the .05 level, the model was ready to be evaluated in terms of its accuracy in predicting

those adults under age 65 who would be continuously without coverage for the subsequent 2-year period. The performance of the model was assessed based upon an independent representative sample that characterizes the nation’s health insurance coverage experience. In this setting, the design of the MEPS is uniquely suited to this more rigorous criterion to assess model performance. This condition was satisfied through development of the prediction model using data from one specific MEPS longitudinal panel and then applying the model to an

independent MEPS longitudinal panel to assess model performance. Since the model was developed using MEPS data from the 2004 to 2005 longitudinal panel, the model was then applied to a different MEPS panel to assess performance. In addition, the model's performance was also assessed in relation to the two alternative prediction models under consideration (Model 2: prior coverage status only; and Model 3, no inclusion of prior coverage status).

Model performance was then evaluated based upon predictive capacity, sensitivity, and specificity, using the distinct predicted probability cutoff thresholds established with the 2004–2005 MEPS longitudinal panel for the three models. For the fully specified model (Model 1), the threshold cutoff point for selecting an oversample of long-term uninsured individuals was .355. Using the 2001 NHIS data in tandem with the associated model coefficients to derive a predicted probability of being continuously uninsured in 2002–2003, all individuals with a predicted probability of 0.355 or greater were targeted for an oversample. In the same manner, the threshold cutoff point established for Models 2 and 3 were 0.268 and 0.428, respectively. Based on MEPS longitudinal data, 11.7% of the US civilian non-institutionalized population between the ages 18–64 were continuously uninsured for the period 2002–2003. Using the predetermined cutoff points for the respective models, the overall percent of the population predicted to be classified as long-term uninsured by Model 1 was most consistent with the population estimate of 11.7% (11.6%), with both Models 2 and 3 yielding predicted population estimates below 10% based on the preestablished cutoff thresholds.

An assessment of the performance of the sensitivity of the alternative models to correctly identify individuals likely to be continuously uninsured in 2002–2003 indicated the logistic model that included prior year insurance coverage profiles in tandem with the significant sociodemographic predictors (Model 1) was superior. More specifically, Model 1 correctly identified 54.9% of those individuals who were continuously uninsured throughout 2002–2003. This was significantly better than the sensitivity of

Model 3, the prediction model that included the same sociodemographic predictors but excluded prior year insurance coverage status. Model 3 only correctly identified 31.2% of the long-term uninsured. Surprisingly, the model that was restricted to only measuring the prior year's health insurance coverage status (Model 2) was able to correctly identify 45.1% of the long-term uninsured. Generally comparable performance was observed when examining the alternative models with respect to specificity. Model 1's performed well with a specificity level of 94.1%, with Model 2 at 95.5% and Model 3 at 93.5%.

The next set of comparisons focused on the predictive capacity of the respective models, as measured as the percent of individuals with predicted probabilities of being long-term uninsured above the threshold cutoff point, who were correctly classified. More specifically, of those individuals predicted to be continuously uninsured throughout 2002–2003, Model 1 correctly predicted 55.5% of the target population. Again, this performance was significantly better than the predictive capacity of Model 3, the prediction model that included the same sociodemographic predictors but excluded prior year insurance coverage status. Model 3 only correctly predicted 38.8% of the target population. Alternatively, the model that was restricted to only measuring the prior year's health insurance coverage status (Model 2) exhibited the best performance in predictive capacity, with a correct prediction rate of 57.1% for the long-term uninsured.

The final set of comparisons in model performance are directly focused on the expected sample necessary to support a 50% increase in sample yield of individuals between the ages of 18–64, who are continuously uninsured over two consecutive calendar years. This enhanced sample size would yield significant improvements in the precision of survey estimates which characterized the long-term uninsured and associated population subgroups. The use of this metric facilitated an evaluation of the efficiency of a model-based oversampling strategy to yield the targeted sample, standardizing the comparison in terms of

sample size requirements under different model specifications. Using an assumption of a base sample requirement of 10,000 individuals aged 18–64 in a MEPS panel responding for their entire 2-year period of eligibility in the survey, the required sample size necessary to achieve a 50% sample size increase above the 1173 long-term uninsured survey participants was derived based on the estimated predictive capacity observed for the alternative models. This sample size specification calls for the inclusion of an additional 587 individuals with the characteristic, resulting in overall target sample yield of 1760 individuals who are long-term uninsured in the survey.

A summary of the overall sample size requirements to achieve a target sample of 1760 chronically uninsured individuals aged 18–64 in the survey is provided in Table 9. Model 1 performed quite well in terms of the necessary overall sample size to meet the target for the policy-relevant population subgroup under consideration. A sample design without access to the predictor variables from a screening interview such as the NHIS or a design without application of oversampling techniques would require an overall sample of 15,000 adults ages 18–64 to achieve the target. Alternatively, all of the model-based oversampling strategies were substantially more effective than the constrained approach, each requiring a substantially lower overall sample to achieve the targeted sample. In addition, the expected overall sample size specification for the model-based oversampling approach inherent in Model 1 was substantially more modest (11,058) and significantly more efficient than the model which excluded a baseline measure of health insurance status (Model 3). Remarkably, the model-based oversampling strategy that required the lowest overall sample was the model that considered only a single baseline insurance coverage status measure (Model 2).

---

## Summary

Policymakers, health-care leaders, and decision makers are particularly sensitive to recent trends in health-care costs, coverage, access, and health-

care quality and are dependent on accurate, reliable national estimates of these health-care parameters to help inform policy and practice. Health-care surveys serve as a critical source of this essential information, and the descriptive and analytical findings they generate are key inputs to facilitate the development, implementation, and evaluation of policies and practices addressing health care and health behaviors. To ensure their utility and integrity, it is essential that these health-care surveys are designed according to high-quality, effective, and efficient statistical and methodological practices and optimal sample designs.

This chapter serves to illustrate several survey methods that enhance the performance and utility of health services research efforts. Attention has been given to the topics of sample and survey designs, nonresponse and attrition, estimation, precision, sample size determination, and analytical techniques to control for survey design complexities in analysis. Several of the topics that are featured in this chapter are further connected by their substantive focus on the measurement of trends in health-care costs, coverage, access, and health-care utilization. In addition to highlighting underlying survey operations, estimates, and outputs, the topics that have been covered also serve to identify potential enhancements that facilitate improvements in design, data collection, estimation strategies, and ultimately analytical capacity for health services research efforts.

A well-designed health-care survey imposes an interdependence between the survey sponsors, the survey designers, the associated statisticians and methodologists, the survey operations, field and management staff, the data processing staff, and the end users, who are primarily the health researchers, policymakers, and the public. The survey methods covered in this chapter should help serve as a roadmap to help realize and strengthen these connections. When all the essential health-care survey contributors work in concert, following the methods covered in this chapter, the overall quality and utility that is achieved in the conduct of health services research should be much greater than the sum of the individual successful components.

## References

- Botman S, Moore T, Moriarity C, Parsons V. Design and estimation for the National Health Interview Survey, 1995–2004. National Center for Health Statistics. *Vital Health Stat.* 2000;2(130)
- Chromy J. Variance estimators for a sequential sample selection procedure. In: Krewski D, Rao J, Platek R, editors. *Current topics in survey sampling*. New York: Academic Press; 1981. p. 329–47.
- Cochran W. *Sampling techniques*. New York: Wiley; 1977.
- Cohen S. Methodology report #11: sample design of the 1997 Medical Expenditure Panel Survey household component. Rockville: Agency for Healthcare Research and Quality; 2000. [http://www.meps.ahrq.gov/data\\_files/publications/mr11/mr11.shtml](http://www.meps.ahrq.gov/data_files/publications/mr11/mr11.shtml)
- Cohen S. Design strategies and innovations in the Medical Expenditure Panel Survey. *Med Care.* 2003;4(7):5–12.
- Cohen S, Yu W. The utility of prediction models to oversample the long term uninsured. *Med Care.* 2009;47(1):80–7.
- Cohen S, DiGaetano R, Goksel H. Methodology report #5: estimation procedures in the 1996 Medical Expenditure Panel Survey household component. Rockville: Agency for Healthcare Research and Quality; 1999. [http://www.meps.ahrq.gov/data\\_files/publications/mr5/mr5.shtml](http://www.meps.ahrq.gov/data_files/publications/mr5/mr5.shtml)
- Cohen SB, Machlin S, Branscome J. Patterns of attrition and reluctant response in the 1996 Medical Expenditure Panel Survey. *J Health Serv Outcome Res Methodol.* 2000;1(2):131–48.
- Cohen S, Ezzati-Rice T, Yu W. Integrated survey designs: a framework for nonresponse bias reduction. *J Econ Soc Meas.* 2005;30(2-3):101–14.
- Cohen SB, Ezzati-Rice T, Yu W. The utility of extended longitudinal profiles in predicting future health care expenditures. *Med Care.* 2006a;44(5):45–53.
- Cohen S, Ezzati-Rice T, Yu W. The impact of survey attrition on health insurance coverage estimates in a National Longitudinal Health Care Survey. *J Health Serv Outcome Res Methodol.* 2006b;6:111–25.
- Cohen J, Cohen S, Banthin J. The Medical Expenditure Panel Survey: a national information resource to support healthcare cost research and inform policy and practice. *Med Care.* 2009;47(7):S1:44–50.
- Cox B, Cohen, S. *Methodological issues for health care surveys*. New York/Basel: 1985: Marcel Dekker
- Groves R, Heeringa S. Responsive design for household surveys: tools for actively controlling survey errors and costs. *J R Stat Soc Ser A.* 2006;169:439–57.
- Groves R, Fowler F, Couper M, Lepkowski J, Tourangeau R. *Survey methodology*. 2nd ed. New York: Wiley; 2009.
- Hansen M, Hurwitz W. The problem of nonresponse in sample surveys. *J Am Stat Assoc.* 1946;41:517–29.
- Korn E, Graubard B. *Analysis of health surveys*. New York: Wiley; 1999.
- Lynn P. *Methodology of longitudinal surveys*. Chichester: Wiley; 2009.
- Madans J, Cohen S. Health surveys: a resource to inform health policy and practice. In: *Health statistics in the 21st century: implications for health policy and practice*. New York: Oxford University Press; 2005.
- Rubin D. *Multiple imputation for nonresponse in surveys*. New York: Wiley; 1987.
- Selden T, Hudson J. Access to care and utilization among children: estimating the effects of public and private coverage. *Med Care.* 2006;44(5):19–26.
- Short P, Graefe D. Battery-powered health insurance? Stability in coverage of the uninsured. *Health Aff.* 2003; 22(6):244–55.
- Stinchcombe A, Jones C, Sheatsley P. Nonresponse bias for attitude questions. *Public Opin Q.* 1981;45: 359–75.
- Stoop I, Billiet J, Koch A, et al. *Improving survey response*. Chichester: Wiley; 2010.
- Vartivarian S, Jang S, Salvucci S, Kasprzyk D. Subsampling nonrespondents: Issue of calculating response rates. In: *Proceedings of the section on survey research methods*. Alexandria: American Statistical Association; 2006. p. 3796–8.





# Two-Part Models for Zero-Modified Count and Semicontinuous Data

# 28

Brian Neelon and Alistair James O'Malley

## Contents

<b>Introduction</b> .....	696
<b>Two-Part Models for Zero-Modified Count Data</b> .....	696
Hurdle Models .....	697
Zero-Inflated Count Models .....	699
Regression Models for Zero-Modified Count Data .....	700
Recent Developments .....	702
<b>Two-Part Models for Semicontinuous Data</b> .....	703
Two-Part Regression Models for Semicontinuous Data .....	704
Recent Developments .....	706
<b>Model Fitting, Testing, and Evaluation</b> .....	707
Zero-Modified Count Models .....	707
Semicontinuous Models .....	709
Model Comparison and Assessment .....	711
<b>Software</b> .....	712
<b>Conclusion</b> .....	712
<b>References</b> .....	713

## Abstract

Health services data often contain a high proportion of zeros. In studies examining patient hospitalization rates, for instance, many patients will have no hospitalizations, resulting in a count of zero. When the number of zeros is greater or less than expected under a standard count model, the data are said to be *zero modified* relative to the standard model. More precisely, the data are *zero inflated* if there is an overabundance of zeros, and *zero deflated* if there are fewer zeros than expected. A similar phenomenon arises with *semicontinuous* data,

---

B. Neelon (✉)

Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA  
e-mail: [brian.neelon@duke.edu](mailto:brian.neelon@duke.edu)

A. J. O'Malley

The Dartmouth Institute for Health Policy and Clinical Practice, Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

Department of Health Care Policy, Harvard Medical School, Boston, MA, USA  
e-mail: [Alistair.J.O'Malley@Dartmouth.edu](mailto:Alistair.J.O'Malley@Dartmouth.edu)

which are characterized by a spike at zero followed by a right-skewed continuous distribution of positive values. When dealing with zero-modified count and semicontinuous data, flexible two-part mixture distributions are often needed to accommodate both the excess zeros and the skewed distribution of nonzero values. A broad array of two-part models has been introduced over the past three decades to accommodate such data. These include hurdle models, zero-inflated models, and two-part semicontinuous models. While these models differ in their distributional assumptions, they each incorporate a two-part structure in which the zero and nonzero observations are modeled in distinct but related ways. This chapter describes recent developments in two-part modeling of zero-modified count and semicontinuous data and highlights their application in health services research.

---

## Introduction

In health services research, it is common to encounter data with an abundance of zeros. For example, in studies examining outpatient clinic visits, patients who report no visits will be assigned a count of zero. Likewise, in studies examining the frequency of screening mammography, patients who have never received a screening mammogram will have a response value of zero. Count-valued outcomes, like those in the previous two examples, are typically modeled using discrete distributions, such as the Poisson or negative binomial distribution. However, there are times when the proportion of zeros is greater or less than what a standard count distribution would predict, and in such cases the data are said to be *zero modified* relative to an ordinary count model. A related phenomenon occurs with *semicontinuous* outcomes, such as medical expenditures, which are characterized by a point mass at zero (representing, say, no expenditures) followed by a right-skewed continuous distribution for the positive values (representing positive expenditures). When dealing with zero-modified count

and semicontinuous data, parametric mixture distributions known as *two-part models* are typically needed to address both the abundance of zeros and the often highly skewed distribution of nonzero values.

Various two-part models have been developed in recent years to address zero-modified count and semicontinuous data, including hurdle models, zero-inflated models, and two-part semicontinuous models. While these models vary in terms of their distributional assumptions and parametric forms, they all incorporate an underlying, two-part structure in which the zero and nonzero observations are modeled through distinct (although sometimes overlapping) sets of parameters.

Sections “[Two-Part Models for Zero-Modified Count Data](#)” and “[Two-Part Models for Semicontinuous Data](#)” of this chapter provide overviews of zero-modified count and semicontinuous models, respectively. Section “[Model Fitting, Testing, and Evaluation](#)” discusses model fitting and evaluation strategies and highlights software packages commonly used to fit such models. The final section provides a summary, discusses potential limitations of two-part models, and points to directions for future research.

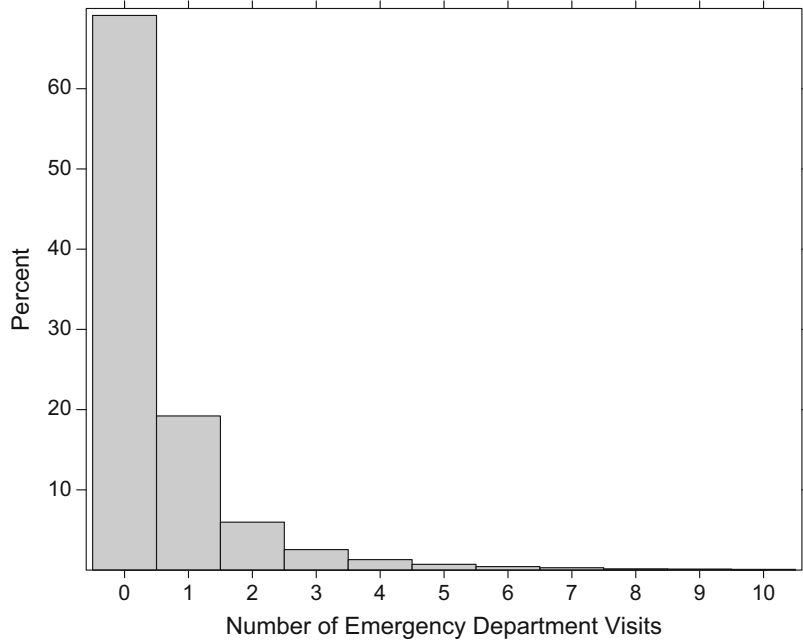
---

## Two-Part Models for Zero-Modified Count Data

Zero-modified count data arise frequently in health services research. Consider, for example, a recent study by Neelon et al. (2012) examining emergency department (ED) visits in Durham, North Carolina, during the 2009 calendar year. Figure 1 presents a partial histogram of the visits up to ten visits. The actual number of visits per patient ranged from 0 to 95, with an average of 0.65 visit per patient. Nearly 70% of the patients made no ED visits during the year, 19% had exactly one visit, 5% had exactly two visits, and the remaining 6% had more than two visits.

Now, suppose one is interested in building a statistical model to describe these data. A first step might be to assume that the data were generated according to a Poisson distribution with mean

**Fig. 1** Partial histogram of ED visits (up to ten visits)



parameter  $\mu = 0.65$ , the average number of ED visits in the sample. That is,

$$\Pr(Y = y) = \frac{0.65^y e^{-0.65}}{y!}, \quad y = 0, 1, \dots, \quad (1)$$

where  $Y$  denotes the number of ED visits. Although this seems like an intuitive (albeit somewhat basic) modeling choice, the model is not especially compatible with the observed data. Under this model, for instance, one would expect 52% zeros and 34% 1's – far fewer zeros and more 1's than were actually observed. When the number of zeros is greater than would be predicted under a standard count distribution, the data are said to be *zero inflated* relative to the standard model. Note that the abundance of zeros by itself is not necessarily problematic. For example, under a Poisson model with mean  $\mu = 0.35$ , one would expect approximately 70% zeros as observed in Fig. 1. However, this same model would predict fewer than 1% of the counts to be greater than two, clearly in conflict with the 6% observed in the data. Ordinary count distributions, therefore, become problematic primarily when there is an abundance of zeros coupled with a

longer than predicted right-tailed distribution of positive counts, since these features impose competing influences on the model. In the Poisson case, for example, the high proportion of zeros tends to lower the mean parameter,  $\mu$ , while large nonzero values tend to increase it. The term “zero inflation,” then, is customarily used to describe data in which a high proportion of zeros, together with a skewed distribution of nonzero counts, leads to a poor-fitting standard count model. More generally, the term *zero modification* is used to encompass both zero inflation and *zero deflation* (i.e., fewer than expected zeros). In the presence of zero modification, special two-part mixture distributions are often needed to provide adequate fit to the data. This section reviews common two-part models for zero-modified count data.

### Hurdle Models

The *hurdle model* (Mullahy 1986; Heilbron 1994) is a two-part mixture model consisting of a point mass at zero followed by a zero-truncated count distribution for the positive observations:

$$\begin{aligned} \Pr(Y = 0) &= 1 - \pi, \quad 0 \leq \pi \leq 1 \\ \Pr(Y = y) &= \frac{\pi p(y; \boldsymbol{\theta})}{1 - p(0; \boldsymbol{\theta})}, \quad y = 1, 2, \dots, \end{aligned} \tag{2}$$

where  $\pi = \Pr(Y > 0)$  is the probability of a nonzero response;  $p(y; \boldsymbol{\theta})$  is an untruncated, or *base*, probability distribution with parameter vector  $\boldsymbol{\theta}$ ; and  $p(0; \boldsymbol{\theta})$  is the base distribution evaluated at 0. This can be written more compactly as  $Y \sim (1 - \pi)1_{(y=0)} + \pi \frac{p(y; \boldsymbol{\theta})}{1 - p(0; \boldsymbol{\theta})} 1_{(y>0)}$ , where  $1_{(\cdot)}$  denotes the indicator function. Models such as the hurdle model are commonly referred to as “two-part” models because the zeros and nonzero counts are modeled separately, thereby accommodating zero modification. The expected value and variance of the hurdle model are given by

$$\begin{aligned} E(Y) &= \eta = \frac{\pi\mu}{1 - p(0; \boldsymbol{\theta})} \text{ and} \\ V(Y) &= \eta(\mu - \eta) + \frac{\pi\sigma^2}{1 - p(0; \boldsymbol{\theta})}, \end{aligned} \tag{3}$$

where  $\mu$  and  $\sigma^2$  denote the mean and variance of the base distribution, respectively. In health services research,  $\pi$  is known as the *utilization probability* – i.e., the probability of using services at least once. When  $1 - \pi = p(0; \boldsymbol{\theta})$ , the hurdle model reduces to its base distribution; when  $(1 - \pi) > p(0; \boldsymbol{\theta})$ , the data are zero inflated relative to the base distribution; and when  $(1 - \pi) < p(0; \boldsymbol{\theta})$ , there is zero deflation. In the extremes,  $\pi = 0$  or 1. When  $\pi = 1$ , there are no zero counts, and the model reduces to a truncated count distribution; when  $\pi = 0$ , there are no users (i.e., all counts equal zero), and the model is degenerate at zero. Typically, one assumes that  $\pi$  is strictly between 0 and 1, so that there is a nonzero utilization probability for all individuals under study, and hence all subjects are viewed as “potential” users, even if some do not actually use services during the study period.

Perhaps the most common choice for the base distribution is the Poisson distribution, which gives rise to the *Poisson hurdle model*:

$$\begin{aligned} \Pr(Y = 0) &= 1 - \pi, \quad 0 \leq \pi \leq 1 \\ \Pr(Y = y) &= \pi \frac{\mu^y e^{-\mu}}{y!(1 - e^{-\mu})}, \quad \mu > 0; \\ & y = 1, 2, \dots, \end{aligned} \tag{4}$$

where  $\mu$  is the mean of the ordinary (i.e., untruncated) Poisson. When  $(1 - \pi) > \exp(-\mu)$ , the data are zero inflated relative to an ordinary Poisson, and when  $(1 - \pi) < \exp(-\mu)$ , there is zero deflation.

Alternative hurdle models can be formed by selecting different base distributions, such as the negative binomial, the generalized power series (Patil 1962; Ghosh et al. 2006) or the generalized Poisson distribution (Consul 1989; Gschlößl and Czado 2008). The *negative binomial hurdle model*, for example, is given by

$$\begin{aligned} \Pr(Y = 0) &= 1 - \pi, \quad 0 \leq \pi \leq 1 \\ \Pr(Y = y) &= \frac{\pi}{1 - \left(\frac{r}{\mu+r}\right)^r} \frac{\Gamma(y+r)}{\Gamma(r)y!} \left(\frac{\mu}{\mu+r}\right)^y \left(\frac{r}{\mu+r}\right)^r, \\ & r, \mu > 0; y = 1, 2, \dots \end{aligned} \tag{5}$$

The negative binomial base distribution is appealing if there is evidence of *overdispersion* relative to the ordinary Poisson – that is, a variance exceeding the mean. The mean and variance of the negative binomial base distribution are given by  $\mu$  and  $\mu(1 + \mu/r)$ , respectively; hence,  $(1 + \mu/r)$  is a measure of overdispersion. As  $r \rightarrow \infty$  the negative binomial converges to a Poisson distribution with mean and variance equal to  $\mu$ . The connection between the negative binomial and Poisson distributions goes even further, since the former can be derived as a Poisson-gamma mixture. In particular, if  $W|\lambda \sim \text{Poi}(\lambda)$  and  $\lambda \sim \text{Ga}(r, \mu/r)$ , then the marginal distribution of  $W$  is negative binomial with mean  $\mu$  and variance  $\mu(1 + \mu/r)$ . Thus, the gamma prior, or “mixing,” distribution for  $\lambda$  induces excess variation relative to the Poisson. More generally, it can be shown that hurdle models are *more* overdispersed than their base distributions if and only if  $(1 - \pi) > p(0; \boldsymbol{\theta})$ , since in this case  $\frac{V(Y)}{E(Y)} > \frac{\sigma^2}{\mu}$ , where  $Y$  is distributed according to Eq. 2 and  $\mu$  and

$\sigma^2$  are the mean and variance of the base distribution, respectively. For example, the negative binomial hurdle distribution is more overdispersed than the ordinary negative binomial if  $(1 - \pi) > 1 - [r/(\mu + r)]^r$  or equivalently  $\pi < [r/(\mu + r)]^r$ . As a corollary, it follows that the Poisson hurdle model is overdispersed relative to the ordinary Poisson if and only if  $(1 - \pi) > \exp(-\mu)$  and underdispersed when  $(1 - \pi) < \exp(-\mu)$ . Thus, the Poisson hurdle model allows for both over- and underdispersion. Underdispersion arises when there are fewer zeros than expected under the ordinary Poisson model (Winkelmann 2008). As  $\mu \rightarrow \infty$ , the number of zeros expected under the ordinary Poisson model decreases, and the potential for underdispersion diminishes. For detailed discussions of over- and underdispersion in zero-modified count models, see Helibrón (1994), Gschlößl and Czado (2008), and Winkelmann (2008).

### Zero-Inflated Count Models

Zero-inflated count models are two-part mixtures consisting of a degenerate distribution at zero and an untruncated count distribution. These include the *zero-inflated Poisson (ZIP) model* (Lambert 1992) and the *zero-inflated negative binomial (ZINB) model* (Green 1994; Mwalili et al. 2008). The ZIP model is given by

$$\begin{aligned} \Pr(Y = 0) &= (1 - \phi) + \phi e^{-\mu}, & 0 \leq \phi \leq 1 \\ \Pr(Y = y) &= \phi \frac{\mu^y e^{-\mu}}{y!}, & \mu > 0; \quad y = 1, 2, \dots; \end{aligned}$$

or, alternatively,

$$Y \sim (1 - \phi)1_{(Z=0)} + \phi \text{Poi}(y; \mu)1_{(Z=1)}, \tag{6}$$

where  $Z$  is a (latent) indicator variable that takes the value 1 with probability  $\phi$ . The mean and variance of the ZIP model are  $E(Y) = \phi\mu$  and  $V(Y) = \phi\mu[1 + (1 - \phi)\mu]$ , respectively, and hence  $V(Y) > E(Y)$  and the model is overdispersed when  $\phi < 1$ . When  $\phi = 1$ , there is no zero inflation, and the model reduces to the ordinary Poisson with  $\Pr(Y = 0) = \exp(-\mu)$ . Conversely, when  $\phi < 1$ ,  $\exp(-\mu) < (1 - \phi) + \phi \exp(-\mu)$ ,

and the zeros are inflated relative to an ordinary Poisson distribution. Thus, unlike the Poisson hurdle model, the ZIP model accommodates only zero inflation. In fact, because zero-inflated count models can be rewritten as hurdle models with mixing probability  $\pi = \phi[1 - p(0; \theta)]$  (Neelon et al. 2010), they can be viewed as special cases of hurdle models in which only zero inflation and overdispersion are allowed. As with hurdle models, other base distributions can be chosen to model the counts in zero-inflated models. For example, the ZINB model is given by  $Y \sim (1 - \phi)1_{(Z=0)} + \phi \text{NB}(y; r, \mu)1_{(Z=1)}$ . For a comprehensive review of zero-inflated models, see Ridout et al. (1998).

Because each part of the mixture accommodates zeros, zero-inflated models such as the ZIP explicitly partition the zeros into two types: *structural* or *ineligibility zeros* (e.g., those that occur because a patient is ineligible for health services) and *chance* or *sampling zeros* (those that occur by chance among eligible patients). In the health services setting, the parameter  $\phi$  is known as the *eligibility probability*, and hence the random variable  $Z$  can be viewed as an “eligibility” indicator taking the value 1 if an individual is eligible for services and 0 otherwise. In this context, the parameter  $\mu$  represents the mean count among eligible subjects (i.e., given  $Z = 1$ ). In other settings, such as infectious disease epidemiology,  $\phi$  is known as the “at-risk” or “susceptibility” probability – i.e., the probability of belonging to an at-risk or susceptible population (Albert et al. 2011; Preisser et al. 2012). Note that the random variable  $Z$  is unobserved, since the observed outcome,  $Y$ , provides no direct information about individuals’ eligibility status, only whether they eventually used services as indicated by  $Y = 0$  or  $Y > 0$ . If  $Z$  were actually observed (e.g., through an eligibility screening process), then  $\phi$  could be estimated using the sample proportion of eligible patients and  $\mu$  by fitting a count model to the subsample of those eligible. The fact that  $Z$  is unobserved means that it is not possible to condition on the eligible group, which, from a policy standpoint, may be the subpopulation of greatest interest. Fortunately, zero-inflated models allow one to estimate  $\phi$  and  $\mu$  even when  $Z$  is

unobserved, a topic discussed in greater detail in section “[Model Fitting, Testing, and Evaluation.](#)”

The choice between ZIP and hurdle models is dictated in large part by the aims of the investigator. If zeros can arise in only one way, then a hurdle model may be desirable. For example, in a study of outpatient service use, it may happen that patients either decline services, in which case  $Y = 0$ , or they use services one or more times, in which case  $Y > 0$ . Here, a hurdle model might reasonably capture the underlying distribution of the counts. In contrast, if patients only use services when they perceive themselves to be “at risk,” then zeros can arise in two ways: among those who are not at risk or among those who are at risk but nevertheless choose not to use services. In this case, a zero-inflated model would seem more appropriate. In some situations, the choice between models is not clear-cut. In these circumstances, Min and Agresti (2005) suggest that hurdle models might provide better fit if there is evidence of zero deflation among subgroups of the population (e.g., among nonsmoking males). Zero-inflated models, on the other hand, imply zero inflation at all covariate values.

### Regression Models for Zero-Modified Count Data

Suppose interest lies in modeling the association between a set of predictors  $\mathbf{x}$  (e.g., age, race, etc.) and a zero-modified response  $Y$ . Hurdle models can be extended to the regression setting by modeling each component of as a function of  $\mathbf{x}$ :

$$\begin{aligned} g[\Pr(Y_i > 0)] &= g(\pi_i) \\ &= \mathbf{x}'_i \boldsymbol{\beta}_1 = \beta_{10} + \beta_{11}x_{1i} + \dots + \beta_{1p}x_{pi} \ln(\mu_i) \\ &= \mathbf{x}'_i \boldsymbol{\beta}_2 = \beta_{20} + \beta_{21}x_{1i} + \dots + \beta_{2p}x_{pi}, \\ i &= 1, \dots, n, \end{aligned} \tag{7}$$

where  $g(\cdot)$  is a binary link function, such as the logit or probit link,  $Y_i$  denotes the response for the  $i$ -th observation,  $\mathbf{x}_i$  is a  $p \times 1$  vector of predictors, and  $\boldsymbol{\beta}_1$  and  $\boldsymbol{\beta}_2$  are corresponding  $p \times 1$  vectors of regression coefficients for each component. Note that Eq. 7 models  $\pi = \Pr(Y > 0)$  rather

than  $1 - \pi = \Pr(Y = 0)$ , since the former is typically of interest. Moreover, for simplicity, identical predictors are assumed for both parts of the model. In general, one might allow for unique predictors for the two components if the goal is to obtain a parsimonious model by removing extraneous variables in one component or if there is a priori scientific reason to believe that the two components are associated with unique sets of predictors.

Choosing a logit link for  $g(\cdot)$  gives rise to the *logistic hurdle regression model*:

$$\begin{aligned} \text{logit}(\pi_i) &= \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}'_i \boldsymbol{\beta}_1 \\ \ln(\mu_i) &= \mathbf{x}'_i \boldsymbol{\beta}_2, \quad i = 1, \dots, n. \end{aligned} \tag{8}$$

Under model (8), the  $l$ -th regression coefficient,  $\beta_{1l}$  ( $1 \leq l \leq p$ ), represents the effect of a one-unit change in the  $l$ -th predictor,  $x_{li}$ , on the log odds of service utilization, adjusting for other predictors. The precise interpretation of  $\beta_{2l}$  is less straightforward, since, conditional on  $Y > 0$ , the counts are modeled via a truncated distribution rather than an ordinary count distribution. Generally speaking, however,  $\beta_{2l} > 0$  implies that the expected count among health services users increases as  $x_{li}$  increases.

Zero-inflated regression models have a similar form:

$$\begin{aligned} g[\Pr(Z_i = 1)] &= g(\phi_i) = \mathbf{x}'_i \boldsymbol{\beta}_1 \\ \ln(\mu_i) &= \mathbf{x}'_i \boldsymbol{\beta}_2, \quad i = 1, \dots, n, \end{aligned} \tag{9}$$

where  $Z_i$  is the eligibility indicator for the  $i$ -th subject as defined in the previous section. Note that the first equation of (9) models  $\phi_i$ , the *eligibility* probability for the  $i$ -th individual, rather than the *utilization* probability, which is represented by  $\pi_i = \phi_i[1 - p(0; \mu_i)]$ . If a logit link is assumed for  $g$ , then  $\beta_{1l}$  denotes the effect of a one-unit change in covariate  $l$  on the log odds of eligibility, while  $\beta_{2l}$  represents the effect of a one-unit change in predictor  $l$  on the log-mean count given eligibility. Or, put another way, for every one-unit change in predictor  $l$ , the mean

count among eligible patients is multiplied by a factor of  $\exp(\beta_{21})$ . The parameter  $\exp(\beta_{21})$  is commonly referred to as the *incidence rate ratio*, or IRR, for the eligible population. As Albert et al. (2011) and Preisser et al. (2012) note, it is more often of interest to make inferences about the entire population comprising both eligible and non-eligible individuals. Consider, for example, the simple case where the model includes a single dichotomous covariate,  $x_i$ , and a logit link is assumed for  $g$  in Eq. 9. In this case, the IRR representing the overall effect of  $x_i$  in the entire population is

$$\begin{aligned} \text{IRR} &= \frac{E(Y_i | x_i = 1)}{E(Y_i | x_i = 0)} \\ &= \exp(\beta_{21}) \left\{ \frac{\exp(\beta_{11}) [1 + \exp(\beta_{10})]}{1 + \exp(\beta_{10} + \beta_{11})} \right\} \end{aligned} \tag{10}$$

where  $\beta_{10}$  is the intercept and  $\beta_{11}$  is the coefficient for  $x_i$  in the first component and  $\beta_{21}$  is the coefficient for  $x_i$  in the second component. When  $\beta_{11} = 0$ , the population IRR is equal to the IRR for the eligible population,  $\exp(\beta_{21})$ . As  $\beta_{11}$  deviates from 0, naively interpreting the IRR for the eligible class as the overall IRR will lead to increasingly biased inferences. For a fuller discussion of this topic, including extensions to multiple categorical and continuous predictors, see Preisser et al. (2012).

A special case of the ZIP regression model is the ZIP( $\tau$ ) model (Lambert 1992) whereby  $\beta_1 = \tau\beta_2 (-\infty < \tau < \infty)$  in Eq. 9, implying that the covariate effects are proportional across model components. If a logit link is assumed for  $g$ , then  $\phi_i = (1 + \mu_i^{-\tau})^{-1}$ . As  $\tau \rightarrow -\infty$ , the probability of observing a zero for the  $i$ -th subject increases, and as  $\tau \rightarrow \infty$ , the probability of a zero decreases. In many applications, the more parsimonious ZIP( $\tau$ ) can lead to efficiency gains in parameter estimation compared to the ordinary ZIP model.

Heilbron (1994) developed a related *zero-altered regression model* that can be used to test for zero modification. The model assumes a single distribution for the  $i$ -th response,  $Y_i$ , but uses separate parameters for  $\Pr(Y_i = 0)$  and  $\Pr(Y_i = y_i | Y_i > 0)$ . For example, the *zero-altered Poisson (ZAP) model* takes the form

$$\begin{aligned} \Pr(Y_i = 0) &= e^{-\mu_{1i}}, \ln(\mu_{1i}) = \mathbf{x}'_i \boldsymbol{\beta}_1 \\ \Pr(Y_i = y_i | Y_i > 0) &= \frac{\mu_{2i}^{y_i} e^{-\mu_{2i}}}{y_i! (1 - e^{-\mu_{2i}})}, \ln(\mu_{2i}) = \mathbf{x}'_i \boldsymbol{\beta}_2. \end{aligned} \tag{11}$$

If one sets  $\mathbf{x}'_i \boldsymbol{\beta}_1 = \gamma + \mathbf{x}'_i \boldsymbol{\beta}_2$ , then testing for zero modification reduces to testing  $\gamma = 0$ . When  $\gamma < 0$ , the data are zero inflated; when  $\gamma > 0$ , the zeros are deflated; and when  $\gamma = 0$ , the model reduces to a standard Poisson model. In this case, model (11) can be fit using a complementary log-log link for  $\Pr(Y_i > 0) = v_i$ :

$$\begin{aligned} \text{cloglog}(v_i) &= \ln[-\ln(1 - v_i)] = \gamma + \mathbf{x}'_i \boldsymbol{\beta}_2 \\ \ln(\mu_i) &= \mathbf{x}'_i \boldsymbol{\beta}_2. \end{aligned} \tag{12}$$

For a general discussion of zero-altered models, including extensions of model (11), see Heilbron (1994).

Several authors have proposed zero-modified regression models for repeated measures and clustered count data. The most common approach is to incorporate random effects into the linear predictors for each part of the model. For example, Hall (2000) developed a repeated measure ZIP model that included a random intercept for the Poisson component. Yau and Lee (2001) later introduced uncorrelated random intercepts for both components of a hurdle model. Min and Agresti (2005) extended the approach to include correlated random intercepts for the two components. In particular, the *logistic hurdle regression model with correlated random intercepts* is given by

$$\begin{aligned} \text{logit}[\Pr(Y_{ij} > 0 | b_{1i})]t &= \text{logit}(\pi_{ij}) \\ &= \mathbf{x}'_{ij} \boldsymbol{\beta}_1 + b_{1i} \\ \ln(\mu_{ij}) &= \mathbf{x}'_{ij} \boldsymbol{\beta}_2 + b_{2i}, \\ j &= 1, \dots, n_i; \\ i &= 1, \dots, n; \\ \mathbf{b}_i &= \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \sim N_2(\mathbf{0}, \boldsymbol{\Sigma}), \end{aligned} \tag{13}$$

where  $Y_{ij}$  is the  $j$ -th response for subject (or cluster)  $i$ ,  $\mathbf{x}_{ij}$  is a corresponding vector of predictors for the  $ij$ -th observation,  $b_{1i}$  and  $b_{2i}$  are random intercepts for the  $i$ -th subject/cluster, and  $N_2(\mathbf{0}, \Sigma)$  denotes a bivariate normal distribution with mean  $\mathbf{0} = (0, 0)'$  and  $2 \times 2$  variance-covariance matrix  $\Sigma$ . Higher dimensional random effects, such as random slopes, can be incorporated as well. The correlated random effects model is appealing if one believes that the process giving rise to a nonzero count is related to the expected count given  $Y > 0$ . For example, returning to the ED study presented at the beginning of the section, it might be reasonable to hypothesize that patients with a high propensity to use the ED at least once are also likely to make repeat visits given some utilization. In such cases, the correlated random effects model can lead to improved model fit over uncorrelated random effects, single random effect, and fixed-effects models – all of which arise as special cases of the correlated model. The correlated zero-inflated model has a comparable form to the hurdle model, but as noted in the previous subsection, the interpretation of the parameters differs. For overviews of zero-modified count models for repeated measures, see Min and Agresti (2005) and Neelon et al. (2010). For a more general discussion of count regression models, including zero-modified models, see Cameron and Trivedi (1998), Winkelmann (2008), and Zuur et al. (2012).

## Recent Developments

Two-part count models have been adapted to cover a wide range of statistical applications, including latent growth curve models, finite mixture models, generalized additive models, variable selection methods, multivariate analysis, and spatial data analysis. For example, Liu (2007) developed a zero-inflated growth model that allows for correlated random intercepts and slopes for both components. Roeder et al. (1999), Dalrymple et al. (2003), and Min and Agresti (2005) developed finite mixture zero-modified models that cluster subjects into distinct classes defined by latent response trajectories. DeSantis and

Bandyopadhyay (2011) proposed a two-state, hidden Markov ZIP model to analyze cocaine dependence, with hypothesized latent states corresponding to “high” or “low” cocaine use. Dobbie and Welsh (2001) and Hall and Zhang (2004) used generalized estimating equations (GEE) to fit population-average (or “marginal”) Poisson hurdle models. Fahrmeir and Osuna Echavarría (2006) developed a generalized additive ZINB model, using penalized splines to model nonlinear trends among the predictors. Lam et al. (2006) proposed a related semi-parametric ZIP model. Hsu (2005) introduced a weighted ZIP (W-ZIP) model to predict the time to recurrence of colorectal polyps among patients randomized to high- and low-fiber diets. Buu et al. (2011) developed a variable selection method for ZIP models that allows for component-specific penalties. Williamson et al. (2007) derived power and sample size calculations for studies involving zero-inflated data. For times series analysis, Hasan and Sneddon (2009) developed first-order autoregressive (AR(1)) and moving average (MA(1)) ZIP models. More recently, Silva et al. (2011) proposed a ZIP model for quantitative trait loci (QTL) mapping.

Several authors have introduced zero-inflated models for the analysis of spatially correlated data. Agarwal et al. (2002) developed a spatial ZIP model that incorporated spatially correlated random effects into the Poisson component. Rathbun and Fei (2006) proposed a similar model in which the structural zeros were fitted using a spatial probit model. Ver Hoef and Jansen (2007) extended the approach to include distinct spatial random effects for both model components. Recently, Neelon et al. (2012) developed a spatial Poisson hurdle model for “areal-referenced” data in which the spatial units consist of aggregated regions of space, such as counties or Census tracts. They introduced spatial random effects for both components of the hurdle model and linked the random effects via a bivariate conditionally autoregressive (CAR) prior that induces dependence between the model components and provides spatial smoothing across neighboring regions. As such, their model can be viewed as a



spatial analogue to the correlated hurdle model given in Eq. 13.

There have been a number of other recent developments as well. These include zero-inflated binomial (ZIB) models (Hall 2000), Hall and Zhang 2004), pattern-mixture Poisson hurdle models for non-ignorable missing data (Hasan et al. 2009; Maruotti 2011), the k-ZIG model for extreme zero inflation (Ghosh et al. 2012), zero-inflated generalized Poisson (ZIGP) models (Gschlößl and Czado 2008; Gupta et al. 1996), zero-inflated power series models (Ghosh et al. 2006), and multivariate extensions of zero-inflated models (Li et al. 1999; Walhin and Bivariate 2001; Majumdar and Gries 2010; Arab et al. 2011). These recent applications highlight the growing use of two-part models for the analysis of complex zero-modified count data.

---

## Two-Part Models for Semicontinuous Data

In many cases, the nonzero response distribution is continuous rather than count valued. Such data are commonly referred to as “semicontinuous” because they consist of a mixture of a degenerate distribution at zero and a right-skewed, continuous distribution for the nonzero values. As an illustration, consider Fig. 2, which shows the distribution of annual mental health expenditures among federal employees from a recent study by Neelon et al. (2011). Over 80% of the patients had no annual expenditures, depicted by the vertical line, while the remaining patients spent upward of 1000 USD during the study period. Other examples of semicontinuous data include medical costs (Manning et al. 1981; Duan et al. 1983; Cooper et al. 2003), hospital length of stay (Xie et al. 2004), health assessment scores (Su et al. 2009), and average daily alcohol consumption (Olsen and Schafer 2001; Liu et al. 2012). In some cases, such as days of hospitalization or questionnaire scores, the response is, strictly speaking, integer valued, but the domain is refined enough to be reasonably approximated by a continuous distribution.

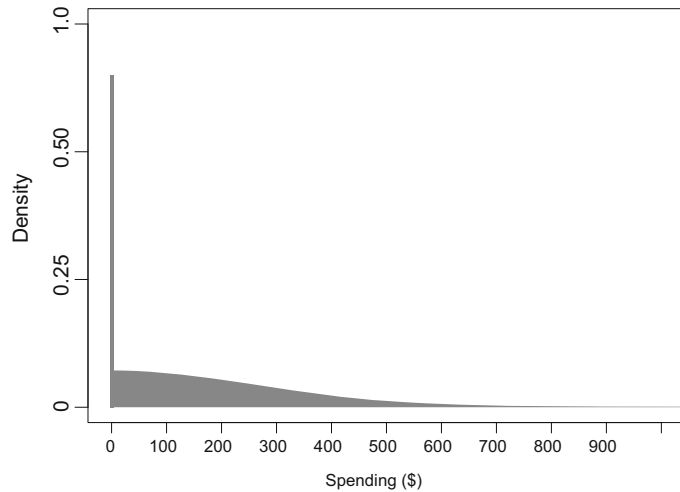
As with zero-modified count data, semicontinuous data can be viewed as arising from two distinct stochastic processes: one governing the occurrence of zeros and the second determining the observed value given a nonzero response. The first process is commonly referred to as the “occurrence” or “binary” part of the data, and the second is often termed the “intensity” or “continuous” part. Two-part mixture models are an ideal choice for such data, since they explicitly accommodate both data-generating processes. A lognormal distribution is frequently chosen to model the nonzero values, giving rise to the *Bernoulli-lognormal two-part model* (Manning et al. 1981):

$$f(y) = (1 - \phi)1_{(y=0)} + [\phi \times \text{LN}(y; \mu, \sigma^2)] 1_{(y>0)}, y \geq 0, 0 \leq \phi \leq 1, \quad (14)$$

where  $\phi = \Pr(Y > 0)$ ,  $\text{LN}(y; \mu, \sigma^2)$  denotes the lognormal density evaluated at  $y$  and  $\mu$  and  $\sigma^2$  denote the mean and variance of  $\ln(Y|Y > 0)$ . Note that Eq. 14 has the same two-part structure as the hurdle model given in Eq. 2 in the previous section and can therefore be viewed as a natural extension of the hurdle model to semicontinuous data. As with hurdle models, when  $\phi = 0$ , the distribution is degenerate at 0; when  $\phi = 1$ , there are no zeros and the distribution reduces to a lognormal density. Typically, one assumes that 0 is strictly between 0 and 1, so that all individuals have a long-run guarantee of a nonzero value. Note that even if 0 truly takes the value 0 for some subjects, model (14) cannot identify such individuals. That is, without further identifying assumptions, the model cannot differentiate the so-called never users from those who happened not to use services during the study period.

As in the count setting, alternative distributions can be used to model the positive values. For example, as part of a study examining inpatient medical expenditures, Manning et al. (2005) proposed a one-part generalized gamma distribution that encompasses the Weibull, exponential, and lognormal distributions as special cases. Building on this work, Liu et al. (2010) developed a two-part generalized gamma model for semicontinuous medical

**Fig. 2** Distribution of annual mental health expenditures among federal employees



costs. More recently, Liu et al. (2012) compared generalized gamma, log-skew-normal, and Box-Cox-transformed two-part models and found that the generalized gamma model provided superior fit in their analysis of daily alcohol consumption.

A related model is the *Tobit model* (Tobin 1958) in which the zeros represent the censoring of an underlying continuous variable  $Y^*$  below a detection limit,  $L$ :

$$\begin{aligned} Y &= 0 && \text{if } Y^* < L \\ Y &= Y^* && \text{if } Y^* \geq L. \end{aligned} \tag{15}$$

Censoring from above and interval censoring are also allowed. Note that in the Tobit model, the zeros arise from censoring of  $Y^*$  values that fall below  $L$ , whereas in two-part semicontinuous models, the zeros are valid observed responses (corresponding to, say, no medical expenditures). The Tobit and two-part models also differ in that the former assumes a single underlying distribution for the data, whereas the two-part model is a mixture of two separate generating processes – one for the zeros and one for the positive values. Recognizing these distinctions, Moulton and Halsey (1995) proposed a *zero-inflated Tobit (ZIT) model* that distinguished between censored zeros and “true” zeros. The model is a mixture of a point mass at zero and a Tobit model, such that with probability  $1 - p$ ,  $Y$  is set to 0, and otherwise  $Y$  is drawn from a Tobit distribution. Because the ZIT

model accommodates two sources of zeros (true zeros and censored zeros), it can be viewed a semicontinuous version of the zero-inflated count models described in section “[Zero-Inflated Count Models.](#)”

### Two-Part Regression Models for Semicontinuous Data

Two-part models for semicontinuous data can be extended to the regression setting by incorporating predictors into each component of the model. For example, the *Bernoulli-lognormal two-part regression model* is given by

$$\begin{aligned} f(y_i) &= (1 - \phi_i)1_{(y_i=0)} \\ &\quad + \phi_i \text{LN}(y_i : \mu_i, \sigma^2) 1_{(y_i>0)}, \quad \text{where} \\ g(\phi_i) &= g[\text{Pr}(Y_i > 0)] = \mathbf{x}'_i \boldsymbol{\beta}_1 \quad \text{and} \\ \mu_i &= \text{E}[\ln(Y_i) | Y_i > 0] = \mathbf{x}'_i \boldsymbol{\beta}_2, \quad i = 1, \dots, n. \end{aligned} \tag{16}$$

When a logit link is assumed for  $g(\cdot)$ , the  $l$ -th regression coefficient,  $\beta_{1l}$ , represents the change in the log odds of a positive response per one-unit change in covariate  $x_{il}$ , adjusting for other predictors. Likewise,  $\beta_{2l}$  represents the adjusted per unit change in the mean of  $\ln(Y_i) | Y_i > 0$ . Note that to convert from the log scale to the original response scale in part 2 of the model, one can

take  $\exp(\beta_{2i})$ , which denotes the multiplicative change in the median of  $Y_i|Y_i > 0$  per unit change in  $x_{i\cdot}$ . Because the expected value of  $Y_i|Y_i > 0$  is given by  $\exp(x'_{i\cdot}\beta_2 + \sigma^2/2)$ , inference involving the untransformed mean response entails estimation of both  $\beta$  and  $\sigma^2$ . If the log-normality assumption fails, nonparametric methods can be used to estimate the untransformed mean in the continuous component of the model (Duan 1983). This topic is discussed in greater detail in section “Model Fitting, Testing, and Evaluation.”

Two-part regression models have also been used to analyze longitudinal and clustered semi-continuous data (Olsen and Schafer 2001; Tooze et al. 2002; Cooper et al. 2007). The most common approach is to introduce correlated random effects for each component, as in model (13) of section “Regression Models for Zero-Modified Count Data.” Assuming a logit link for  $g(\cdot)$  and a lognormal distribution for the positive values leads to the logistic-lognormal correlated random effects model:

$$\begin{aligned} \text{logit}(\phi_{ij}) &= \text{logit} [\Pr(Y_{ij} > 0 | b_{1i})] = \mathbf{x}'_{ij}\beta_1 + b_{1i} \\ \mu_{ij} &= E [\ln(Y_{ij}) | Y_{ij} > 0, b_{2i}] = \mathbf{x}'_{ij}\beta_2 + b_{2i}, \\ j &= 1, \dots, n_i; \quad i = 1, \dots, n; \\ \mathbf{b}_i &= (b_{1i}, b_{2i}) \sim N_2(0, \Sigma), \end{aligned} \tag{17}$$

where  $Y_{ij}$  denotes the  $j$ -th response for the  $i$ -th subject (or cluster);  $b_{1i}$  and  $b_{2i}$  are correlated subject-specific random intercepts for the binary and continuous components, respectively; and  $\Sigma$  is a  $2 \times 2$  variance-covariance matrix. The model can be easily extended to include higher-order random effects.

As in the count setting, the correlated model is appealing if one believes that the process giving rise to the positive values is related to the observed value given a positive response. For example, in a study of hospital length of stay, patients who are likely to be admitted to the hospital may also tend to have longer stays than those with lower propensities for admission. This would imply a positive association between the probability of admission (component 1 of the model) and the length of stay given admission (component 2).

Modeling the correlation between  $b_{1i}$  and  $b_{2i}$  directly accommodates the between-component association, thus providing a realistic characterization of the underlying data-generating process.

There are other advantages to modeling the between-component association, however. Most importantly, ignoring the between-component association can lead to biased estimates in the second part of the model (Su et al. 2009). To see this, consider the two-part lognormal model given in Eq. 17, which can be recoded in terms of two random variables:

$$\begin{aligned} R &= \begin{cases} 0 & \text{if } Y = 0 \\ 1 & \text{if } Y > 0 \end{cases} \\ V &= \begin{cases} \text{Undefined} & \text{if } R = 0 \\ \log(Y) & \text{if } R = 1 \end{cases}, \end{aligned} \tag{18}$$

where subscripts have been omitted to simplify notation. The random variable  $R = 1_{(Y>0)}$  can be viewed as a response indicator for the second component of the model.

Recall that the target population for the continuous part is the set of all subjects with positive responses – that is, for whom  $Y > 0$  (or equivalently for whom  $R = 1$ ). Valid inferences can be achieved by selecting a random sample  $\mathcal{V} = \{V_1, V_2, \dots, V_n\}$  from this target population. However, when some individuals have a response value of 0, a subset of  $\mathcal{V}$  (say,  $\mathcal{V}^*$ ) is undefined. These undefined observations can be viewed as akin to missing data. If the two components are truly uncorrelated, then  $\mathcal{V}^*$  is missing completely at random (MCAR). In this case, the model for  $R$  includes only an intercept and therefore has no bearing on the model for  $V$ . Consequently, a model fitted to the observed values of  $\mathcal{V}$  will yield population-representative estimates.

If the association between the components can be explained entirely by observed data, then the elements of  $\mathcal{V}^*$  are missing at random (MAR). In other words,  $R$  and  $\mathcal{V}^*$  are conditionally independent given the observed data. Modeling  $R$  and  $V$  separately will once again yield unbiased estimates as long as the model for  $V$  is correctly specified and includes all predictors relevant to  $R$ . In some instances, investigators may wish to

include only a subset of the predictors in part 2 of the model, in which case the model for  $V$  will, by necessity, exclude predictors associated with  $R$ . Here, one can use the model for  $R$  to form sampling weights and fit a weighted regression to  $V$ . Alternatively, one can impute  $\mathcal{V}^*$  and base the analysis on the observed and imputed data. The key point is that if  $\mathcal{V}^*$  is MAR, then modeling  $R$  and  $V$  separately will not induce bias so long as the model for  $V$  is correct and incorporates, in some fashion, the relevant predictors for  $R$ .

If, however,  $R$  and  $\mathcal{V}^*$  remain correlated after adjusting for covariates, then  $\mathcal{V}^*$  is *not missing at random* (NMAR). Here, fitting separate models for  $R$  and  $V$  induces *selection bias* in the parameter estimates for  $V$ . For example, if the two components are positively correlated, higher-valued  $V$ 's will tend to have increased nonzero response probability,  $\Pr(R = 1)$ , conditional on observed covariates. As a result, at fixed values of the observed covariates, there will be an overrepresentation of large response values among the observed cases in  $\mathcal{V}$ . Ignoring the association between the two components and basing the part-2 analysis solely on observed cases will bias the fixed-effects intercept upward and may lead to bias in other part-2 parameters as well, depending on the structure of the between-component association (Su et al. 2009). One way to correct for this bias is to fit a correlated two-part model analogous to Eq. 17. The resulting model can be viewed as a *shared-parameter model* (Wu and Carroll 1988) that accounts for unmeasured subject-level factors that induce correlation between  $R$  and  $\mathcal{V}^*$ . Note that this approach again relies on a conditional independence assumption whereby, this time,  $R$  and  $\mathcal{V}^*$  are assumed to be stochastically independent given both the observed data *and* the random effects. While it is impossible to verify whether  $\mathcal{V}^*$  is MAR or NMAR, it is often safer to assume NMAR, unless enough covariates have been measured to reasonably account for the dependence between  $R$  and  $\mathcal{V}^*$ .

For further details on selection bias in two-part semicontinuous models, see Su et al. (2009). For a related discussion regarding selection bias in hurdle count models, see Neelon et al. (2012). For further discussion of missing data

mechanisms, see Little and Rubin (2002). For a general overview of shared-parameter models for non-ignorable missing data, see Albert and Follmann (2009).

## Recent Developments

There have been a number of recent developments in semicontinuous regression modeling. Liu et al. (2008) developed a *multilevel two-part model* that incorporates correlated random effects at multiple levels of clustering – for example, longitudinal measurements on patients clustered within clinic. Here, clinics constitute the first clustering level, since patients are nested within clinics; patients then form the second level of clustering, since there are repeated measurements for each subject.

Another active area of research involves *two-part growth mixture models* for examining longitudinal trends among latent subgroups of individuals (Neelon et al. 2011; Muthén 2001). Growth mixture models assume that the data are generated through a two-step process: first, individuals are placed into one of  $K$  latent classes defined by a set of average trajectory curves – one for each component of the two-part model; then, around these average trajectories, individuals are randomly assigned their own, subject-specific curves defined by a set of random effects with class-specific variance parameters. As such, these models can be viewed as finite mixtures of the two-part correlated random effects model expressed in Eq. 17.

Other recent developments include bivariate two-part models (Su et al. 2012), two-part models for the joint analysis of longitudinal and survival outcomes (Liu 2009); Hatfield et al. 2011), two-part models for estimating expected cumulative cost of illness in the presence of censoring (Basu and Manning 2010), and Bayesian extensions of two-part semicontinuous models (Neelon et al. 2011; Liu 2009; Hatfield et al. 2011; Zhang et al. 2006). This recent work highlights a growing interest in parametric two-part modeling and solidifies its current role as a vibrant area of statistical research.

### Model Fitting, Testing, and Evaluation

There is a broad spectrum of estimation, testing, and model evaluation techniques suitable for statistical inference in two-part models, and the choice of method depends on both the type of model being fit and the analytic aims of the investigators. This section highlights common approaches to parameter estimation in two-part models, with the aim being to provide an informal overview of these techniques. Readers are encouraged to seek out the cited references for more technical discussions of these methods.

### Zero-Modified Count Models

There are a number of approaches to parameter estimation in zero-modified count models, including maximum likelihood (ML), generalized estimating equations (GEE), penalized quasi-likelihood, estimation-maximization (EM) algorithms, and Bayesian methods. For the uncorrelated hurdle models given in Eqs. 2, 7, and 8, parameter estimation proceeds by fitting the two model components separately. For example, for the logistic-Poisson hurdle model (8),  $\beta_1$  is estimated by fitting a logistic regression to  $\pi_i = \Pr(Y_i > 0)$ , while  $\beta_2$  is estimated by fitting a truncated Poisson model for  $Y_i | Y_i > 0$ . Newton-Raphson or Fisher scoring algorithms are typically used for ML estimation. For large samples, asymptotically approximate confidence intervals can be obtained using well-established normal theory results. Predicted values for a future response,  $y_i^*$ , can also be generated as functions of the regression estimates:

$$\hat{y}_i^* = \frac{\hat{\pi}_i \hat{\mu}_i}{1 - e^{-\hat{\mu}_i}}, \quad \text{where } \hat{\pi}_i = \frac{e^{x_i' \hat{\beta}_1}}{1 + e^{x_i' \hat{\beta}_1}} \quad (19)$$

and  $\hat{\mu}_i = e^{x_i' \hat{\beta}_2}$ .

Bootstrapping or large-sample Taylor series approximations can then be used to obtain confidence intervals for the predicted values.

For zero-inflated models, recall that the latent “eligibility” indicator  $Z$  – and hence the eligibility

probability  $\phi$  – is unobserved. Consequently, ML estimation is commonly implemented using the EM algorithm. Under this approach, the latent indicator  $Z$  is treated as missing data. The expectation step involves computing the expected value (with respect to  $Z$ ) of the logged “complete data” likelihood as expressed, for example, in the last line of Eq. (6). In the maximization step, the expected complete data log-likelihood is maximized with respect to the model parameters. Alternatively, since the zero-inflated model can be reparameterized as a hurdle model (Neelon et al. 2010), Newton-Raphson algorithms can be used to obtain the ML estimates. Of the two choices, Lambert (1992) found that the EM algorithm generally outperformed Newton-Raphson. However, for the ZIP( $\tau$ ) regression model, Lambert (1992) notes that the EM algorithm is not useful, since the parameters  $\beta$  and  $\tau$  are not easily estimated even when  $Z$  is observed. She recommends Newton-Raphson procedures in this case. A more technical discussion of these procedures can be found in her paper.

Recall from section “Regression Models for Zero-Modified Count Data” that the zero-altered model can be used to test for the presence of zero modification. ML estimation for zero-altered models proceeds via Newton-Raphson or related Fisher scoring methods. If there is evidence of zero inflation, the investigator may subsequently elect to fit a zero-inflated model. In this case, Vuong’s test (Vuong 1989) or the score test developed by Ridout et al. (2001) can be used to choose between ZIP and ZINB models. Xiang et al. (2007) recently extended the testing procedure to account for repeated measures.

There is a wide range of model fitting strategies for clustered and longitudinal data, including mixed models and GEE. GEE (Liang and Zeger 1986) is a quasi-likelihood approach in which regression estimates are first obtained from score-type estimating equations that include a “working covariance” matrix to account for within-cluster association. Next, asymptotic standard errors that are robust to possible mis-specifications of the working covariance structure are derived. For hurdle models, GEE estimation proceeds by separately estimating the parameters for the binary

and truncated count components. Dobbie and Welsh (2001) used GEE to fit a Poisson hurdle model to clustered count data. Hall and Zhang (2004) extended the approach to zero-inflated models by combining GEE with an EM-type expectation step, resulting in a two-step “expectation-solution” (ES) procedure (Rosen et al. 2000). In the E-step, the expectation of the complete data log-likelihood with respect to the latent indicator  $Z$  is computed; in the S-step, GEE is used in lieu of maximum likelihood to obtain parameter estimates and robust standard errors separately for each component of the model.

For the zero-modified random effects models described in Eq. 13, Min and Agresti (2005) proposed a two-stage approach in which numerical integration, such as Gaussian quadrature, is first used to estimate the marginal likelihood integrated across the random effects; then, in the second stage, Fisher scoring is used to maximize the estimated marginal likelihood. More recently, Kim et al. (2012) used restricted maximum quasi-likelihood (RMQL) to fit a correlated negative binomial hurdle model.

Several authors have used the EM algorithm for fitting longitudinal finite mixture (or “latent class”) models. Roeder et al. (1999) used EM to fit a latent class trajectory model as part of a study examining risk factors for long-term criminal behavior. Dalrymple et al. (2003) adopted a similar approach to study longitudinal trends in sudden infant death syndrome, or SIDS. Min and Agresti (2005) used the EM algorithm to fit a discrete random effects model in an analysis of pharmaceutical side effects.

Bayesian methods are also well suited for inference involving zero-modified count data. In Bayesian inference, model parameters are treated as random variables and assigned prior distributions that quantify one’s uncertainty about their values prior to observing the data. Common prior distributions for regression models include normal distributions for fixed-effect parameters, inverse-gamma distributions for error variances, and inverse-Wishart distributions for random effect covariance matrices. These prior distributions are then combined with the current data via Bayes’ theorem to obtain posterior distributions. In this

way, Bayesian methodology provides a natural scheme for learning from prior experience. For zero-modified count models, the posterior distributions generally do not have closed forms, and hence Markov chain Monte Carlo (MCMC) algorithms, such as Gibbs sampling (Gelfand and Smith 1990), are often used for posterior inference. At convergence, the MCMC draws form a Monte Carlo sample from the joint posterior distribution of all model parameters, which can then be used to obtain parameter estimates and corresponding interval estimates (credible intervals), thus avoiding the need for asymptotic assumptions. Moreover, because MCMC produces draws from the entire joint posterior distribution of the model parameters, estimation of complex functions of parameters is straightforward. For example, the Bayesian framework is ideal for estimating and obtaining uncertainty intervals for quantities such as the population IRR given in Eq. 10. In the maximum likelihood setting, one would have to perform bootstrapping or derive a Taylor series approximation to obtain standard errors and confidence intervals for such quantities.

In recent years, there has been growing interest in Bayesian methods for fitting zero-modified models. Rodrigues (2003) proposed a data-augmented Gibbs sampling algorithm to fit a ZIP model. Ghosh et al. (2006) used a similar approach to fit zero-inflated generalized power series models, which include the ZIP as a special case. Neelon et al. (2010) developed Bayesian model fitting strategies for repeated measures hurdle, ZIP, and ZAP models and compared various prior specifications, model comparison strategies, and approaches to assessing model fit. Ghosh et al. (2012) used Gibbs sampling to fit the  $k$ -ZIG model, which accommodates extreme zero inflation. Several authors have proposed Bayesian methods for analyzing zero-modified spatial data (Neelon et al. 2012; Rathbun and Fei 2006; Ver Hoef and Jansen 2007). In particular, Neelon et al. (2012) used hybrid Gibbs and Metropolis-Hastings steps to fit a spatially correlated Poisson hurdle model. For more on Bayesian estimation of zero-inflated count models, see Winkelmann (2008), Neelon et al. (2010, 2012), and Zuur et al. (2012).

## Semicontinuous Models

Similar procedures can be used for parameter estimation in two-part semicontinuous models. For example, Duan et al. (1983) used maximum likelihood to estimate a fixed-effects probit-log-normal model. The two components were estimated separately by fitting a probit regression model for the binary part and a lognormal regression model for the continuous part. Maximum likelihood can also be used to fit the Tobit and ZIT models. For the ZIT, Moulton and Hasley (1995) first restructured the model as a hurdle-type model, with the first component comprising both “true” and censored zeros and the second component consisting of positive responses that were assumed to follow a truncated lognormal distribution. They then used a quasi-Newton-Raphson procedure to obtain ML estimates. The EM algorithm could also be applied in this context, since the true zeros in the ZIT model are comparable to the structural zeros in zero-inflated count models.

For clustered semicontinuous data, one can use GEE to fit population-average two-part models. The two components can be estimated separately by fitting one GEE-estimated model for the binary part and another for the continuous part. For two-part mixed models with uncorrelated random effects, ML estimates can be derived by fitting separate random effects models for each component. For the correlated two-part model given in Eq. 17, Olsen and Schafer (2001) used a sixth-order Laplace approximation together with an approximate Fisher scoring algorithm to derive parameter estimates. Tooze et al. (2002) adopted a slightly different approach, using adaptive Gaussian quadrature to first approximate the marginal likelihood and then applying quasi-Newton-Raphson to obtain parameter estimates.

Several authors have proposed Bayesian approaches for fitting two-part semicontinuous models (Neelon et al. 2011; Cooper et al. 2003, 2007; Deb et al. 2006; Ghosh and Albert 2009). For example, Neelon et al. (2011) used a data-augmented MCMC algorithm to fit a probit-lognormal correlated two-part model. For more on Bayesian inference in two-part semicontinuous

models, see Cooper et al. (2003) and Neelon et al. (2011).

There is an extensive literature regarding estimation of the untransformed mean response in part 2 of semicontinuous models, a quantity of primary interest in many studies (Duan 1983; Manning 1998; Manning and Mullahy 2001). Recall that in the two-part lognormal regression model (16),  $\exp(\mathbf{x}'_i\boldsymbol{\beta}_2)$  represents the median response (given  $\mathbf{x}_i$ ) among the positive observations. The untransformed mean response, meanwhile, is given by

$$\begin{aligned}\psi(\mathbf{x}_i) &= E(Y_i | Y_i > 0, \mathbf{x}_i) \\ &= \exp(\mathbf{x}'_i\boldsymbol{\beta}_2 + \sigma^2/2),\end{aligned}\quad (20)$$

where  $\sigma^2$  is the lognormal variance. Thus, if interest lies in estimating  $\psi(\mathbf{x}_i)$ , it is necessary to estimate  $\sigma^2$ . A consistent estimator of  $\psi(\mathbf{x}_i)$  is given by  $\hat{\psi}(\mathbf{x}_i) = \exp(\mathbf{x}'_i\hat{\boldsymbol{\beta}}_2 + \hat{\sigma}^2/2)$ , where  $\hat{\boldsymbol{\beta}}_2$  is the ordinary least squares (OLS) estimate of  $\boldsymbol{\beta}_2$  and  $\hat{\sigma}^2$  is the mean squared error obtained from regressing the log-transformed response on  $\mathbf{x}$ . However, if the log-normality assumption is violated – for example, if the data arise from a mixture of lognormal distributions – then  $\hat{\psi}(\mathbf{x}_i)$  is not a consistent estimator for  $\psi(\mathbf{x}_i)$ . To accommodate departures from log-normality, Duan (1983) developed a consistent nonparametric estimator known as the *smearing estimator*, which is expressed as

$$\begin{aligned}\hat{E}(Y_0 | Y_0 > 0, \mathbf{X} = \mathbf{x}_0) &= \exp(\mathbf{x}'_0\hat{\boldsymbol{\beta}}_2) \frac{1}{n_+} \sum_{i: Y_i > 0} \exp(\hat{\varepsilon}_i), \\ &= \exp(\mathbf{x}'_0\hat{\boldsymbol{\beta}}_2) \hat{S},\end{aligned}\quad (21)$$

where  $Y_0$  denotes the untransformed response for an individual with covariate profile,  $\mathbf{x}_0$ ,  $n_+$  is the number of nonzero observations, and  $\hat{\varepsilon}_i = \ln(Y_i) - \mathbf{x}'_i\hat{\boldsymbol{\beta}}_2$  is the residual for the  $i$ -th nonzero observation. The expression  $\hat{S}$  is known as the “smearing factor.” The method generalizes to any monotone differentiable function  $g(Y)$ . Because the smearing estimator is nonparametric, it makes no explicit assumption about the distributional form

of  $Y$ , only that  $E[g(Y_i) | Y_i > 0, \mathbf{x}_i]$  is a linear function of  $\beta_2$  and that the errors are independent and identically distributed with mean zero with homogeneous variance  $\sigma^2$ . When the errors are heteroscedastic – for example, when they depend on covariates – the smearing estimator is biased (Manning 1998). Three approaches have been proposed to account for heteroscedasticity when constructing a smearing estimator: (1) estimate unique smearing factors for different covariate subgroups (Manning 1998), (2) apply separate smearing factors to different parts of the response distribution (Buntin and Zaslavsky 2004) or (3) use  $\hat{S} = E[\exp(\hat{e}_i) | Y_i > 0, \mathbf{x}_i]$  as a corrected smearing factor (Jones 2011) which can be obtained by regressing the exponentiated estimated residuals on  $\mathbf{x}$  and using the predicted values as the smearing factors at the corresponding values of  $\mathbf{x}$ . Recently, Welsh and Zhou (2006) developed a heteroscedastic smearing estimator for the untransformed *marginal* mean,  $E(Y_i | \mathbf{x}_i)$ , averaged over both the zero and nonzero observations.

Note that retransformations to the  $Y$ -scale pose no difficulty for Bayesian inference: after drawing MCMC samples of model parameters on the transformed scale, simply retransform and take the average to estimate the posterior mean on the original data scale. However, unless advanced Bayesian nonparametric techniques are employed (Ferguson 1973), an explicit parametric form for the likelihood must be assumed.

Quasi-likelihood generalized linear models (GLMs) offer an alternative approach to estimating the untransformed mean in part 2 of the model (Manning and Mullahy 2001; Buntin and Zaslavsky 2004; Blough et al. 1999). Here, the untransformed mean is modeled as  $\psi(\mathbf{x}_i) = h(\mathbf{x}_i; \beta_2)$ , where  $h$  is an inverse-link function (e.g., the exponential function). By modeling  $\psi(\mathbf{x}_i)$  directly, GLMs avoid the need to transform  $Y$  altogether. Next, the variance of  $Y(Y > 0)$  is modeled as a function of covariates, typically using a power function of the form  $V(Y_i | Y_i > 0, \mathbf{x}_i) \propto \psi(\mathbf{x}_i)^\lambda$ . The approach does not specify a distribution for  $Y$ , making it robust to mis-specifications that might otherwise occur. The method can also be used to directly estimate the marginal mean,  $E$

$(Y_i | \mathbf{x}_i)$ , yielding a simpler, one-part GLM that incorporates both zero and nonzero values.

Estimation for GLMs proceeds by nonlinear weighted least squares, with weights proportional to the inverse variances of the observations (Buntin and Zaslavsky 2004). The choice of  $\lambda$  is important, since it can affect the efficiency of the parameter estimates. Choosing  $\lambda = 0$  implies constant variance;  $\lambda = 1$  implies a ‘‘Poisson-type’’ variance proportional to the mean; and  $\lambda = 2$  results in a ‘‘gamma-type’’ variance. To help guide this choice, one can apply the Park test (Park 1966) which exploits the fact that

$$\ln[V(Y_i | Y_i > 0, \mathbf{x}_i)] = \text{constant} + \lambda \ln[\psi(\mathbf{x}_i)]. \quad (22)$$

To apply the Park test, the squared residuals from a candidate model are regressed on the log-transformed predicted values,  $\hat{y}_i$ :

$$\ln \left[ (y_i - \hat{y}_i)^2 | y_i > 0 \right] = \alpha + \lambda \ln(\hat{y}_i) + e_i, \quad i = 1, \dots, n, \quad (23)$$

where  $e_i$  is a mean-zero error term. An estimate of  $\lambda$  close to zero suggests constant variance, an estimate close to 1 suggests a Poisson-like variance, and an estimate close to 2 suggests a gamma-type structure.

In deciding between a GLM and a transformed parametric model, one can employ the following decision procedure, adapted from Manning and Mullahy (2001):

1. Fit an OLS regression to the transformed positive values.
2. If the residuals are highly kurtotic, then the parametric two-part model is generally preferable, since high kurtosis can lead to imprecision (high variability) in quasi-likelihood GLM parameter estimates. To guard against model mis-specifications, smearing should be applied when estimating the untransformed mean. In the presence of heteroscedasticity, multiple covariate- or response-dependent smearing factors should be applied to reduce bias.



3. If there is minimal kurtosis, fit a series of quasi-likelihood GLMs and apply the Park test to determine the optimal value of  $\lambda$ .
4. To avoid over-fitting, use penalized model comparison or cross validation techniques, such as split-sample analyses, to choose between competing models.

The choice between models is guided by non-statistical considerations as well. For example, if there is interest in estimating both the probability of a positive response and the mean response among positive observations, then a two-part model (either parametric or quasi-likelihood GLM) may be preferable to a one-part model. Further, if it is reasonable to assume that the two components are correlated, then a correlated parametric two-part model, as in Eq. 17, might be most appropriate. For a more detailed comparison of quasi-likelihood GLMs and transformed parametric models, see Manning and Mullahy (2001) and Buntin and Zaslavsky (2004).

### Model Comparison and Assessment

There are several model comparison measures that can be used to select among competing two-part models, including the *Akaike information criterion* (AIC) (Akaike 1974) and the *Bayesian information criterion* (BIC), also known as the *Schwarz criterion* (Schwarz 1978). AIC and BIC are referred to as “penalized” criteria because they combine a measure of model fit, typically twice the negative log-likelihood, with a penalty for model complexity, expressed as a function of the number of parameters. Smaller values of AIC and BIC are considered preferable. A related measure for quasi-likelihood and GEE models is the *quasi-likelihood under independence*, or QIC, criterion (Pan 2001). In the Bayesian setting, a common model comparison statistic is the *deviance information criterion* (DIC) (Spiegelhalter et al. 2002) which can be used to compare Bayesian hierarchical (i.e., random effect) models. As with the other selection criteria, DIC balances an assessment of model fit with a penalty for complexity. For random effects models, the dimension of the

parameter space depends on the degree of heterogeneity between individuals, with greater heterogeneity implying more “effective” parameters. DIC was specifically designed to estimate the number of effective parameters in Bayesian hierarchical models. Celeux et al. (2006) recently adapted the measure to accommodate additional latent variable models, such as finite mixtures.

A second Bayesian comparison measure is the *Bayes factor* (Kass and Raftery 1995) which offers perhaps the most principled approach to Bayesian model selection. However, because Bayes factors rely on the marginal likelihood of the data under a presumed model, they are not defined for improper (infinite variance) prior distributions. To accommodate improper priors, alternative criteria such as the *intrinsic Bayes factor* (Berger and Pericchi 1996) have been proposed. The *pseudo Bayes factor* (Gelfand and Dey 1994) offers a computationally convenient numerical approximation to the Bayes factor, but it has been criticized recently due to its reliance on the computationally unstable harmonic mean (Raftery et al. 2007). Several other Bayesian comparison measures have been proposed specifically in connection with zero-inflated count models, including the *group-marginalized DIC* (Millar (2009) and the *predictive log-score loss function* (Ghosh et al. 2012).

To further assess model fit in the Bayesian setting, one can apply *Bayesian posterior predictive checks*, whereby the observed data are compared to data replicated from the posterior predictive distribution (Gelman et al. (1996). If the model fits well, the replicated data should resemble the observed data. To quantify the degree of similarity, one typically chooses a “discrepancy statistic,” such as a sample moment or quantile, which captures some important aspect of the data. The *Bayesian predictive p-value* denotes the probability that the model-predicted statistic is more extreme than the observed sample value (i.e., the value expected under the correct model). A Bayesian p-value close to 0.50 represents adequate model fit, while p-values near 0 or 1 indicate lack of fit.

For more information on Bayesian model comparison and assessment strategies, see Millar (2009), Neelon et al. (2010, 2011), and Ando (2010).

---

## Software

There are a number of software packages that can be used for fitting zero-modified count and semicontinuous models. The statistical software program R (R Development Core Team 2012) has several packages for fitting zero-modified count models, including the `pscl` (Zeileis et al. 2008; Jackman 2012) package, which performs ML estimation of zero-inflated and hurdle models; `glmmADMB` (Fournier et al. (2012); Skaug et al. (2012)) for fitting random effect zero-inflated and hurdle models; and `MCMCglmm` (Hadfield 2010) for Bayesian estimation of hurdle, zero-inflated, and zero-altered models. SAS 9.1.3 Help and Documentation (2000) offers PROC COUNTREG for fitting zero-inflated count regressions, PROC GENMOD for GEE models, and PROCs NLMIXED and GLIMMIX for random effect zero-modified count models. Stata Statistical Software (2011) uses the `zip` and `zinb` commands for fitting ZIP and ZINB models, `HPLOGIT` and `HNBLOGIT` for hurdle models (Hilbe 2005a, b), and `gllamm` for fitting random effect ZIP models (Rabe-Hesketh et al. 2005). For Bayesian inference, the freeware package WinBUGS (Lunn et al. 2000) can be used to fit various zero-modified count models, including hierarchical models. See Neelon et al. (2010, 2012) for examples.

Many of these packages can also be used to fit semicontinuous models. For example, SAS PROC NLMIXED can be used to fit random effect semicontinuous models (Tooze et al. 2002). The freeware package ML (Lillard and Panis 1998) can be used to fit multilevel two-part models; see Liu et al. (2008) for an application. Mplus software (Muthén and Muthén 1998) is useful for fitting finite mixture and growth mixture two-part models; Muthén (2001) provides an illustration. SAS PROC GENMOD and the Stata command `glm` can be used to fit quasi-likelihood one- and two-part models. Buntin and Zaslavsky

(2004) provide example code for fitting such models. Finally, WinBUGS can be used to fit Bayesian two-part semicontinuous models; see Cooper et al. (2003, 2007) Cooper et al. (2007) and Ghosh and Albert (2009) for examples. Readers should visit the appropriate software websites for updates and current versions of these packages.

---

## Conclusion

Two-part models play an important role in health services research settings where data are characterized by both a high proportion of zeros and a skewed distribution of positive values. By modeling the zero and nonzero values in distinct ways, two-part models offer a flexible parametric approach to the analysis of zero-modified count and semicontinuous data. In many cases, such flexibility can yield improved model fit over traditional one-part models. At the same time, the reliance on parametric assumptions can be a liability, particularly in the case of semicontinuous data. Misguided assumptions about the response distribution will naturally lead to biased inferences. As in any regression analysis, careful attention to modeling assumptions is paramount to achieving unbiased parameter estimates. If these assumptions appear to be violated, distribution-free quasi-likelihood or other semi-parametric approaches may be preferable.

There are a number of active areas of research involving two-part models. These include two-part spatial and spatiotemporal models for semicontinuous data, shared-parameter models for informatively censored zero-modified counts, and inverse-probability weighting methods for population-average two-part models. These developments highlight just a few of the potential opportunities for methodological research involving two-part models.

Lastly, given the scope of the methods described above, this chapter should be viewed as an introductory overview of two-part modeling. Readers are encouraged to consult the references cited herein for further discussions of two-part models and their ongoing application to health services research.

## References

- Agarwal DK, Gelfand AE, Citron-Pousty S. Zero-inflated models with application to spatial count data. *Environ Ecol Stat.* 2002;9(4):341–55. Available from <http://www.ingentaconnect.com/content/klu/eest/2002/00000009/00000004/05102063>
- Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control.* 1974;19(6):716–23.
- Albert P, Follman D. Shared-parameter models. In: Fitzmaurice G, Davidian M, Ver-beke G, Molenberghs G, editors. *Longitudinal data analysis*. Boca Raton: Chapman & Hall/CRC Press; 2009. p. 433–52.
- Albert JM, Wang W, Nelson S. Estimating overall exposure effects for zero-inflated regression models with application to dental caries. *Stat Methods Med Res.* 2011. Available from <http://smm.sagepub.com/content/early/2011/09/08/0962280211407800.abstract>
- Ando T. *Bayesian model selection and statistical modeling*. Boca Raton: Chapman Hall/CRC Press; 2010.
- Arab A, Holan SH, Wikle CK, Wildhaber ML. Semiparametric bivariate zero-inflated Poisson models with application to studies of abundance for multiple species. *ArXiv e-prints.* 2011. Available from <http://arxiv.org/abs/1105.3169v1>
- Basu A, Manning WG. Estimating lifetime or episode-of-illness costs under censoring. *Health Econ.* 2010;19(9):1010–28. <https://doi.org/10.1002/hec.1640>.
- Berger JO, Pericchi LR. The intrinsic Bayes factor for model selection and prediction. *J Am Stat Assoc.* 1996;91(433):109–22. Available from <http://www.jstor.org/stable/2291387>
- Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. *J Health Econ.* 1999;18(2):153–71. Available from <http://www.sciencedirect.com/science/article/pii/S0167629698000320>
- Buntin MB, Zaslavsky AM. Too much ado about two-part models and transformation?: comparing methods of modeling Medicare expenditures. *J Health Econ.* 2004;23(3):525–42. Available from <http://www.sciencedirect.com/science/article/pii/S0167629604000220>
- Buu A, Johnson NJ, Li R, Tan X. New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Stat Med.* 2011;30(18):2326–40. <https://doi.org/10.1002/sim.4268>.
- Cameron AC, Trivedi PK. *Regression analysis of count data*. No. 9780521635677 in Cambridge Books. Cambridge University Press; 1998. Available from <http://ideas.repec.org/b/cup/cbooks/9780521635677.html>
- Celeux G, Forbes F, Robert CP, Titterton DM. Deviance information criteria for missing data models. *Bayesian Anal.* 2006;1(4):651–74.
- Consul P. *Generalized Poisson distributions: properties and applications*. New York: Marcel Dekker; 1989.
- Cooper NJ, Sutton AJ, Mugford M, Abrams KR. Use of Bayesian Markov chain Monte Carlo methods to model cost-of-illness data. *Med Decis Mak.* 2003;23(1):38–53. Available from <http://mdm.sagepub.com/content/23/1/38.abstract>
- Cooper NJ, Lambert PC, Abrams KR, Sutton AJ. Predicting costs over time using Bayesian Markov chain Monte Carlo methods: an application to early inflammatory polyarthritis. *Health Econ.* 2007;16(1):37–56. <https://doi.org/10.1002/hec.1141>.
- Dalrymple ML, Hudson IL, Ford RPK. Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. *Comput Stat Data Anal.* 2003;41(3–4):491–504. [https://doi.org/10.1016/S0167-9473\(02\)00187-1](https://doi.org/10.1016/S0167-9473(02)00187-1).
- Deb P, Munkin MK, Trivedi PK. Bayesian analysis of the two-part model with endogeneity: application to health care expenditure. *J Appl Econ.* 2006;21(7):1081–99. <https://doi.org/10.1002/jae.891>.
- DeSantis SM, Bandyopadhyay D. Hidden Markov models for zero-inflated Poisson counts with an application to substance use. *Stat Med.* 2011;30(14):1678–94. <https://doi.org/10.1002/sim.4207>.
- Dobbie MJ, Welsh AH. Modelling correlated zero-inflated count data. *Aust N Z J Stat.* 2001;43(4):431–44. <https://doi.org/10.1111/1467-842X.00191>.
- Duan N. Smearing estimate: a nonparametric retransformation method. *J Am Stat Assoc.* 1983;78(383):605–10. Available from <http://www.jstor.org/stable/2288126>
- Duan N, Manning J Willard G, Morris CN, Newhouse JP. A comparison of alternative models for the demand for medical care. *J Bus Econ Stat.* 1983;1(2):115–26. Available from <http://www.jstor.org/stable/1391852>
- Fahrmeir L, Osuna EL. Structured additive regression for overdispersed and zero-inflated count data. *Appl Stoch Model Bus Ind.* 2006;22(4):351–69. <https://doi.org/10.1002/asmb.631>.
- Ferguson TS. A bayesian analysis of some nonparametric problems. *Ann Stat.* 1973;1(2):209–30. Available from <http://www.jstor.org/stable/2958008>
- Fournier DA, Skaug HJ, Ancheta J, Ianelli J, Magnusson A, Maunder MN, et al. AD model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optim Methods Softw.* 2012;27(2):233–49. <https://doi.org/10.1080/10556788.2011.597854>.
- Gelfand AE, Dey DK. Bayesian model choice: asymptotics and exact calculations. *J R Stat Soc Ser B Stat Methodol.* 1994;56(3):501–14. Available from <http://www.jstor.org/stable/2346123>
- Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc.* 1990;85(410):398–409. Available from <http://www.jstor.org/stable/2289776>
- Gelman A, li Meng X, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. *Stat Sin.* 1996;6:733–807.
- Ghosh P, Albert PS. A Bayesian analysis for longitudinal semicontinuous data with an application to an acupuncture clinical trial. *Comput Stat Data Anal.* 2009;53(3):699–706. <https://doi.org/10.1016/j.csda.2008.09.011>.

- Ghosh SK, Mukhopadhyay P, Lu JC. Bayesian analysis of zero-inflated regression models. *J Stat Plann Infer.* 2006;136(4):1360–75. Available from <http://www.sciencedirect.com/science/article/pii/S0378375804004008>
- Ghosh S, Gelfand AE, Zhu K, Clark JS. The k-ZIG: flexible modeling for zero-inflated counts. *Biometrics.* 2012; 68(3):878–85. <https://doi.org/10.1111/j.1541-0420.2011.01729.x>.
- Green W. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working paper EC-94-10, Department of Economics. New York: New York University; 1994.
- Gschlößl S, Czado C. Modelling count data with overdispersion and spatial effects. *Stat Pap.* 2008;49:531–52. <https://doi.org/10.1007/s00362-006-0031-6>.
- Gupta PL, Gupta RC, Tripathi RC. Analysis of zero-adjusted count data. *Comput Stat Data Anal.* 1996;23(2):207–18. Available from <http://EconPapers.repec.org/RePEc:eee:csdana:v:23:y:1996:i:2:p:207-218>
- Hadfield JD. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J Stat Softw.* 2010;33(2):1–22. Available from <http://www.jstatsoft.org/v33/i02/>
- Hall DB. Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics.* 2000;56(4): 1030–9. <https://doi.org/10.1111/j.0006-341X.2000.01030.x>.
- Hall DB, Zhang Z. Marginal models for zero inflated clustered data. *Stat Model.* 2004;4(3):161–80. Available from <http://smj.sagepub.com/content/4/3/161.abstract>
- Hasan MT, Sneddon G. Zero-inflated Poisson regression for longitudinal data. *Commun Stat – SimulCompu.* 2009;38(3):638–53.
- Hasan MT, Sneddon G, Ma R. Pattern-mixture zero-inflated mixed models for longitudinal unbalanced count data with excessive zeros. *Biom J.* 2009;51(6):946–60. Available from <https://doi.org/10.1002/bimj.200900093>
- Hatfield LA, Boye ME, Carlin BP. Joint modeling of multiple longitudinal patient-reported outcomes and survival. *J Biopharm Stat.* 2011;21(5):971–91. Available from <http://www.tandfonline.com/doi/abs/10.1080/10543406.2011.590922>
- Heilbron DC. Zero-altered and other regression models for count data with added zeros. *Biom J.* 1994;36(5): 531–47. <https://doi.org/10.1002/bimj.4710360505>.
- Hilbe J. HNBLOGIT: stata module to estimate negative binomial-logit hurdle regression; 2005a. Statistical Software Components, Boston College Department of Economics. Available from <http://ideas.repec.org/c/boc/bocode/s456401.html>
- Hilbe J. HPLOGIT: stata module to estimate Poisson-logit hurdle regression. Statistical Software Components, Boston College Department of Economics; 2005b. Available from <http://ideas.repec.org/c/boc/bocode/s456405.html>
- Hsu CH. Joint modelling of recurrence and progression of adenomas: a latent variable approach. *Stat Model.* 2005;5(3):201–15. Available from <http://smj.sagepub.com/content/5/3/201.abstract>
- Jackman S. pscl: classes and methods for R developed in the political science computational laboratory. Stanford: Stanford University; 2012. R package version 1.04.4. Available from <http://pscl.stanford.edu/>
- Jones AM. Models for health care. In: Hendry D, Clements M, editors. *Oxford handbook of economic forecasting.* Oxford: Oxford University Press; 2011. p. 625–54.
- Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc.* 1995;90(430):773–95. Available from <http://www.jstor.org/stable/2291091>
- Kim S, Chang CC, Kim K, Fine M, Stone R. BLUP (REML) estimation of a correlated random effects negative binomial hurdle model. *Health Serv Outcome Res Methodol.* 2012;12:302–19. <https://doi.org/10.1007/s10742-012-0083-0>.
- Lam KF, Xue H, Bun CY. Semiparametric analysis of zero-inflated count data. *Biometrics.* 2006;62(4):996–1003. <https://doi.org/10.1111/j.1541-0420.2006.00575.x>.
- Lambert D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics.* 1992;34(1):1–14. Available from <http://www.jstor.org/stable/1269547>
- Li CS, Lu JC, Park J, Kim K, Brinkley PA, Peterson JP. Multivariate zero-inflated Poisson models and their applications. *Technometrics.* 1999;41(1):29–38. <https://doi.org/10.2307/1270992>.
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986;73(1): 13–22. Available from <http://biomet.oxfordjournals.org/content/73/1/13.abstract>
- Lillard LA, Panis CWA. Multiprocess multilevel modeling, version 2, user's guide and reference manual. Los Angeles: EconoWare; 1998–2003.
- Little RJA, Rubin DB. *Statistical analysis with missing data.* 2nd ed. Hoboken: Wiley; 2002.
- Liu H. Growth curve models for zero-inflated count data: an application to smoking behavior. *Struct Equ Model Multidiscip J.* 2007;14(2):247–79. <https://doi.org/10.1080/10705510709336746>.
- Liu L. Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Stat Med.* 2009;28(6):972–86. Available from <https://doi.org/10.1002/sim.3497>
- Liu L, Ma JZ, Johnson BA. A multi-level two-part random effects model, with application to an alcohol-dependence study. *Stat Med.* 2008;27(18):3528–39. Available from <https://doi.org/10.1002/sim.3205>
- Liu L, Strawderman RL, Cowen ME, Shih YCT. A flexible two-part random effects model for correlated medical costs. *J Health Econ.* 2010;29(1):110–23. Available from <http://www.sciencedirect.com/science/article/pii/S0167629609001386>
- Liu L, Strawderman RL, Johnson BA, O'Quigley JM. Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study. *Stat Methods Med Res.* 2012. Available from

- <http://smm.sagepub.com/content/early/2012/04/01/0962280212443324.abstract>
- Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput.* 2000;10(4):325–37. <https://doi.org/10.1023/A:1008929526011>.
- Majumdar A, Gries C. Bivariate zero-inflated regression for count data: a Bayesian approach with application to plant counts. *Int J Biostat.* 2010;6(1):27. Available from <http://ideas.repec.org/a/bpj/ijbist/v6y2010i1n27.html>
- Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. *J Health Econ.* 1998;17(3):283–95. Available from <http://www.sciencedirect.com/science/article/pii/S0167629698000253>
- Manning WG, Mullahy J. Estimating log models: to transform or not to transform? *J Health Econ.* 2001;20(4):461–94. Available from <http://www.sciencedirect.com/science/article/pii/S0167629601000868>
- Manning W, Morris C, Newhouse J, Orr L, Duan N, Keeler E, et al. A two-part model of the demand for medical care: preliminary results from the health insurance study. In: van der Gaag J, Perlman M, editors. *Health, economics, and health economics*. Amsterdam: North-Holland; 1981. p. 103–23.
- Manning WG, Basu A, Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. *J Health Econ.* 2005;24(3):465–88. Available from <http://www.sciencedirect.com/science/article/pii/S0167629605000056>
- Maruotti A. A two-part mixed-effects pattern-mixture model to handle zero-inflation and incompleteness in a longitudinal setting. *Biom J.* 2011;53(5):716–34. Available from <https://doi.org/10.1002/bimj.201000190>
- Millar RB. Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes' factors. *Biometrics.* 2009;65(3):962–9. <https://doi.org/10.1111/j.1541-0420.2008.01162.x>.
- Min Y, Agresti A. Random effect models for repeated measures of zero-inflated count data. *Stat Model.* 2005;5(1):1–19. Available from <http://smj.sagepub.com/content/5/1/1.abstract>
- Moulton LH, Halsey NA. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics.* 1995;51(4):1570–8. Available from <http://www.jstor.org/stable/2533289>
- Mullahy J. Specification and testing of some modified count data models. *J Econ.* 1986;33(3):341–65. Available from <http://www.sciencedirect.com/science/article/pii/0304407686900023>
- Muthén BO. Two-part growth mixture modeling; 2001. Unpublished Manuscript. Available from [http://pages.gseis.ucla.edu/faculty/muthen/articles/Article\\_094.pdf](http://pages.gseis.ucla.edu/faculty/muthen/articles/Article_094.pdf)
- Muthén BO, Muthén LK. *Mplus (Version 7)*. Muthén & Muthén; 1998–2012.
- Mwalili SM, Lesaffre E, Declerck D. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Stat Methods Med Res.* 2008;17(2):123–39. Available from <http://smm.sagepub.com/content/17/2/123.abstract>
- Neelon BH, O'Malley AJ, Normand SLT. A Bayesian model for repeated measures zero inflated count data with application to outpatient psychiatric service use. *Stat Model.* 2010;10(4):421–39. Available from <http://smj.sagepub.com/content/10/4/421.abstract>
- Neelon B, O'Malley AJ, Normand SLT. A bayesian two-part latent class model for longitudinal medical expenditure data: assessing the impact of mental health and substance abuse parity. *Biometrics.* 2011;67(1):280–9. Available from <https://doi.org/10.1111/j.1541-0420.2010.01439.x>.
- Neelon B, Ghosh P, Loeb PF. A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2012; Published online ahead of print. Available from <https://doi.org/10.1111/j.1467-985X.2012.01039.x>
- Olsen MK, Schafer JL. A two-part random-effects model for semicontinuous longitudinal data. *J Am Stat Assoc.* 2001;96(454):730–45. <https://doi.org/10.1198/016214501753168389>.
- Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics.* 2001;57(1):120–5. Available from <http://www.jstor.org/stable/2676849>
- Park RE. Estimation with heteroscedastic error terms. *Econometrica.* 1966;34(4):888. Available from <http://www.jstor.org/stable/1910108>
- Patil GP. Maximum likelihood estimation for generalized power series distributions and its application to a truncated binomial distribution. *Biometrika.* 1962; 49(1–2):227–37. Available from <http://biomet.oxfordjournals.org/content/49/1-2/227.short>
- Preisser JS, Stamm JW, Long DL. Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries Res.* 2012;46:413–23.
- R Development Core Team. R: a language and environment for statistical computing. Vienna; 2012. ISBN 3-900051-07-0. Available from <http://www.R-project.org/>
- Rabe-Hesketh S, Skrondal A, Pickles A. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *J Econ.* 2005;128(2):301–23. Available from <http://www.sciencedirect.com/science/article/pii/S0304407604001599>
- Raftery AM, Newton MA, Satagopan JM, Krivitsky PN. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In: Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, et al., editors. *Bayesian statistics 8*. Oxford: Oxford University Press; 2007. p. 1–45.
- Rathbun S, Fei S. A spatial zero-inflated poisson regression model for oak regeneration. *Environ Ecol Stat.* 2006;13:409–26. <https://doi.org/10.1007/s10651-006-0020-x>.

- Ridout M, Demétrio C, Hinde J. Models for count data with many zeros. Proceedings from the International Biometric Conference, Cape Town; 1998. Available from [https://www.kent.ac.uk/smsas/personal/msr/webfiles/zip/ibc\\_fin.pdf](https://www.kent.ac.uk/smsas/personal/msr/webfiles/zip/ibc_fin.pdf)
- Ridout M, Hinde J, DemAtrio CGB. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*. 2001;57(1):219–23. Available from <https://doi.org/10.1111/j.0006-341X.2001.00219.x>
- Rodrigues J. Bayesian analysis of zero-inflated distributions. *Commun Stat Theory Methods*. 2003;32(2):281–9. Available from <http://www.tandfonline.com/doi/abs/10.1081/STA-120018186>
- Roeder K, Lynch KG, Nagin DS. Modeling uncertainty in latent class membership: a case study in criminology. *J Am Stat Assoc*. 1999;94(447):766–76. Available from <http://www.jstor.org/stable/2669989>
- Rosen O, Jiang W, Tanner M. Mixtures of marginal models. *Biometrika*. 2000;87(2):391–404. Available from <http://biomet.oxfordjournals.org/content/87/2/391.abstract>
- SAS 9.1.3 Help and Documentation. Cary; 2000–2004. Available from: <http://sas.com/>
- Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461–4. Available from <http://www.jstor.org/stable/2958889>
- Silva FF, Tunin KP, Rosa GJM, Silva MVBD, Azevedo ALS, Verneque RdS, et al. Zero-inflated Poisson regression models for QTL mapping applied to tickresistance in a Gyr x Holstein F2 population. *Genet Mol Biol*; 2011;34:575–82. Available from [http://www.scielo.br/scielo.php?script=sci\\_arttext&pid=S1415-47572011000400008&nrm=iso](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-47572011000400008&nrm=iso)
- Skaug H, Fournier D, Nielsen A, Magnusson A, Bolker B. *glmmADMB: generalized linear mixed models using AD Model Builder*; 2012. R package version 0.7.2.12. Available from <http://glmmadmb.r-forge.r-project.org>
- Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B Stat Methodol*. 2002;64(4):583–639. <https://doi.org/10.1111/1467-9868.00353>.
- Stata Statistical Software: Release 12. College Station; 2011. Available from <http://stata.com/>
- Su L, Tom BDM, Farewell VT. Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics*. 2009;10(2):374–89. Available from <http://biostatistics.oxfordjournals.org/content/10/2/374.abstract>
- Su L, Brown S, Ghosh P, Taylor K. Modelling household debt and financial assets: a Bayesian approach to a bivariate two-part model; 2012.
- Tobin J. Estimation of relationships for limited dependent variables. *Econometrica*. 1958;26(1):24–36. Available from <http://www.jstor.org/stable/1907382>
- Tooze JA, Grunwald GK, Jones RH. Analysis of repeated measures data with clumping at zero. *Stat Methods Med Res*. 2002;11(4):341–55. Available from <http://smm.sagepub.com/content/11/4/341.abstract>
- Ver Hoef JM, Jansen JK. Spacetime zero-inflated count models of harbor seals. *Environmetrics*. 2007;18(7):697–712. Available from <https://doi.org/10.1002/env.873>
- Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*. 1989;57(2):307–33. Available from <http://www.jstor.org/stable/1912557>
- Walhin JF, Bivariate ZIP. *Models. Biom J*. 2001;43(2):147–60. Available from 10.1002/1521-4036(200105)43:2<147::AID-BIMJ147> 3.0.CO;2-5
- Welsh AH, Zhou XH. Estimating the retransformed mean in a heteroscedastic two-part model. *J Stat PlannInfer*. 2006;136(3):860–81. Available from <http://www.sciencedirect.com/science/article/pii/S0378375804003337>
- Williamson JM, Lin HM, Lyles RH. Power calculations for ZIP and ZINB models. *J Data Sci*. 2007;5:519–34. Available from <http://www.jds-online.com/v5-4>
- Winkelmann R. *Econometric analysis of count data*. 5th ed. Berlin: Springer; 2008. Available from [http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+368353176&sourceid=fwb\\_bibsonomy](http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+368353176&sourceid=fwb_bibsonomy)
- Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*. 1988;44(1):175–88. Available from <http://www.jstor.org/stable/2531905>
- Xiang L, Lee AH, Yau KKW, McLachlan GJ. A score test for overdispersion in zero-inflated poisson mixed regression model. *Stat Med*. 2007;26(7):1608–22. Available from <https://doi.org/10.1002/sim.2616>
- Xie H, McHugo G, Sengupta A, Clark R, Drake R. A method for analyzing longitudinal outcomes with many zeros. *Ment Health Serv Res*. 2004;6:239–46. <https://doi.org/10.1023/B:MHSR.0000044749.39484.1b>. Available from <https://doi.org/10.1023/B:MHSR.0000044749.39484.1b>
- Yau KKW, Lee AH. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Stat Med*. 2001;20(19):2907–20. Available from <https://doi.org/10.1002/sim.860>
- Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. *J Stat Softw*. 2008;27(8):1–25. Available from <http://www.jstatsoft.org/v27/i08/>
- Zhang M, Strawderman RL, Cowen ME, Wells MT. Bayesian inference for a two-part hierarchical model: an application to profiling providers in managed health care. *J Am Stat Assoc*. 2006;101(475):934–45. Available from <http://www.jstor.org/stable/27590773>
- Zurr AF, Saveliev AA, Ieno EN. *Zero inflated models and generalized linear mixed models with R*. Newburgh: Highland Statistics Ltd; 2012. Available from <http://www.highstat.com/book4.htm>



Theresa Henle, Gregory J. Matthews, and Ofer Harel

## Contents

<b>Introduction</b> .....	718
<b>Introducing the Basics</b> .....	719
Types of Disclosures and an Overview of Terms .....	719
Privacy for Different Types of Data .....	721
Balancing Privacy Versus Utility .....	722
<b>Privacy-Preserving Techniques</b> .....	723
Unperturbed and Perturbed Methods .....	723
Basic Methods for Limiting Disclosure Risk .....	723
More Sophisticated SDC Approaches .....	725
<b>Measuring Privacy</b> .....	728
K-Anonymity .....	728
Differential Privacy .....	729
<b>Conclusion</b> .....	730
<b>References</b> .....	730

### Abstract

When medical data are collected and disseminated for research purposes, the organization which releases the data has an ethical, and in most cases a legal, responsibility to maintain the confidentiality of the data relating to individuals involved. Striking a balance between

getting data to researchers and maintaining this confidentiality is becoming an increasingly tricky proposition. Methods developed in the field of statistical disclosure control aim to thwart potential disclosures of private information while still allowing researchers the ability to use the data. This chapter presents a survey of the main types of potential disclosure risks, an overview of the widely used disclosure control methods, and the most common techniques for measuring privacy.

---

T. Henle · G. J. Matthews  
Department of Mathematics and Statistics, Loyola  
University, Chicago, IL, USA  
e-mail: [theresahenle@gmail.com](mailto:theresahenle@gmail.com); [gjm112@gmail.com](mailto:gjm112@gmail.com)

O. Harel (✉)  
Department of Statistics, University of Connecticut, Storrs,  
CT, USA  
e-mail: [ofer.harel@uconn.edu](mailto:ofer.harel@uconn.edu)

## Introduction

In 1997, Governor William Weld of Massachusetts arrived to his office to find his medical record waiting for him in his mailbox. Just months before, he had authorized the release of state employee medical records for research purposes and assured the public of the safety of their release, since all explicit identifiers had been removed from the data (i.e., name, address, social security number, etc.). However, a young MIT graduate student by the name of Latanya Sweeney thought otherwise. Using publicly available voting records and the attributes that remained in the data (ZIP code, birthdate, and gender), Sweeney was able to positively identify Governor Weld's medical record. Then to make a point, Sweeney mailed Governor Weld's medical records to him directly (Sweeney 2002b, 2).

Sweeney's work exposed the vulnerability of medical data and triggered a widespread response by data publishers and policy makers alike. With the ever-expanding amount of publicly available data along with the increasing power of statistical tools, the confidentiality of data has become a growing concern. To abate the threat of releasing medical data which risks the privacy of individuals, the field of statistical disclosure control was born.

Statistical disclosure control (SDC) aims to develop and assess techniques to safely release data to interested parties, such as researchers (Matthews and Harel 2011). If done correctly, the distributed data should retain its utility to the researcher while fully ensuring the privacy of the participants involved. It is important to clarify that a breach of privacy in this case is not a result of a "hacking" incident but rather discovery of information through statistical techniques. In the case of hacking, an "intruder" or "adversary" gains unauthorized access to data in order to discover sensitive information. In statistical disclosure, an intruder is also interested in identifying sensitive information about specific individuals for malicious intent, but they go about learning private information using data which they are authorized to possess (Shlomo 2015, 201). An individual party who seeks to

learn some private attribute or attributes of the entities in a data set, often referred to as a "data snooper" in the statistical privacy literature, will use certain techniques, often in conjunction with other sources of data, to reveal private information about individual entities in the data set that were not meant to be known.

The field of statistical disclosure control has continued to grow in complexity as technology and access have made data more readily available. Today, data are ubiquitous: from the development of electronic health records (EHR) to personal electronic devices and wearables. The availability of data is an invaluable resource to researchers for improving our understanding of diseases, treatments, and the human body. However, releasing data to researchers while maintaining the privacy of the individuals involved is a unique balancing act.

Many agencies rely on publicly released data to perform their research. The United States Census Bureau and National Institutes of Health (NIH) are two of the largest sources of publicly available data. While these agencies seek to share data with as many researchers as possible, they are also required to protect the privacy of participants. Therefore, "investigators submitting a research application requesting \$500,000 or more of direct costs in any single year to NIH on or after October 1, 2003 are expected to include a plan for sharing final research data for research purposes, or state why data sharing is not possible" (Matthews et al. 2010).

It is clear that these organizations take privacy seriously, but what exactly is meant by privacy in these settings? Professor Alan Weston of Columbia University defined privacy as the right "to determine what information about ourselves we will share with others" (Fellegi 1972). What one person may wish to keep private, another may reveal publicly, but it is of utmost importance that the individual makes the choice to release the information rather than a third party. Therefore, when these preferences are unknown, researchers must default to assuming that participants desire privacy.

From a legal standpoint, steps have been taken to protect certain types of individuals' data. For



instance, in the mid-1990s, the US Congress passed several important pieces of legislation regarding private information. The Health Insurance Portability and Accountability Act (HIPAA 1996) was passed to protect medical data of individuals, and the Family Educational Rights and Privacy Act (FERPA) was designed to protect the educational data of individuals. HIPAA requires that obvious identifiers, such as name, birthdate, and ZIP code, be removed from medical records prior to data release unless explicitly authorized by an individual. However, even with these baseline privacy measures, data can still be at risk of disclosure. Legal guidelines alone are inadequate for ensuring the protection of sensitive data, and therefore greater measures must be taken to effectively maintain data privacy.

---

## Introducing the Basics

### Types of Disclosures and an Overview of Terms

The vulnerability of a given data set hinges on the structure and type of information contained within it. Rows in a data set are herein referred to as records, with each record containing a variety of attributes. An attribute is a value associated with some variable that reveals information about the record. Typically, in a medical data set, a record would be a person, and an attribute would be a descriptive characteristic of that person (e.g., age). Attributes need not be sensitive in nature, but because of their potential to identify an individual, they should be considered when establishing methods of privacy protection. There are three main types of data attributes present in medical health data, and the presence of these attributes can lead to certain privacy disclosure risks (Gkoulalas-Divanis and Loukides 2015). Disclosure risks define the process by which a data set can be breached.

The three types of attributes in medical data are direct identifiers, quasi-identifiers, and sensitive attributes. Direct identifiers are the most dangerous type of attributes from a privacy perspective,

as they “uniquely identify patients” (Gkoulalas-Divanis and Loukides 2015, 19).

For example, data containing names, social security numbers, or addresses make it easy for an attacker to directly identify an individual and are prohibited from being released under HIPAA guidelines. In all there are 18 direct identifiers that are unlawful to release in medical data. Quasi-identifiers also have the potential to identify a patient but require several working in combination in order to do so. Oftentimes, quasi-identifiers exist partially within the data set of interest and partially in a public data set from a separate source. By cross-tabulating between multiple sources, an adversary can identify individuals. Examples of quasi-identifying attributes could be demographic information (e.g., race, sex, age) or diagnosis codes. Lastly, sensitive attributes are the type of patient information that researchers are the most interested in protecting because it is often an information a patient “is not willing to be associated with” and therefore very often the target of an adversary’s attack (Gkoulalas-Divanis and Loukides 2015, 19). While the “specification of sensitive attributes is generally left to data owners” (Gkoulalas-Divanis and Loukides 2015, 19), common examples of sensitive attributes in medical data include serious diseases, such as mental illness and life-threatening conditions (Gkoulalas-Divanis and Loukides 2015, 19).

The presence of these attributes can lead to disclosure risks and ultimately threaten patients’ privacy.

The main types of disclosure risk are:

1. Identity disclosure (or reidentification)
2. Attribute disclosure
3. Inferential disclosure

Identity and attribute disclosures are the two most commonly cited types of disclosure risk. Identity disclosure occurs when an adversary is able to identify an individual based on their record within a data set. If direct identifiers are present, or the correct combination of quasi-identifiers, a patient is at risk of identity disclosure. According to Latanya Sweeney, “It has been estimated that over 87% of US citizens can be re-identified based

on a combination of only three demographics (ZIP code, gender and date of birth)” (Sweeney 2000, 1). Attribute disclosure occurs when an adversary is able to learn and reveal sensitive attributes about an individual in the data set. Identity disclosure is often a precursor to attribute disclosure: first, if a record in the data set is linked to an individual, then a private attribute about that individual is learned. When data are in tabular form, attribute disclosure is most likely to occur in a column that contains a degenerate distribution of cell counts, as opposed to a column with more uniformly distributed counts. In general, “a row or column with large cell counts would have less risk of identity or attribute disclosure as compared to a row or column with small counts” (Shlomo 2015, 215).

While attribute disclosure often takes place after an identity disclosure has occurred, there are other ways in which attribute disclosure can occur such as group attribute disclosure, disclosure by differencing, and disclosure by linking tables. In these scenarios, individuals need not be identified in order for disclosure risk to occur, such as in the case of group attribute disclosure, where sensitive information is exposed about a group within the data set, rather than an individual person. For example, say that in a certain data set, all people within a small ZIP code all have a diagnosis of high blood pressure. If you know of an individual in this data set who lives in that particular ZIP code, you now also know that they have a high blood pressure diagnosis. Note that no identity disclosure has taken place here as no particular record was matched to any individual; however, an attribute disclosure has still taken place.

Another way for attribute disclosure to occur is from what is called disclosure by differencing. An example of this is where two nested tables (i.e., one table is a subset of another table) are subtracted from one another exposing sensitive information previously unknown (Shlomo 2015, 214). This is often a problem when data are accessed through flexible table generation. In such a scenario, a user is not given access to the data set as a whole but must submit queries to a database. This is problematic from a privacy perspective when a

user can difference two separate query results to gain confidential information about a single person in the study. For example, a user might submit a query based on all men under the age of 34, and then subsequently submit a query based on all men under the age of 35. If the difference is 1, then you have identified a unique combination of attributes of an individual in the data set. Note that in many cases, if the user were to submit a query to the database based *ONLY* on men who were *EXACTLY* 34, in this case *ONLY* one record, many query systems will suppress a cell in a table if the value is below some prespecified threshold precisely because a small cell count can lead to attribute disclosures. By submitting nested queries and differencing, a user is gaining access to information that would potentially be suppressed on its own. The best way to avoid disclosure by differencing is to release a single data set as opposed to providing a system that allows for flexible table generation.

Similar to disclosure by differencing, disclosure by linking tables occurs when two tables originate from the same source and therefore have the potential to be linked by common cells or common margins. This can potentially allow an adversary to discover the SDC technique which guards the data and with it the original data values. The best way to avoid disclosure by linking tables is to ensure that the margins and cells of tables be made consistent (Shlomo 2015, 214).

The last type of disclosure risk is inferential disclosure. Inferential disclosure relies on probability and/or modelling to expose attributes with a high degree of confidence. One way in which inferential disclosure can occur is by way of regression model if the model has “very high predictive power” (i.e., the dependent and explanatory variables are highly correlated). This specific case of inferential disclosure is called model, or predictive, disclosure. Willenborg and de Waal (2001) explain predictive disclosure using microdata containing information about an individual’s gender, age, occupation, location, and income. If an adversary knows certain characteristics (i.e., gender, age, occupation, and location) about a specific individual in the data set, they can build a regression model to predict an unknown value

(say income). Through modelling, the adversary has achieved the “predictive distribution of the target’s income” (Willenborg and de Waal 2001); however, many would argue that such inferences are the goal of statistical modelling, and therefore it extends beyond reasonable privacy protection.

Disclosing by differencing, which was previously discussed, can also be considered a type of inferential disclosure. For this reason, inferential disclosure is often associated with web-based interactive data. The recent emergence of interest in inferential disclosure has catalyzed a need for stricter forms of privacy guarantees and has formed the basis for differential privacy, a popular privacy measurement discussed later in this chapter (Shlomo 2015, 203).

There are examples showing that an individual does not even need to be in a data set in order to be at risk of inferential disclosure. For example, say a person were to release their genetic sequencing data (as was done in the Human Genome Project) and they are found to have a specific gene. Then later a subsequent study with new participants reveals that specific gene makes a person extremely likely develop a rare form of cancer. The person who released their genetic data was not a part of the study which determined the effects of this specific gene, but they are now subject to its inference: that they will likely develop this form of rare cancer. However, according to the current privacy guidelines, researchers are only beholden to protect the privacy of those individuals included in the published data set.

### Privacy for Different Types of Data

There are two common structures for how published data sets are presented: microdata and tabular data. Microdata are “data containing observations on the individual level” such as social surveys or general health surveys. Tabular data “contains frequency counts or magnitude data,” which is more typical of business surveys (Shlomo 2015, 202). Methods for protecting privacy in tabular data can be classified as either

pre-tabular, post-tabular, or some combination of these two methods.

Pre-tabular methods are implemented on original microdata before it is transformed into a tabular data set. Post-tabular methods modify data for privacy purposes after the data set is already in its tabular form. The most common forms of post-tabular disclosure control techniques are methods utilizing random rounding and cell suppression.

Data collection techniques can also impact the type and degree of vulnerability associated with a data set. Data can be generated either through sampling from a larger population or by collecting complete information on the population through what is called a census. Though sampling is more common because it requires fewer resources, census data are popular for government publications. Census data contain unique challenges in preventing identity and attribute disclosure because there is no uncertainty about membership in the data. Conversely, sampling as part of data collection process obscures the ability to make inference on frequency counts, thereby reducing the possibility of identity disclosure.

Lastly, the introduction of new types of data has opened the door to new concerns over data vulnerability. With advancements in technology, large quantities of data are being generated from processes that simply did not exist until recently such as location data collected from cell phone or individuals’ genome-wide data. With the Global Positioning System (GPS) capability of modern cell phones, it has become easy to track the locations of millions of people at once, yielding massive quantities of location data. Location data are problematic from a privacy perspective as it has the potential to jeopardize confidential information about individuals such as where they live or where they work.

Genetic data, such as data used in genome-wide association studies (GWAS), for example, are another type of data with unique and ever-expanding complexities. Since the inception of the Human Genome Project in 1990, and its completion in 2003, scientists have made great strides in understanding the human genetic structure. The availability of human genetic data is crucial for the continued growth of genetic research. However,

this new source of personal information has been accompanied by its very own set of privacy concerns. Homer et al. (2008) demonstrated “the ability to accurately and robustly determine whether individuals are in a complex genomic DNA mixture” (Homer et al. 2008, 1). Homer argued the need for more stringent methods for sharing and combining individual genotype data across studies, since “sharing only summary data do not completely mask identity” (Homer et al. 2008, 9). Gymrek et al. (2013) demonstrated a shocking breach of privacy when they were able to identify a man whose genetic data had been used in the Human Genome Project. By using a sequence of that man’s genetic data along with information about his location and age, they were able to cross-reference genealogical databases and public records to discover the identity of the individual (Gymrek et al. 2013).

### Balancing Privacy Versus Utility

Protection of privacy is not the only concern a data distributor must consider when preparing a data set. Establishing privacy protections can often come at a loss of data utility, causing the data to become unusable or inaccurate. This is true regardless of the type of data used or how a publisher defines privacy in their data set. In general, “it is always possible to increase the privacy of any specific data release, but this almost assuredly comes with a loss of data utility” (Matthews et al. 2010). Therefore, publishers should think carefully about the balance between privacy and utility when preparing a data set for publication. Further, if the distributor knows that the data will be used for a specific purpose, this is often helpful information in choosing an appropriate disclosure control method.

There are two major frameworks for how to measure data utility on data sets where SDC techniques have been applied. The more general of the two is information loss measures, which do not presume any specific intended use for the data. Alternatively, the utility-constrained approach considers the way the data are intended to be used and preserves data utility for that task specifically.

When the intended use of the data is unknown, publishers can utilize information loss measures to quantify data utility. These measures seek to minimize data distortion in a broad sense, making the data more versatile but relinquishing the promise of utility for any specific task. Information loss measures compare the difference in utility of the altered data set to the original data set. This difference in utility is task specific, meaning the altered data set may perform accurately for some desired analysis but inaccurately for others. For example, if a data set was altered such that marginal totals remained constant, it is likely that when testing for average values, one would achieve perfect retention of data utility. However, in the same case, the relationship between variables may not be maintained, and therefore a subsequent regression analysis would be ineffective at reflecting the true nature of the data. This is often the case when electronic health records and data sets that contain multiple variables of interest need to be shared (e.g., demographics and diagnosis codes). The two attribute types cannot be anonymized separately; however, it is difficult to “preserve data utility when anonymizing both of the attribute types together” (Gkoulalas-Divanis and Loukides 2015, 30). For this reason, it is desirable that publishers reveal the type of privacy-preserving method used and that researchers consider the effect that method may have on the tests they wish to perform.

When the intended use of the data is known, the data distributor will likely opt for utility-constrained approach when measuring data utility. The specific type of utility constraints employed depends entirely upon the intended use of the data, but in general a utility constraint prevents the anonymization procedure from generating data that will produce vastly different results when compared to the original data. For example, a data publisher may want to add noise to a variable but may check that the resulting sample mean of the modified data is relatively close to the true sample mean of the original data set. Further, constraints preventing combinations of variables that are not possible (i.e., a record of a woman with prostate cancer) could also be considered here.

## Privacy-Preserving Techniques

The following section provides an overview of the most common and basic of privacy-preserving techniques utilized by researchers. These techniques provide the foundation for the more sophisticated techniques to follow. In general, there are two basic approaches for statistical disclosure control: (1) restricting access to the data, “for example, by limiting its use to approved researchers within a secure data environment (safe access),” or (2) implementing statistical disclosure techniques to protect the data prior to release (“safe data”). It is typical for a publisher to use some combination of both approaches when releasing sensitive health data; however in this section the focus is on creating “safe data” (Shlomo 2015, 201).

## Unperturbed and Perturbed Methods

Statistical disclosure control (SDC) methods protect the privacy of medical health data by preventing adversaries from uncovering sensitive information. SDC methods can be broken down into two categories: perturbed and unperturbed methods. Perturbed methods work by adding noise to data, thereby obscuring the true values. This can be done either through a probability distribution approach or a value distortion approach. The probability distribution approach identifies the distribution of the data and samples from that distribution to create a new data set of plausible values. The value distortion approach perturbs data by building decision tree classifiers for the data where each element is assigned random noise. Then, the perturbed data are sampled to match the distribution of the original data set. In general, the value distortion approach is considered to be more effective than the probability approach; however random additive noise if properly filtered can lead to privacy compromises. In general, perturbed methods are more difficult to implement, both because they require higher statistical sophistication and the added inconvenience that one would need the details on how the data were perturbed in order to analyze it.

Unperturbed methods do not alter the data but rather seek a limitation of detail in order to preserve privacy. Methods surveyed here include suppression, generalization, rounding, sampling, and disassociation. The main advantage of unperturbed techniques is that the risk of altering relationships between variables is less than with perturbed techniques. This is because unperturbed techniques protect the data by reducing the detail rather than altering the data through noise addition. However, when a study calls for specific detailed information, unperturbed methods may obscure the data in such a way that deem it no longer useful.

Before selecting an SDC method to implement, the first step is to remove obvious identifiers in your data set. Such identifiers include name, social security number, birth data, and home address. As previously stated, “87% of the population in the United States have reported characteristics that made them unique based only on ZIP code, gender and date of birth” (Sweeney 2002b, 2). Removing identifiers is necessary to preserving data privacy, but by itself it is not usually enough to protect the privacy of individuals. Recalling the example from the introduction, the Massachusetts Group Insurance Commission released data under the impression that it was safe, having removed all obvious identifiers. However, Sweeney (2002b) was able to cross-reference the released medical records with publicly available voting records and identified the specific medical record of former Massachusetts Governor William Weld.

## Basic Methods for Limiting Disclosure Risk

Among the many disclosure control techniques, the simplest are generalization, suppression, rounding, sampling, randomization, and additive noise.

### Generalization

Generalization works by binning similar values of sensitive variables into overarching generalized terms. For example, rather than providing separate

diagnosis codes for different forms of cancer and risk small counts that are more easily exposed, generalization would bin all cancer diagnoses into one or more subsets that contain higher counts. The generalized term is still semantically consistent for the specific diagnosis, such as replacing lymphoma with “cancer.” However, the generalized term does not offer as much detail, thereby obscuring sensitive values and preventing an attacker from distinguishing a specific diagnosis code from within the generalized term. Generalization is best implemented when the number of quasi-identifier attributes is small and when the intended use applies to a range of data rather than a specific class. The more attributes involved, the greater number of generalized terms required to ensure privacy, which will lead to the degeneration of data utility. When a user seeks information about a group or range of values, such as people from a certain geographic area, generalization provides privacy without any utility loss.

Generalization is susceptible to composition attacks when multiple independent data sets are available. If two equivalence classes share only one sensitive value, an adversary can deduce sensitive information by differencing. For example, the raw data set may contain information about the age of an individual. Rather than reporting exact age, generalization would report, for instance, the age group (e.g., 20–29, 30–39, etc.).

### Suppression

Generalization is a favorite technique due to its “faithful” information properties. Although the granularity of detail may not be fine, the accuracy of values is pristine, and the relationship between variables is not disturbed. Suppression is an extreme case of generalization where the most generalized term is utilized. Therefore, possible generalization is preferred because it is a superior technique in preserving data utility. Top-bottom coding is another specialized case of generalization that applies specifically to extreme values. For example, there may only be one person in the study that is 99 years old; however they may be ten individuals over the age of 80. An agency may record specific age for individuals in the study less than 80 years old but utilize the

generalized term “over 80” for those ten individuals. Generalization provides the basis for more complex partitioning privacy preservation models such as *k*-anonymity, *l*-diversity, and *t*-closeness (Li et al. 2015, 187). These techniques will be discussed in the following section.

### Sampling

The last of the unperturbed methods is sampling. A familiar technique for data collection, sampling is also very useful in privacy preservation. In Skinner et al. (1994), they make the case that “population uniqueness will be a sufficient condition for an exact match to be verified as correct.” In other words, samples obscure population uniqueness and stifle an adversary’s ability to cross-reference uniqueness between data sets in a linkage attack. Sampling also does extremely well in balancing privacy with utility, as proper sampling techniques should yield data that are an accurate representation of the population. More so, sampling is an “easy technique to implement and the resulting sampled data are relatively easy to analyze” (Matthews and Harel 2011).

### Randomization

Perturbation techniques work by modifying the contents of the data in some way as the basis for privacy preservation. Randomization is the most basic perturbation technique and can be used for both microdata and tabular data sets. In randomization, noise is randomly added to the original values (or aggregated values) obscuring the true values contained within an individual’s record and making it difficult for an adversary to infer sensitive information. The simplest application of randomization would be random noise generated from an independent and identical distribution with a positive variance and mean of zero. In this case, the addition of random noise “will not change mean of the variable for large data sets, but will introduce more variance,” (Shlomo 2015, 210) which may harm the ability of a researcher to make accurate statistical inferences. Randomization is best used within “small homogenous sub-groups in order to use different initiating perturbation variance for each sub-group” (Li et al. 2015, 180). The use of subgroups for noise

addition is also beneficial in maintaining accurate relationships between variables in the data (Matthews et al. 2010).

### **Rounding**

Rounding is another perturbation method generally applied to tabular data sets. As the name implies, in rounding, observations are rounded up or down to the nearest multiple of a pre-determined rounding base. For example, if the rounding base was 0.1 and the observed value was 0.3, the probability of rounding up would be 0.3, whereas the probability of rounding down would be 0.7. Another method is controlled rounding, “which allows the sum of the rounded values to be the same as the rounded value of the sum of the original data” (Shlomo 2015, 218). A problem with random rounding occurs however when cells generated in different tables lack consistency. When this happens, “the true cell count can be learned by generating many tables containing the same cell and observing the perturbation patterns” (Shlomo 2015, 218). An alternative to controlled rounding is semi-controlled random rounding which “ensures that rounded internal cells aggregate to the controlled rounded total” (Shlomo 2015, 218), thereby enforcing consistency across all generated tables.

## **More Sophisticated SDC Approaches**

### **Micro-agglomeration, Substitution, Subsampling, and Calibration**

MASSC (Micro-agglomeration, Substitution, Subsampling, and Calibration) combines various simple techniques to create a more robust approach to data privatization. The names of the procedure lay out the four steps: micro-agglomeration, substitution, subsampling, and calibration. In micro-agglomeration, records are sorted by the level of risk, dependent on the presence of identifying variables. High-risk identifying variables are called core variables, as compared to noncore identifying variables which generally pose less risk to privacy. Core identifying variables pose a greater risk because they are generally easier for an intruder to obtain. The

greater risk to a record exists when core identifying variables are present and are unique. Disclosure control techniques are then applied to groups of records based on their risk category. Substitution techniques are used to perturb the data. Substitution methods include random rounding, randomization, data swapping, and synthetic data (the last two methods mentioned here are discussed in detail below). The data are then sampled from the perturbed data set to add another layer of privacy protection and to “help reduce the bias caused by substitution (Singh et al. 2003). A unique and desirable property of MASSC is “that both disclosure risk and information loss can be controlled for simultaneously” (Matthews et al. 2010).

### **Data Swapping**

Data swapping is a privacy-preserving technique popular for its ease of use. Although the technique was originally intended for use on contingency tables, it has become a popular technique for microdata as well. The procedure involves “the swapping of values of variables for records that match on a representative key” (OECD 2008, 126). In other words, given a data set with a sensitive variable, such as cancer diagnoses where it is necessary to protect against attribute disclosure, some records containing that diagnosis code will swap with another record exclusively within that variable. Variables that are not considered sensitive will be untouched by this process, for the record swapping applies only to the variable of concern. An example of this can be viewed through the following table. In the real data set, the sensitive variable is the participant’s cancer diagnosis. In the swapped data, the second and third rows are swapped within the cancer column, so that the participant in row 2 now is associated with a cancer diagnosis and the participant in row 3 is no longer (Fig. 1).

Data swapping is best used when one is simply interested in univariate statistics. Since records are swapped one for one, the marginal totals remain intact, making univariate statistics unchanged. Multivariate relationships, on the other hand, between the affected variable and the other variables in the data set may not be correctly

**Fig. 1** A simple example of data swapping

Real Data			Swapped Data		
age	sex	cancer	age	sex	cancer
50	F	yes	50	F	yes
40	M	yes	40	M	no
60	M	no	60	M	yes
70	F	no	70	F	no

maintained in the data swapping process. However because only the sensitive variable is affected, multivariate analysis can be effectively conducted by simply excluding the sensitive variable.

When implementing data swapping, one must be wary of swaps that may result in impossible or improbable records. An example of this would be if a data set contained the variables gender and diagnosis code and swapping resulted in a record suggesting that a female was diagnosed with prostate cancer.

As mentioned, data swapping is effective for both tabular and microdata sets. However, implementation procedures may differ depending on the type of data used. Since microdata provides subject-level information rather than variable aggregates, “many more swaps must be made to preserve the level of privacy” (Matthews et al. 2010). Determining the number of swaps necessary was deemed “computationally impractical” by Fienberg and McIntyre (2004), and therefore totals should be preserved only approximately for best practice. As previously stated, arguably the greatest advantage to this method is that it is very easy to implement. All that is required to utilize this method is microdata and a random number generator (Moore 1996).

### Rank-Based Proximity Swapping

A more contemporary alternative to the traditional data swapping method is a rank-based proximity swapping proposed by Greenberg (1987) and popularized by Moore (1996). Unlike data swapping, values of sensitive

variables are swapped with records where the value of the sensitive variable falls within a certain range of the original record. This restriction allows the relationships between the sensitive variable and the other variables in the data set to be more effectively maintained than in traditional data swapping where the process of swapping is strictly random.

### Data Shuffling

Sarathy and Muralidhar (2002) proposed a further extension of data swapping called data shuffling. Data shuffling utilizes a conditional distribution approach where all of the marginal distributions remain intact. More so, pairwise monotonic relationships in the original data are maintained. They are therefore able to increase the privacy protection without sacrificing the high level of utility achieved through data swapping (Sarathy and Muralidhar 2002). For this reason, this method has become standard for many, including the United States Bureau of the Census and the Office for National Statistics in the UN (Lauger et al. 2014).

### Randomized Response

Randomized response is a technique for survey data closely related to the previously discussed technique of randomization (Warner 1965; Greenberg et al. 1969). In randomized response, respondents will answer a question truthfully with some given probability (e.g., a coin flip). Otherwise, they are instructed to answer the question with the opposite of the truthful answer.



This technique is most useful when the questions being asked require the respondent to reveal sensitive information about themselves that they may not be comfortable answering truthfully.

As a simple example, consider this sensitive survey question “Are you an injection drug user?” Before answering the question, the respondent flips a coin, with the outcome remaining unknown to the administrator of the survey, for whether or not they will answer truthfully. If the coin lands heads, the respondent is directed to answer the question truthfully. However, if the coin lands tails the respondent is directed to answer the question untruthfully (i.e., “Yes” if they ARE NOT an injection drug user and no if they ARE an injection drug user.) Since the probability of a truthful and untruthful answers is known, these type of data are useful for many types of analyses, and the uncertainty in the answers provided introduces a level of confidentiality for the data because an adversary cannot be sure whether the response is actually correct or not.

### **PRAM**

The randomized response method could also be applied to mask raw microdata, even when there is no speculation of false response. This special case of randomized response is called Post-randomization Method (PRAM). In PRAM, “for each observation, the real value of a sensitive field would be released with some probability and its opposite would be released with some other probability” (Matthews et al. 2010, 6). The result is essentially a randomized addition of noise. The difference between PRAM and randomized response is that “PRAM is applied after completion of a survey and formation of the data set, whereas randomized response is applied during the interviewing” (Willenborg and De Waal 2001, 32). While in randomized response, the random mechanism is “independent of the true score” in PRAM, “the true value is known and one can therefore condition on this value when defining the probability mechanism used to perturb the data” (Matthews et al. 2010, 7). This distortion of the data, however, requires the researcher to have information about the randomization mechanism in order to effectively analyze the data.

### **Synthetic Data**

Synthetic data are a perturbation approach wherein artificial data sets are generated from the original data through the process of multiple imputation (Rubin 1993). Multiple imputation is a technique which is traditionally used in missing data settings where missing values are filled in by sampling from an appropriate distribution (Rubin 1987). Multiple imputation requires the creation of multiple completed data sets, each time replacing the originally missing cells with plausible values. Then the statistical analysis of interest is performed on each completed data set, and the results are combined across the data sets using Rubin’s combining rules (Harel and Zhou 2007).

When creating synthetic data, however, the purpose is not to impute missing values but rather to create usable data sets which conceal sensitive information. This is accomplished by viewing sensitive attributes of the data as missing values and replacing them using multiple imputation techniques. In the case of fully synthetic data generation, all sensitive variables in the original data set are viewed of as missing, and the posterior predictive distribution is used to generate a synthetic “population.” This is repeated to create several fully imputed data sets, each of which is considered a synthetic population. Lastly random samples are drawn from each synthetic population, and this collection of data sets are released to the public. A popular alternative to fully synthetic data is partially synthetic data. This technique is similar to fully synthetic data, however, imputing values only for sensitive attributes, rather than on the entire collection of sensitive variables. An agency may select individual attributes or entire variables, depending on their privacy needs, and the resulting data set would then contain both real and synthetic data values.

In either fully or partially synthetic data, researchers can perform an analysis on each data set and combine their results using rules set forth in Raghunathan et al. (2003) (for full synthetic data sets) or Reiter (2003) (for partially synthetic data sets), which are slightly modified versions of Rubin’s combining rules.

Synthetic data sets are desirable for their ease of analysis. Similar to multiple imputation,

researchers can compute analysis across several synthetic data sets and pool their results for a combined estimate. However, since synthetic data relies on artificial data, this can leave researchers pondering the validity of their findings. Raghunathan et al. (2003) and Reiter (2005) set out to assure researchers that synthetic data have merit by showing that for accurate imputation models, resulting analyses yield almost identical results to that of the original data. However, “if the model for imputation is incorrect or inaccurate, the resulting analysis from the synthetic data will yield parameter estimates that are much different than those estimated from the actual data. As such, synthetic data sets are only as good as the models used for imputation” (Matthews et al. 2010, 10).

Though the idea of synthetic data sets was slow to catch on, it has become a widely used and highly successful disclosure control technique. The most highly visible user of this technique is the United States Census Bureau. They have used partially and fully synthetic data in several of their publicly released data sets, including the yearly release of “On the Map” data (Shlomo 2015, 228). This data generated by personal GPS devices provides information on the locations of individuals. However, it would be rather easy to identify individuals based on their home and place of work, making it a statistical disclosure concern. However, through the use of synthetic data sets, the Census Bureau has been able to release this data without risking the privacy of the individual’s involved (Shlomo 2015, 228).

---

## Measuring Privacy

Statistical disclosure techniques are designed to protect the privacy of individuals by masking sensitive attributes and preventing disclosure risk and, at the same time, producing data sets that are useful for analysis and inference. While assessing data utility is relatively straightforward (i.e., how similar is the analysis when using the raw data vs. the analysis when using the protected data), the

assessment of privacy is substantially more difficult. This is due to the many different kinds of disclosures that exist and that measures of privacy will be different depending on the type of disclosure.

Measures of privacy based on reidentification assess the probability of accurately identifying a subject in the published data set. Spruill (1982) studied the privacy of some masking procedures (e.g., normal random error, random rounding, data swapping, etc.). They proposed a measure of confidentiality based on the percentage of records in the published data set that could be linked to the original record. Paass (1988) discusses a measure of privacy based on matching subjects in the published data set to some additional available information, and their proposed measure of privacy is based on the percentage of records that are at risk for identification. They concluded that the best way to protect privacy is to release as few variables as possible, since the greater number of variables, the more difficult it is to protect against a privacy attack. Larger data sets (i.e., data sets with many variables) require substantial modifications to the data in order to maintain a robust level of privacy though this comes at the cost of potentially dramatic reductions in data utility. They also note that the addition of random noise does little to protect privacy in this framework.

## K-Anonymity

K-anonymity is an additional privacy measure for data that has had suppression and generalization techniques applied to it. In general, k-anonymity promises a level of anonymization for any given record in the data by focusing on quasi-identifiers. As previously mentioned, quasi-identifiers are “a set of attributes in a data set that could be used for matching with an external database” (Matthews et al. 2010, 16). Quasi-identifiers put an individual at greatest risk for disclosure when certain combination of attributes is rare or, in the worst case, unique. More formally, k-anonymity states that every set of quasi-identifiers that appears in the

Raw Data			Privacy Preserved Data		
Age Group	Gender	Procedure	Age Group	Gender	Procedure
60-65	F	Appendectomy	60-70	F	Appendectomy
60-65	F	Cataracts	60-70	F	Cataracts
60-65	M	Cataracts	60-70	M	Cataracts
60-65	M	Biopsy	60-70	M	Biopsy
65-70	M	Hip Replacement	60-70	M	Hip Replacement
65-70	M	Craniotomy	60-70	M	Craniotomy
70-75	M	Whipple	70-80	M	Whipple
70-75	M	Whipple	70-80	M	Whipple
70-75	F	Appendectomy	70-80	F	Appendectomy
75-80	F	Cataracts	70-80	F	Cataracts
75-80	F	Hip Replacement	70-80	F	Hip Replacement

**Fig. 2** An example of making data 2-anonymous

data set must appear at least  $k$ -times. Thus there is at most a  $1/k$  chance of reidentifying a particular record (Sweeney 2002a).

In application, take the following two data sets with variables age group, gender, and surgical procedure. Generalization has been applied to the age group variable in the “Privacy Preserved Data” so that there are fewer overall age groups and less potential to uniquely identify an individual based on their age. In the raw data set, the combination of age group = 70–75 and gender = F is a unique combination. However, after the data are generalized, every combination of quasi-identifiers (age and sex) appears at least two times. Therefore, the privacy of this data can be measured as 2-anonymous by the principle of  $k$ -anonymity. Note, however, that both 70–80-year-old men had the Whipple procedure, thus causing an attribute disclosure even though no record was uniquely identified (Fig. 2). Extensions of  $k$ -anonymity include  $l$ -diversity (Machanavajjhala et al. 2007) and  $t$ -closeness (Li et al. 2007).

## Differential Privacy

Differential privacy was proposed in Dwork (2006) and provides formal privacy guarantees and results in one of the strongest versions of privacy. The basic idea of differential privacy is that no single observation in a data set should be overly influential in terms of a function of the data. This means that for a given function of the data, the value of this function will not change “very much” if ANY one single record in the data is modified. Data sets that differ by only one record are referred to as *neighboring data sets*. (There are actually two distinct meanings for neighboring data sets: one refers to a record being modified, and the other refers to a record being removed. Here the second definition is used.) Exactly how much values of the function are allowed to change is controlled by the parameter epsilon ( $\epsilon$ ), with smaller values guaranteeing more privacy and larger values guaranteeing less. Guaranteeing that the result of a function of the data does not change “very

much” is accomplished by creating a randomized version of the function rather than the exact value of the function. This results in very strong privacy. Practically speaking, this type of privacy guarantees that if an adversary knows all records in the data set except for 1, they will still not be able to learn very much about the last unknown observation, and this would be true for ANY set of observations.

Example data set: 1,2,3,4,100

As an example, imagine that a data set contained five observations, and one of these observations was a large outlier. The mean of this data set is 22. However, rather than release the value of 22, a randomized version of the mean is released by simply adding some noise to the true value of the sample mean. If no noise was added and the true value of the sample mean was released, if an intruder knew the first four values in this data set and the mean of 22, the intruder can learn the exact value of the remaining data value. However, since the released value of the mean is random, the exact value of the remaining data point is uncertain. The exact amount of noise that is necessary to add is based on a data releaser’s choice of the  $\epsilon$  parameter and what is referred to as the sensitivity of the function. The sensitivity of the function is the absolute value of the largest possible difference in the function computed on the actual data and a neighboring data set across ALL neighboring data sets.

As an example of sensitivity, if we consider the neighboring data base with the outlier removed, the mean is now 2.5. This yields a sensitivity of  $|22-2.5| = 19.5$  as this is the largest difference across all neighboring databases.

One of the simplest and most popular ways to achieve  $\epsilon$ -differential privacy is to add Laplace noise to the true value of the function of interest calculated on the full data set where the mean of the Laplace distribution is 0 and the variance is determined by the value of  $\epsilon$  and the sensitivity of the function.

Extensions of differential privacy include several relaxed versions including  $(\epsilon, \delta)$  – indistinguishability (Nissim et al. 2007) and probabilistic differential privacy (Machanavajjhala et al. 2008). Matthews et al. (2010) and Matthews and Harel

(2012) place the problem of measuring privacy in a hypothesis testing framework and use the receiver-operating characteristic (ROC) curve to assess the privacy of a database.

---

## Conclusion

It is estimated that 2.5 quintillion bytes of data are collected every day (DN Capital 2015). These massive quantities of data allow researchers and businesses to perform analyses that were previously unthinkable. However, as the amount of data that are collected is increased, concerns about data privacy will naturally follow. Malicious data users often possess the capabilities to expose sensitive attributes and reveal the identities of individuals in a publicly available data set. This is especially problematic in medical data, where sensitive attributes might refer to a serious illness or diagnosis. Therefore, it is of the utmost importance that proper consideration be given to protecting patient privacy prior to releasing medical data, which requires consideration beyond simply removing direct identifiers. It is imperative that statistical disclosure control techniques be applied to data to ensure a standard of privacy.

---

## References

- DN Capital – Venture Capital. Beyond ‘big data’ to data driven decisions. 2015. [Dncaptical.com/thoughts/beyond-big-data-to-data-driven-decisions/](https://dncaptical.com/thoughts/beyond-big-data-to-data-driven-decisions/).
- Dwork C. Differential privacy. In: ICALP. Springer Verlag; 2006. p. 1–12. MR2307219.
- Fellegi IP. On the question of statistical confidentiality. *J Am Stat Assoc.* 1972;67(337):7–18.
- Fienberg SE, McIntyre J. Data swapping: variations on a theme by Dalenius and Reiss. In: Domingo-Ferrer J, Torra V, editors. *Privacy in statistical databases*. Vol. 3050 of lecture notes in computer science. Berlin/Heidelberg: Springer; 2004. p. 519. [https://doi.org/10.1007/978-3-540-25955-8\\_2](https://doi.org/10.1007/978-3-540-25955-8_2).
- Gkoulalas-Divanis A, Loukides. A survey of anonymization algorithms for electronic health records. In: Gkoulalas-Divanis A, Loukides G, editors. *Medical data privacy handbook*. Cham: Springer International Publishing; 2015. p. 17–34.
- Greenberg B. Rank swapping for masking ordinal micro-data. Technical report, U.S. Bureau of the Census (unpublished manuscript), Suitland; 1987.

- Greenberg BG, Abul-Ela A-LA, Simmons WR, Horvitz DG. The unrelated question randomized response model: theoretical framework. *J Am Stat Assoc.* 1969;64(326):520–39. MR0247719.
- Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science.* 2013;339:321–4.
- Harel O, Zhou X-H. Multiple imputation: Review and theory, implementation and software. *Statistics in Medicine* 2007;26, 3057–3077. MR2380504
- Health Insurance Portability and Accountability Act (HIPAA); Pub.L. 104–191, 110 Stat. 1936, enacted August 21, 1996.
- Homer N, Szlinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genet* 2008;4(8): e1000167. <https://doi.org/10.1371/journal.pgen.1000167>
- Lauger A, et al. Disclosure avoidance techniques at the U.S. census bureau: current practices and research. Research report series. 2014. [www.census.gov/srd/CDAR/cdar2014-02\\_Discl\\_Avoid\\_Techniques.pdf](http://www.census.gov/srd/CDAR/cdar2014-02_Discl_Avoid_Techniques.pdf)
- Li N, Li T, Venkatasubramanian S. t-closeness: privacy beyond k-anonymity and l-diversity. In: *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on; 2007.* p. 106–15.
- Li H, et al. Differentially private histogram and synthetic data publication. In: Gkoulalas-Divanis A, Loukides G, editors. *Medical data privacy handbook*. Cham: Springer International Publishing; 2015. p. 35–58.
- Machanavajjhala A, Kifer D, Gehrke J, Venkatasubramanian M. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* 2007;1 (1), 3.
- Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L. Privacy: theory meets practice on the map. In: *International Conference on Data Engineering*. Cornell University Computer Science Department, Cornell; 2008. p. 10.
- Matthews GJ, Harel O. Data confidentiality: a review of methods for statistical disclosure limitation and methods for assessing privacy. *Statist Surv.* 2011;1–29. <https://doi.org/10.1214/11-SS074>.
- Matthews GJ, Harel O. Assessing the privacy of randomized vector valued queries to a database using the area under the receiver-operating characteristic curve. *Health Serv Outcome Res Methodol.* 2012;12 (2–3):141–55.
- Matthews GJ, Harel O, Aseltine RH. Assessing database privacy using the area under the receiver-operator characteristic curve. *Health Serv Outcome Res Methodol.* 2010;10(1):1–15.
- Moore Jr R. Controlled data-swapping techniques for masking public use microdata. *Census Tech Report.* 1996.
- Nissim K, Raskhodnikova S, Smith A. Smooth sensitivity and sampling in private data analysis. In: *STOC '07: Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing; 2007.* p. 75–84. MR2402430.
- OECD Statistics. Glossary of statistical terms – data swapping definition, stats. 2008. [Oecd.org/glossary/detail.asp?ID=6904](http://Oecd.org/glossary/detail.asp?ID=6904)
- Paass G. Disclosure risk and disclosure avoidance for microdata. *J Bus Econ Stat.* 1988;6(4):487–500.
- Raghunathan TE, Reiter JP, Rubin DB. Multiple imputation for statistical disclosure limitation. *J Off Stat.* 2003;19(1):1–16.
- Reiter JP. Inference for partially synthetic, public use microdata sets. *Survey Methodology* 2003;29 (2), 181–188.
- Reiter JP. Releasing multiply imputed, synthetic public use micro- data: an illustration and empirical study. *J Royal Stat Soc Series A Stat Soc.* 2005;168(1):185–205. MR2113234.
- Rubin DB. Multiple imputation for nonresponse in surveys. Hoboken: Wiley; 1987. MR0899519.
- Rubin DB. Comment on “statistical disclosure limitation”. *J Off Stat.* 1993;9:461–8.
- Sarathy R, Muralidhar K. The security of confidential numerical data in databases. *Inf Syst Res.* 2002;13 (4):389–403.
- Shlomo N. Statistical disclosure limitation for health data: a statistical agency perspective. In: Gkoulalas-Divanis A, Loukides G, editors. *Medical data privacy handbook*. Cham: Springer International Publishing; 2015. p. 201–30.
- Singh A, Yu F, Dunteman G. MASSC: a new data mask for limiting statistical information loss and disclosure. In: *Proceedings of the Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality; 2003.* p. 373–94.
- Skinner C, Marsh C, Openshaw S, Wymer C. Disclosure control for census microdata. *Journal of Official Statistics* 1994;10, 31–51.
- Spruill NL. Measures of confidentiality. *Proceedings of the section on survey research methods, American Statistical Association.* 1982
- Sweeney L. Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.
- Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. *Int J Uncertainty Fuzziness Knowledge Based Syst.* 2002a;10 (5):571–88. MR1948200.
- Sweeney, L. Simple demographics often identify people uniquely. Carnegie Mellon University, data privacy working paper 3. 2002b.
- Sweeney L. K-anonymity: a model for protecting privacy. *Int J Uncertainty Fuzziness Knowledge Based Syst.* 2002c;10(5):557–70. MR1948199.
- Warner SL. Randomized response: a survey technique for eliminating evasive answer bias. *J Am Stat Assoc.* 1965;60(309):63–9.
- Willenborg L, de Waal T. *Elements of statistical disclosure control*. New York: Springer; 2001. MR1866909.



Cynthia Robins

## Contents

<b>Introduction</b> .....	734
<b>What Is Qualitative Research?</b> .....	734
<b>A Sampling of Qualitative Health Research Studies</b> .....	735
<b>Methods of Qualitative Data Collection</b> .....	736
Informed Consent .....	736
Data Collection Approaches .....	736
To Record or Not to Record? .....	746
<b>Data Analysis</b> .....	747
Simplifying the Data .....	747
<b>Summary</b> .....	750
<b>References</b> .....	750

### Abstract

Qualitative methods were introduced into the world of health services research about three decades ago, and have begun to gain traction among researchers only in the last decade and a half. Despite the growing interest in what qualitative research can tell us about the human understanding of and experience of illness, skepticism remains among some scholars about the value-added of non-numeric research. Indeed, say some, if the findings from qualitative studies are not generalizable,

can it really be called “research” at all? A philosophical debate about qualitative versus quantitative research, however, is not within the purview of this chapter. Rather, the pages that follow have a threefold objective: First, to set forth the epistemological assumptions of qualitative research, which are fundamentally different from their quantitative counterparts (and thus non-comparable); second, to provide the reader with a brief review of seminal works in qualitative health research and to discuss what factors have contributed to the growing interest in such approaches; and, lastly, to provide readers with some basic tools of qualita-

C. Robins (✉)  
Westat, Rockville, MD, USA  
e-mail: [cynthiarobins@westat.com](mailto:cynthiarobins@westat.com)

tive data collection and analysis that can serve as templates for their own qualitative health studies. The overarching goal of the chapter is to argue that when conducted systematically by well-trained scholars, qualitative research has the potential to offer us valuable insights into the socio-cultural factors that underlie the interpretation of diseases, the illness experience, and the search for meaningful and effective treatments.

---

## Introduction

Qualitative research involves the analysis of nonnumeric data obtained through data collection methods such as in-depth interviews, focus groups, and observations. Although it is gaining traction in the field of health research, qualitative methods are a relatively recent addition and many health researchers are still unsure about the value it brings to the field. This chapter thus has a threefold objective: first, to briefly set forth how qualitative research differs philosophically from quantitative research. Although readers may be aware that qualitative and quantitative data collection methods are different, it is also important to understand that the epistemologies that drive each approach are very distinct. Second, this chapter will review the recent history of the use of qualitative methods in health research. It will look at the social and political processes that co-occurred with the rise of qualitative research in health and healthcare, as well as briefly describe some of the signal studies in the field over the last 20–25 years. Finally, the chapter will provide the reader with an overview of the fundamentals of qualitative research, including data collection techniques and the basics of the analytic process. The goal of this chapter is to demonstrate that when skillfully and appropriately implemented, qualitative research can offer critical insights into such phenomena as patients' experiences, service providers' views of disease processes and treatments, as well as key socio-cultural factors that underlie the structure and delivery of health care in different communities.

## What Is Qualitative Research?

Much ink has been spilled over the years on what scholars have often referred to as the “qualitative-quantitative divide” or the “science wars,” with scholars arguing about the supremacy of one approach over the other (classic examples include Popper 1934; Kuhn 1962; Sokal 1996). Those arguments will not be revisited here; instead, this chapter starts from the assertion shared by Hopper (2008), Morse (1991), and others that qualitative research is not “better” or “worse” than its scientific counterparts, but legitimate in its own right and on its own terms. In research, as in life, some things simply serve different ends. To ask, “Which is better, an electric drill or a reciprocating saw?” is a pointless question without knowing the project that is to be undertaken. Once the project objectives are clearly defined, however, there is a right answer. Selecting the wrong tool for the job, perhaps because it is the one the researcher likes or knows the best, can have disastrous consequences for the work at hand.

For the reader who is not well versed in philosophy and who simply wants to know if he or she should consider using qualitative methods on a project, the following brief distinction is worth considering. Quantitative research is rooted in a philosophy of positivism, i.e., the belief that there is objective truth in the world that can be discovered through the application of the scientific method (see discussion in Ponterotto 2005: 128–129). Through controlled experiments, careful measurements, and agreed-upon numeric indicators (e.g., p-values, confidence intervals), science aims to gather increasing amounts of information about the world. Scientific advances, thus, are viewed as getting us closer to a full understanding of an empirical reality.

Qualitative research, by contrast, has its roots in a philosophy of knowledge often called phenomenology or social constructivism (Morse and Field 1996; Alvesson and Skoldberg 2009). This philosophical position is quite distinct from positivism in that it asserts that human beings' interactions with the world are always mediated by a socially or culturally provided system of symbols – language, beliefs, values, and rules for behavior.

For example, cultural anthropologists, such as Geertz (1973) and others, operate from the fundamental position that because all of our experiences are filtered through a cultural lens, there is no way to get at what some might refer to as “truth”: What we believe to be truth is someone else’s heresy. Qualitative research thus aims not to uncover “the” truth, but rather “their” truth, often with the explicit aim of creating a foundation of understanding between populations in conflict. The objective of qualitative research generally “is to experience, reflect, organize, understand, and communicate” (Estroff 1981, xvi).

Different paradigms lead to different questions and thus different ways to answer those questions. The reader should not wonder if quantitative or qualitative research is “better,” even if a dissertation advisor prefers one approach or the other. The question really is: Which is the right tool to meet the research objectives? If the research questions seek to understand volumes or counts (e.g., “How much. . .,” “How many. . .,” “How often. . .”), the reader should look to quantitative data collection techniques. If the interest is in what the world looks like from another’s point of view, perhaps with an eye towards understanding motivations (e.g., “Why do. . .,” “How do. . .”), then qualitative approaches are likely the best option. Once the research objectives are clearly defined, the choice – or even choices (a researcher may use multiple methods) – will become obvious. What remains is for the researcher to learn how to use the tool properly.

---

## A Sampling of Qualitative Health Research Studies

The use of qualitative methods to learn how people make sense of their medical experiences – either as recipients or providers of health care – is a fairly new phenomenon, dating back only to the mid-1980s. Arguably one of the greatest contributors to this epistemological shift was the patients’ rights movement, which sought to challenge the hegemony of the medical system. The mental health consumer rights movement, for example, was an early catalyst for change, as

former patients of psychiatric hospitals decried some of the abuses they had endured under the guise of psychiatric medicine. Members of this movement, such as activist-writer Judi Chamberlain, wanted to tell their side of the story, i.e., to share their experiences and perspectives being “treated” under lock and key. Chamberlain’s landmark work, *On Our Own: Patient-Controlled Alternatives to the Mental Health System* (Chamberlain 1978), generously incorporated first-person accounts from mental health consumer/survivors and, in so doing, made a compelling argument that there could be two sides to the medical story.

At roughly the same time, qualitative research methods were being incorporated more broadly into health research. One of the earliest such endeavors was *Making It Crazy* (Estroff 1981), anthropologist Sue Estroff’s account of how individuals with psychiatric disorders were putting together their lives outside of the state mental hospital. Estroff’s research methods included participant observation, in-depth interviews, and ad hoc encounters in the community with formerly hospitalized psychiatric patients. The result of her time “in the field” is an ethnography that offers the reader critical insights into the patients’ perspectives on psychiatric medications, work, and their relationships with others in the community. Other important works were Emily Martin’s (1987) *The Woman in the Body*, which examined how the language used to describe women’s reproductive systems influences the medical establishment’s approach to pregnancy, childbirth, and menopause, and Joan Cassell’s (1991) *Expected Miracles*, an ethnography of surgeons and their perceptions of and behaviors around their work. Cassell’s study offered one of the earliest examples of what she referred to as “studying up,” i.e., research into the lives of powerful members of a society rather than the dispossessed.

Occurring about the same time was the adoption of anthropological methods by scholars in other fields, notably within the field of nursing research. This movement is perhaps best epitomized by the work of Janice Morse, a registered nurse who went on to receive advanced degrees in both nursing and anthropology. In the mid- to



late-1980s, Morse edited several seminal works (Morse 1988, 1989a, b) that introduced nursing scholars to the epistemology and methods of qualitative research. These approaches proved essential to cross-cultural nursing, where the nurses' and patients' understanding of illness and appropriate treatment might be worlds apart. Effective care could best be provided, nursing scholars argued, when these different perspectives were taken into account.

By the early 1990s, qualitative health research – while still not fully accepted by the health research establishment – was becoming both ubiquitous and highly influential. Efforts to combat the rapid spread of HIV/AIDS both in the USA and in other countries and cultures demanded research methods that could uncover how a group's behaviors and beliefs about the disease were contributing to transmission. Anthropologists and other social scientists using qualitative methods rose to the occasion. Paul Farmer's (1993) ethnography of the interpretation of HIV/AIDS in Haiti early in the epidemic was a landmark work, demonstrating the critical role of both history and culture in people's illness experiences. The rapid pace of globalization over the last two decades – and the concomitant potential for pandemics – has only increased the essentialness of qualitative research methods in the health fields (Ramin 2009; Ebola Anthropology Response Platform).

Seminal journals, such as *Qualitative Health Research*, first published in 1990, and, roughly a decade later, the *International Journal of Qualitative Methods* as well as the online [Forum for Qualitative Research](#), have provided important avenues for scholars to share their health research findings and learn about new and innovative approaches to qualitative methods. There are also increasingly well-attended research conferences, including the Qualitative Methods Conference and the Qualitative Health Research Conference (held alternating years), both sponsored by the International Institute of Qualitative Methodology at the University of Alberta, Canada, and the International Congress of Qualitative Inquiry held annually at the University of Illinois in Urbana. These forums are testament to

the acceptance that qualitative methods have something valuable to offer health care practitioners and researchers.

---

## Methods of Qualitative Data Collection

### Informed Consent

Before collecting any data, the researcher must ensure that he or she follows the guidelines for the protection of human subjects. Key to this is the informed consent process, whereby the study participant is told what his or her rights are as a research subject before any data are collected. The core elements of the informed consent process are provided in Fig. 1 and include a description of the study sponsor and how the data will be used, the risks and benefits to the participants, and the voluntary nature of participation, among others.

These elements must be provided to participants in a written informed consent form that the study participant and researcher will both sign and date before data are collected. It is also good practice to review these key elements verbally with participants before beginning an interview or focus group discussion. Examples of how the information can be verbally reviewed with participants can be found in the sample in-depth interview protocol (Fig. 2) and sample focus group guide (Fig. 3).

### Data Collection Approaches

There are three primary qualitative data collection strategies that will help researchers understand how the study subjects experience the world and, in turn, make meaning of those experiences: focus groups, in-depth interviews, and participant observation. Each of these is described in turn.

### In-Depth Interviews

In-depth interviews (IDIs) are known by a number of other terms, including semi-structured interviews, unstructured interviews, one-on-one interviews, and guided conversations, among

### Key Elements of Informed Consent for Research Participants

All study participants need to be given the following information, as applicable, before any data are collected.

- They must be told that the the study involves research
- The purposes of the research must be explained to the participants (e.g., the research is for a dissertation; the study is being funded by a particular agency and why)
- How long the subject's participation in the study will last (e.g., one hour interview)
- A description of the procedures to be followed (e.g., participant will be asked questions, the interview will be recorded with the participant's permission)
- Identification of any procedures which are experimental
- Participants must be informed about any risks or discomfort they may experience during the study. If the study involves more than minimal risk, participants must be told what compensation or treatments will be available to them and whom to contact.
- Participants also must be told if they can expect any benefits from participating in the study (if none, disclose that to the participants as well)
- A disclosure of appropriate alternative procedures or courses of treatment, if any, that might be advantageous to the subject
- The researcher should explain his or her procedures to maintain the confidentiality of the subjects. This includes how data will be stored as well as steps the researcher will take to make sure not to disclose the participants' identities in any written or presented materials.
- All participants should be reminded that their participation in the study is voluntary, and that there are no penalties or loss of benefits if they decide not to participate or drop out of the study.

A checklist of the elements of informed consent can be obtained from the U.S. Department of Health and Human Services, Office for Human Research Protections website:

<https://www.hhs.gov/ohrp/regulations-and-policy/guidance/checklists/index.html> - accessed 9.5.17

**Fig. 1** Elements of informed consent

others. All of these terms, however, can mislead the outside observer, who may believe that the researcher and interviewee are having an hour-long discussion bounded by few, if any, parameters. Although the researcher may use an IDI guide that at first glance appears lean, each guide must be carefully crafted to clearly and narrowly frame the topic for the respondent. The guide must also include targeted probes to help the interviewer ensure that, within the general frame, the respondent addresses the areas of

greatest interest to the study. IDIs thus require a skilled interviewer who has superior active listening skills and who fully understands how the interviews are intended to support the goals and objectives of the study. He or she must have the intellectual flexibility to move simultaneously between the respondent's narrative and the study aims, gently guiding the narrative back to the frame when needed, but also listening for new and relevant information that may merit additional probing.

A hypothetical example will help to illustrate the process. Sjogren's syndrome has been characterized as an "invisible illness," a disease that may be disabling to the individual who has it, but that offers few visible symptoms to the outside observer (Donohue and Siegel 2000). Sjogren's symptoms can range from the annoying, such as dry eyes, mouth, and skin, to the disabling, including crippling fatigue, joint pain, and even lymphoma (<http://www.sjogrens.org>). People living

### **Introduction**

Hello, my name is [NAME ]. Thank you for agreeing to talk with me today about how Sjogren's has impacted your social life and experiences. This study, which I am conducting for my dissertation at University, is being funded by [ORGANIZATION].

### **Informed Consent**

Before we get started there are a couple of things I need to mention. First, this is a research project and your participation is voluntary. You can stop the interview at any time; if I ask you a question you would prefer not to answer, just tell me and we'll move on to the next one. Second, I will do everything I can to maintain your confidentiality. I will not attach your name to any data files and I will never use your name in any of my writings from this study. I may use quotes from the people I interview, but the names of interviewees will not be attached to those quotes. I will also remove any information from that quote that might identify you to others.

There are no direct benefits to you from participating in this research, although your story will contribute to my efforts to create a resource manual for others living with Sjogren's. The main risk to you from participating in this study is that you might experience some emotional distress from telling your story. I have a list of resources I will give you at the end of the interview if you would like.

Finally, with your permission, I would like to audio record our interview today. This is so that I do not have to take many notes while we are talking and I can focus on the story you are sharing with me. The recording will also help me to be more accurate when analyzing all of the interviews.

Do you have any questions before we get started? [ANSWER ANY QUESTIONS]

Do I have your permission to audio record the interview? [IF YES, TURN ON THE AUDIO RECORDER]

### **Interview Questions**

I am interested in learning what it is like to live with Sjogren's, which some people have referred to as an "invisible illness." By that they mean the disease can have profound effects on the person who has it, but it offers few obvious clues to outside observers that the person is ill. What I'd like to do today is have you tell me a story about your experiences living with Sjogren's in a world that may not know you are sick. You can start your story wherever you like, and you can talk as long as you like,

**Fig. 2** (continued)

but tell me everything you think is important for me to fully understand your experiences living and coping with this invisible illness.

**PROBES (IF NEEDED):**

- In what ways, if any, has this unseen illness affected
  - ...your professional life?
  - ...your home life with family members?
  - ...your social life?
- How long did it take for you to get a diagnosis after you first began experiencing symptoms of the disease? Why do you think that was?
- How well do you think the medical community recognizes symptoms of the disease?

What have you done that has been most effective in getting your work colleagues, family, and friends to understand what it's like living with Sjogren's?

What, if anything, do you wish you had had – or would still like to have – to help others understand your experiences?

Is there anything else about your experience living with an invisible illness like Sjogren's that you haven't talked about, but that you think is important for me to hear to fully understand your experiences?

TURN OFF AUDIO RECORDER AND THANK THE INTERVIEWEE

**Fig. 2** Sjogren's IDI guide

with Sjogren's may have to make a number of significant lifestyle changes, but often without the support of family or friends, who think the person looks "perfectly healthy." This illness captures the attention of a hypothetical researcher, who wants to interview people with Sjogren's to understand their experiences working and living with a disease that no one can see. She hopes to develop a guidebook that can offer sufferers some coping strategies, including talking points that will help the person with the disease explain the illness to people in their social network. Thus, in addition to hearing about her subjects' social experiences, she also wants to hear from her interviewees what steps they have taken that have been successful in explaining their condition to others, as well as any additional supports they might like to have.

The first question in the IDI guide must set the parameters of the interview for study participants, but also give them sufficient leeway to be able to share their experiences and their points of view. Thus it may look like the following:

I am interested in learning what it is like to live with Sjogren's, which some people have referred to as an "invisible illness." By that they mean the disease can have profound effects on the person who has it, but it offers few obvious clues to outside observers that the person is ill. What I'd like to do today is have you tell me a story about your experiences living with Sjogren's in a world that may not know you are sick. You can start your story wherever you like, and you can talk as long as you like, but tell me everything you think is important for me to fully understand your experiences living and coping with this invisible illness.

This opening statement is by no means “unstructured” or even “semi-structured” because the interviewee is told precisely the bounds within which her narrative should remain: She is being asked to describe the social aspects of the illness, i.e., what it is like to live with a disease that others cannot see. She is *not* being asked to give a full accounting of her symptoms, the specialists she sees, or the treatments she is undergoing.

In a perfect world, each interviewee would spontaneously relate a story that fully addresses all areas of interest to the researcher. But because this is an imperfect world, the protocol should

include probes so that the interviewer makes sure the respondent addresses the key domains of the research. Possible probes for this hypothetical study might include the following:

- *In what ways, if any, has this unseen illness affected*
  - . . .*your professional life?*
  - . . .*your home life with family members?*
  - . . .*your social life?*

Notice that these three probes cover the key dimensions of interest (work, family, friends), but

### Introduction

Hello, my name is [NAME]. Thank you all for agreeing to participate in this focus group discussion today about how Sjogren's has impacted your social lives and experiences. This study, which I am conducting for my dissertation at University, is being funded by [ORGANIZATION].

### Informed Consent

Before we get started there are a couple of things I need to mention. First, this is a research project and your participation is voluntary. If you decide you no longer want to participate, you can leave the discussion at any time; if I ask you a question you would prefer not to answer, just tell me and I'll move on to the next person. Second, there are no right or wrong answers to any of the questions that I ask today. You may disagree with what someone else says during the group, and that's ok. It's important that I hear different perspectives. Third, I will do everything I can to maintain your confidentiality. I will not attach your names to any data files and I will never use your names in any of my writings from this study. I may use quotes from the focus groups, but the names of interviewees will not be attached to those quotes. I will also remove any information from that quote that might identify you to others.

There are no direct benefits to you from participating in this research, although your story will contribute to my efforts to create a resource manual for others living with Sjogren's. The main risk to you from participating in this study is that you might experience some emotional distress from telling your story. I have a list of resources I will give you at the end of the interview if you would like.

Finally, with your permission, I would like to audio record our interview today. This is so that I do not have to take many notes while we are talking and I can focus on the story you are sharing with me. The recording will also help me to be more accurate when analyzing all of the interviews.

Do you have any questions before we get started? [ANSWER ANY QUESTIONS]

Do I have your permission to audio record the interview? [IF YES, TURN ON THE AUDIO RECORDER]

**Fig. 3** (continued)

### Warm-Up Exercise

I'd like to start off by taking just a couple of minutes for us to get to know each other. So if you would, please tell us just your first name and , briefly, something that you think is unique about yourself – an interesting hobby, somebody famous that you once met, or an interesting place that you have visited. [GO AROUND THE ROOM; MODERATOR SHOULD GO LAST]

### Discussion Questions

First, I'd like to get a sense of how long each of you has been living with Sjogren's.

#### *Symptoms*

Sjogren's has often been called an "invisible illness," that is, a disease in which the symptoms can have profound effects on the individual who has it, but in ways that may not be obvious to outside observers. Let's talk about this idea for a little bit. What symptoms do you all regularly experience that may affect your daily life, but that you don't think are noticed by people you work, live, or socialize with.

#### *Social Impacts*

Think for a moment about your professional lives, your home life and family, or your social activities with friends: Tell me about an instance in which you had to make a lifestyle adjustment to accommodate your symptoms, but that you didn't think was fully understood by others, such as your work colleagues or family and friends.

#### *Strategies*

What have any of you done that has been effective in getting your work colleagues, family, and friends to understand what it's like living with Sjogren's?

What, if anything, do you wish you had had – or would still like to have – to help others understand your experiences?

#### **Close**

Is there anything else about your experiences living with an invisible illness like Sjogren's that you haven't talked about, but that you think is important for me to hear to fully understand your experiences?

**Fig. 3** Sjogren's focus group guide

also remind the interviewee that the focus of the research is on the impact of the invisibility of the illness, in short, the social effects. As an example, perhaps in responding to the third probe, the interviewee describes how she can no longer do a variety of physical activities because of extreme joint pain: she can no longer garden, take weekend hikes with friends, go for her morning run, or walk the dog. Clearly the loss of an array of

activities that she once enjoyed is important to the respondent. The researcher can be sympathetic to this wider loss, but needs the respondent to hone in on the one activity that relates to the social impact of the illness, namely, hiking with friends. Thus, an appropriate probe at this juncture might be: *How well do your friends understand why you stopped going on weekend hikes?* A simple probe such as this is respectful of the

respondent's need to describe these myriad losses, but in a way that steers the narrative back to the research focus.

The researcher should also be aware that the above probes may not be exhaustive and that interviewees may add a dimension to their experiences that the researcher did not anticipate. Perhaps three of the first four interviewees start off their narratives by recounting how many years it took for a doctor to finally recognize their symptoms and provide a diagnosis. The research team is not interested in hearing about the clinical manifestations of the illness per se, but these narratives suggest that the symptoms may be invisible to the medical community as well. Thus, two new questions for subsequent interviews might be:

- *How long did it take for you to get a diagnosis after you first began experiencing symptoms of the disease? Why do you think that was?*
- *How well do you think the medical community recognizes symptoms of the disease?*

Qualitative researchers should always be alert to the possibility that data collection will add entirely new dimensions to their understanding of the issue and be prepared to modify the interview protocol, as needed.

Recall, too, that in this example, the researcher's aim is to create a guidebook for people living with Sjogren's syndrome, one that includes successful coping strategies and other resources that readers might find useful. Two additional questions might be included in this protocol:

- *What have you done that has been most effective in getting your work colleagues, family, and friends to understand what it's like living with Sjogren's?*
- *What, if anything, do you wish you had had – or would still like to have – to help others understand your experiences?*

Finally, because this is a bounded narrative, one steered in a particular direction both by the

questions and the interviewer; it is always a good idea to give the respondent a last opportunity to talk about something that may have been given short shrift during the interview:

- *Is there anything else about your experience living with an invisible illness like Sjogren's that you haven't talked about, but that you think is important for me to hear to fully understand your experiences?*

Interviewees generally will not take this as an open invitation to talk about their illness experiences for another hour, for two key reasons: First, the protocol was structured so as to give them sufficient latitude to tell their stories; and, second, this summary question reiterates that the boundaries of the discussion are around the social invisibility of the illness. The full interview guide, along with the critical elements of informed consent, is shown in Fig. 2.

With a skilled interviewer, the above example should generate 45 min to an hour's worth of rich, detailed narrative. And after conducting another 12–15 such interviews, likely the researcher will have sufficient information to at least begin to create the desired end-product. Should there be critical information gaps, additional in-depth interviews can be conducted to complete the product.

### Focus Groups

Focus groups are small group (6–10 person) discussions in which a moderator uses a carefully designed protocol to elicit participants' input on the topic of interest (Morgan and Krueger 1997). While IDIs offer depth on an issue, focus groups provide the *breadth* necessary when beginning to explore a particular issue. This is a particularly valuable data collection approach in the formative stages of a project, when the study team is still learning the range of study participants' experiences and perspectives on the topic. Focus groups may also be the data collection method of choice when project resources (money, time) are limited. Sometimes this is unavoidable, although the

researcher should remain cognizant that the lack of depth necessarily limits what one can say about the findings.

Two aspects of the group dynamic need to be considered when developing the discussion protocol for a focus group (see Fig. 3). The first is that even though they have consented to participate, some participants may be a little nervous, uncertain how much they want to reveal about themselves in this group of strangers. Thus, the protocol should include a brief (5 min) “icebreaking” exercise to get rid of any lingering participant butterflies and to begin to create connections between those in the room. A particularly effective strategy is to ask participants to tell the group something unique or interesting about themselves, such as a hobby they have, someone famous they once met, or some unusual place they have visited. Having the moderator also participate in this exercise is an excellent way for him or her to establish rapport with the group members before reassuming control of the discussion.

The second aspect of the group dynamic that must be taken into account is that the protocol questions – and the moderator – must balance the desire for detailed information against the need to hear from as many participants as possible. In the hypothetical Sjogren’s study, the initial questions to a focus group may look something like the following:

- *First, I’d like to get a sense of how long each of you has been living with Sjogren’s.*
- *[Next] Sjogren’s has often been called an “invisible illness,” that is, a disease in which the symptoms can have profound effects on the individual who has it, but in ways that may not be obvious to outside observers. Let’s talk about this idea for a little bit. What symptoms do you all regularly experience that may affect your daily life, but that you don’t think are noticed by people you work, live, or socialize with.*
- *Think for a moment about your professional lives, your home life and family, or your social activities with friends: Tell me about an*

*instance in which you had to make a lifestyle adjustment to accommodate your symptoms, but that you didn’t think was fully understood by others, such as your work colleagues or family and friends.*

These questions endeavor to get at the same issues as those covered in the IDI, but in a way that does not allow any one person to tell his or her life story. For example, the second question about symptom experience is clearly directed to the group (“symptoms which you *all* regularly experience”) and implies that some of these symptoms may be shared and so discussed. The third question also restricts any participant’s input to a single example – enough to give the group (and the research team) a sense of the *breadth* of experiences of people living with Sjogren’s. Summary questions can be roughly identical to those used in the IDI:

- *What have any of you done that has been effective in getting your work colleagues, family, and friends to understand what it’s like living with Sjogren’s?*
- *What, if anything, do you wish you had had – or would still like to have – to help others understand your experiences?*
- *Is there anything else about your experiences living with an invisible illness like Sjogren’s that you haven’t talked about, but that you think is important for me to hear to fully understand your experiences?*

An important thing to remember is that because the researcher must necessarily limit each person’s input to the discussion, it will limit the depth around any one person’s contribution to the research topic – often, some important details about a person’s story may be missing. This is the trade-off of conducting focus groups instead of in-depth interviews, so make sure this is the right data collection strategy to answer the research questions. If the researcher has to conduct focus groups because there are constraints on project resources, there may be a temptation to over-



interpret the data, e.g., the analyst may see differences between groups that are, at best, lightly supported by the evidence. Analysts should remember to work with the information they do have and let unanswered questions serve as the basis for their next data collection effort.

### Participant Observation/Ethnography

This data collection strategy is invaluable when the researcher believes that subjects' experiences and perceptions can best be understood in the context in which those experiences occur. The researcher gains an understanding how the world looks through their eyes by observing their behaviors in the location of interest and asking countless questions, some targeted, some spontaneous (Murchison 2010).

A new researcher may find it tricky trying to create interview protocols for this kind of study, in part because so much about the context is unknown, anticipating what specific questions to ask and of whom can feel like an exercise in futility. In addition, the field site oftentimes is not in a location that lends itself to scheduled in-depth interviews or focus groups. That said, the researcher does know the core study goals and, very generally, the roles of those within the context who might be able to address them. Instead of trying to develop a series of interview guides applicable to every conceivable situation, the researcher might consider developing a table of question domains by interviewee role. The table ensures that the researcher will remain focused on the goals and objectives of the study, but in a way that provides the latitude required for ad hoc encounters in the field. In addition, having a single, focused study document can prove helpful if the work is being conducted by a team.

Another hypothetical example can illustrate this approach: A community clinic is struggling to meet the needs of local residents because residents are reluctant to go there. Community members say they are often treated rudely by staff, and avoid the clinic altogether so as not to be subjected to the abuse. Without an alternative source of care nearby, however, many residents end up not receiving any medical care at all. Indeed, surveys conducted with community members

show high rates of morbidity from otherwise very treatable conditions, such as diabetes and high blood pressure.

Participant observation would be an excellent research strategy for trying to understand *what* is happening in these aversive encounters, *why* it is occurring, and if the findings point to a possible solution. Locations where the researcher might consider conducting observations could include the clinic waiting room, intake stations where staff make the initial patient contact, weekly team meetings of various staff (e.g., administrators, clinicians, and support staff), and locations throughout the community where the researcher can hear from local residents (e.g., senior centers, community library). An example of the kinds of question domains that might be relevant to this hypothetical study and the categories of people who might be able to speak to each domain, is illustrated in Fig. 4.

This example table is by no means exhaustive, but suggests areas where there may be a disconnect between the various participants. For example, the administration may need a high patient volume to ensure sufficient reimbursements to keep the clinic operating; clinicians, however, may find the required volume overwhelming because it severely limits the amount of time they can spend with each patient. Intake staff and clinicians may get frustrated with patients who repeatedly return to the clinic with the same issues, clearly not having followed the treatment recommended during the last visit. At the same time, patients do not understand why physicians expect them to be able to follow-through on medication regimens when the community does not have a pharmacy. Moreover, patients with mobility challenges are not always able to drive to the closest pharmacy to pick up their prescriptions.

The data produced through participant observation are not as neat and tidy as those produced through IDIs or focus groups. Although the field researcher may be able to conduct the occasional audio-recorded interview and have it transcribed, much of the resulting data will be in the form of comprehensive observation notes written by the researcher on a daily basis. Notes should include some obvious things, such as observations made

QUESTION DOMAINS	Clinic Administrators	Clinicians	Intake Staff	Community Residents
<b>CLINIC ENVIRONMENT</b>				
Resource availability, e.g., medical supplies, space, equipment	X	X	X	
Staffing, e.g., staff-to-patient ratios, tenure/turnover	X	X	X	
Clinic atmosphere, e.g., cleanliness, welcoming, noise levels	X	X	X	X
<b>EXPECTATIONS</b>				
Patient volume	X	X	X	
Length of appointments		X		X
Time spent in waiting room			X	X
Appointment outcomes, e.g., diagnosis, treatment		X		X
Treatment adherence		X		X
<b>COMMUNITY CONTEXT</b>				
Resources, e.g., pharmacies, public transportation, other clinics			X	X
History of clinic in the community	X	X	X	X

**Fig. 4** Question domains

by the researcher while sitting in the waiting room: What were interactions like between patients and intake staff? Did the participants seem to be polite with each other or was tension evident? And what was the evidence for either of these observations? How long were patients sitting in the waiting room? What did it look, feel, and smell like while sitting there, i.e., did the researcher find it to be a welcoming environment or not so much? Why? Notes should also be recorded of any ad hoc interviews, whether in the clinic or in the community. Although it likely will not be possible to write verbatim notes while talking with people during these spontaneous

encounters, the researcher should write up as complete a recounting of the conversation as memory allows and as soon as possible after the interview. Finally, the researcher should include her own thoughts and feelings in the observation notes. Perhaps she finds the clinic staff insufferable, believing them to be rude to the patients. Conversely, perhaps she finds the patients themselves unpleasant, believing them to be demanding too much from harried physicians. Regardless, it is important that the researcher keep in mind the biases she brings to the work as well as how those biases can easily color her interpretation of the data. Realistically, it is

highly unlikely that one clinic managed to hire all of the unbearable doctors, nurses, physicians' assistants, and administrative staff in the area. The researchers must then ask herself, what might be the structural contributors to the staffs' bad behavior? Are they overworked? Is the pay lower than other similar positions in the area? Do they feel like they are unable to make a positive difference in their patients' lives? The researcher may not like – indeed, *should* not feel compelled to like – the individuals with whom she is working. But it is critical to acknowledge those feelings and move beyond them so that systemic challenges – and thus possible solutions to the problems – can be identified.

### **To Record or Not to Record?**

Researchers new to qualitative methods often express discomfort about using an audio recorder during an interview or even a focus group discussion. Particularly when interviewing people about sensitive subjects (e.g., illness, sexuality), the recorder can seem like a monstrous intrusion on the interviewee's private experiences. Nevertheless, recording is the best way to create an accurate record of what was said during the interview and thus ensure that the analysis is based not on secondary data (e.g., the interviewer's notes and remembrances), but on the primary results of the data collection effort (e.g., the recording itself or interview transcripts). Edward Ives *The Tape Recorded Interview* (Ives 1995) is a particularly useful guide for researchers, but the following brief tips may facilitate the reader's use of an audio recording device.

### **Discuss the Desire to Record Early in the Process**

Except on holidays and birthdays, many people do not care to be surprised. If the study plan is to record the interviews, respondents should be told this at the recruitment stage of the project: "I will be conducting an approximately one-hour interview that, with your permission, I would like to audio record." If the recording is optional, this

gives the respondent some time to consider if he or she is okay with being recorded. If the recording is not optional (e.g., the funder/client may stipulate in a contract that focus groups are to be recorded), this allows potential participants the opportunity to opt out early if they do not wish to be recorded.

### **Allow Interviewees or Participants to State Things off the Record**

Interviews can be very cathartic at times, leading respondents to get something off their chests that they then wish they hadn't. The researcher should let participants know that if they end up saying something they want to have expunged, it will be deleted from the recording, any notes about it will be scrubbed, and that information will never make it into the report. Sometimes respondents may say, "I need to say something, but it has to be off the record." The interviewer should TURN off the audio recorder, let them say what they need to say, and then ask permission to turn the recorder back on. Interviewees can be much more comfortable knowing they have some editorial control over what they say.

### **Let Participants Create a Pseudonym**

Because study participants' names will never be used in final reports or journal, it makes no difference to the researcher whether they use their real name when being interviewed or not. But some individuals feel more comfortable with the added layer of anonymity that a pseudonym can bring. If a topic, particularly in a focus group discussion, is especially sensitive, the researcher should consider offering participants the option of coming up with their own names for purposes of the discussion.

### **Store Audio Files in a Secure Location**

Neither the researcher nor his/her interviewees should feel confident that an audio file on a portable recording device will not be accessed by others. Not only do such devices lack security features, but also they are small and easily misplaced or lost. Study participants should be told

how the researchers will secure their information, including where the file will be maintained and how quickly the file will be deleted from the portable device.

### **In the End, the Recorder Usually Becomes Invisible to the Participant**

Finally, it bears noting that in most instances, the researcher feels more awkward about using the audio recorder than does the interviewee. People agree to share their stories with researchers because they have something to say that they want others to hear. The use of a recording device offers an assurance that the details of their stories will not get lost and that their experiences will be faithfully recounted.

---

## **Data Analysis**

Health researchers new to qualitative methods may find themselves immediately overwhelmed by the volume of data generated from in-depth interviews and focus groups. It is not unreasonable, for example, to expect a 1-hour interview to result in a 20–25 page transcript; thus, even a project with only 20 interviews can leave a researcher awash in 400–500 pages of text. Throw in a few focus groups and notes from field observations and the numbers increase dramatically. Perhaps due to this inordinate volume of data, qualitative analysis is often thought of as one of the most mysterious aspects of qualitative research: Work colleagues have called it “magic,” while others have referred to it as “art.” Really, it is neither. Qualitative analysis is the process of systematically reading through one’s data (and re-reading it, numerous times) looking for details and patterns in respondents’ narratives that address the study’s research questions. There are numerous books and articles that offer excellent guidance on both philosophical and logistical aspects of qualitative analysis (Boeije 2010; Bernard and Ryan 2010; Roller and Lavrakas 2015; Thorne et al. 2004); thus, no attempt will be made here to redo what has already been done exceedingly well. Instead, this article provides the reader

with a description of the fundamentals of the analytic process, more details of which can be found in the previously cited references.

### **Simplifying the Data**

The prospect of reading through several hundred pages of text multiple times to find answers to one’s research questions is a daunting prospect indeed. Thus, the analyst’s first goal must be to distill that indistinct mass of narrative into smaller, like units for further analysis, a simplification process that Miles and Huberman (1994) called “data reduction.” In many instances, particularly in applied health research where the objective is to find answers to very specific questions, the data can be distilled on the basis of predetermined categories or themes that are often embedded in the very questions asked of the respondents.

### **Deductive Simplification**

Using the hypothetical Sjogren’s study as an example, IDI probes and focus group questions asked participants to describe the effects of their illness on three dimensions of their lives: their professional, home, and social lives. Thus, the analyst’s first step towards simplifying the data might involve identifying those sections in each transcript where the interviewee described how her illness had impacted her work, home, and/or social activities. Those descriptions may have come in response to a direct question from the interviewer or may have emerged spontaneously during the interviewee’s recounting of her experiences living with the illness. Regardless, identifying those sections of the transcripts that deal with each of these dimensions means the hundreds of pages of text have now been separated into at least *four* “piles” of text: that having to do with the impacts of the illness on interviewees’ (1) work lives, (2) home lives, (3) social lives, and (4) text having to do with everything else. “Everything else” (4) may be further simplified by identifying those sections of narrative where interviewees answered other interviewer questions, such as

(4.a.) effective coping strategies, (4.b.) strategies that did not work so well for them, (4.c.) resources that they wish they had available to them, and (4.d.) text still not yet categorized. As the reader can see, in this example, simply using the concepts covered in the interview guide (which, not coincidentally, parallel the research questions), the analyst can readily parse hundreds of pages of data into smaller, more manageable “units” of data for analysis.

Distillation by mapping extant categories onto the data, essentially a deductive approach, is not so much “analysis” as it is a necessary precursor to the analytic process. That is, there is nothing particularly analytic about locating all of the transcript sections in which interviewees describe how Sjogren’s has affected their social lives. However, it is only by reading through all of this similar text that the analyst can then begin to discern patterns in interviewees’ descriptions – i.e., analyze – the ways in which this invisible disease leaves its social mark. Indeed, the analyst ultimately may find at least two threads in these narratives: those cases in which interviewees were no longer able to participate in their peer group’s activities and their social lives collapsed and those, perhaps fewer, instances in which interviewees described a strengthening of their core social relationships. This bifurcate finding may lead the analyst down a further analytic path as he or she endeavors to determine the factors that contribute to any individual experiencing one social trajectory or the other. In sum, deductive data simplification does not preclude inductive (see below) data analysis.

### **Inductive Simplification**

What if the research questions are not nearly so clear-cut as the ones proposed in this article? What if, instead of wondering how Sjogren’s affects the work, home, and social dimensions of people’s lives, the research goal is simply to capture the broad experience of living with Sjogren’s? Instead of asking interviewees to describe what it is like “*living with an invisible illness*” (which, as noted previously, necessarily implies asking how the person with the illness interfaces with others), the researcher asks,

“*Please tell me what it has been like for you living with Sjogren’s?*” Your probes may be less directive than ours, asking, “*How does the illness affect you day-to-day?*” rather than, “*How has living with this illness affected your social life?*” In this case, simplifying the data requires reliance on an inductive analytic approach, in which the meaningful categories emerge from the reading of the analyst’s data rather than being predetermined by the research questions. Inductive simplification may mean the analyst needs to read all of the interviewees’ transcripts, at least once, possibly twice, before he or she can *begin* to find recurring themes in their narratives. Many may describe impacts of the illness on their work and social lives and, as a result, an initial cut in the data is created along these two dimensions. But the analyst may also find that interviewees often describe being disappointed in themselves when they find they are no longer able to do not only high-energy activities, such as hiking or playing tennis, but even simple tasks to which they once gave not a moment’s thought. Carrying a basket of laundry, turning a wrench to release the oil drain bolt on the car, even walking up a flight of stairs – once effortless activities have become onerous, if not impossible, to perform. Interviewees describe a loss of self-efficacy that to them is as disturbing, if not more so, than the loss of their social lives. After reading several similar descriptions, the researcher might create a provisional category, perhaps called “Sense of Self,” and begin to look for additional text that recounts similar feelings and experiences on the part of the narrator. As with deductive simplification, once these subsets of data are defined, the analysts can dive further into each, looking for additional similarities and differences in how individuals describe these like experiences.

Unlike a purely deductive approach, where data reduction is a precursor to analysis, data analysis is part and parcel of inductive data simplification. Meaningful cuts in the data are not predetermined by the research questions or interview guides, but must be determined by the researcher through multiple careful readings of the data and their subsequent interpretation.

Nevertheless, data reduction is still only the first step in the process, whether it begins through induction or deduction. Subsequent analytic efforts will explore the data for additional patterns, such as themes or concepts that are shared by all respondents, or multiple different perspectives on the same issue (e.g., differential impacts on one's social life). Whenever possible, finding a potential explanation for such differences is the next step in the analytic process. The previous social impacts example described hypothetical interviewees who said their social worlds came undone as a result of their illness and others who said they grew even closer to their core group of friends. The analyst might first look for demographic differences in each of these groups as a possible way to account for the different effects: perhaps the latter interviewees are significantly older than the former or perhaps the first group are single while the second group are married. The analyst might also look to each speaker's narrative for additional clues that could account for the differences: words such as "outgoing," "active," "gregarious," and "social" may characterize the first group's narratives, while such terms are largely absent among the second group.

### **A Note on Data Coding**

Over the last 20 years, qualitative researchers have increasingly incorporated software into their approach to data analysis. Sophisticated programs such as NVivo, Atlas.ti, Dedoose, and others are allowing researchers to analyze more data, more quickly, and in a way that is far more transparent than the old-fashioned paper-and-colored-markers approach. Importantly, though, the fundamentals of qualitative data analysis do not change simply because a computer is involved. The analyst must still read through all of the data; reduce the reams of information into manageable, "like" units through either deductive or inductive simplification processes; read the like units to identify narrative themes that are shared by the interviewees or that diverge; and, when possible, seek an explanation to account for those differences. The software does, however, make several of these steps easier.

First, because the analyst is using electronic codes for data reduction rather than colored markers, extracting similarly coded text can be as quick and easy as the click of a button or the writing of a simple program (data query). The analyst thus can spend less time looking for text and more time reading it to see if there are important nuances in interviewees' narratives. Second, automation allows studies to collect and analyze much larger volumes of data than would be feasible if the work were being done by hand. In 2010, for example, the U.S. Department of Defense supported a Comprehensive Review Working Group (CRWG) to examine active-duty and reserve service members' views about the potential impact on unit cohesion, morale, and readiness if Don't Ask, Don't Tell (the 1993 law barring openly gay individuals from serving in the military) were repealed. In addition to conducting the largest-ever survey of service members and military spouses, the effort included the analysis of hundreds of focus group transcripts, two thousand open-ended survey comments, and literally thousands of comments sent to a DOD inbox. All data collection and analysis took place within a ten-month timeframe, a feat that was possible only with the support of an excellent qualitative data analysis program (Robins and Eisen 2017).

Third, these programs allow the users to link respondent characteristics (e.g., demographic data, geographic location, organizational affiliation) to interview documents such that the analyst can quickly examine the data for any patterns by respondent type. In the Don't Ask, Don't Tell study, for example, the team was able to explore respondent sentiment regarding repeal (positive, negative, does not care) by respondent gender, service (e.g., Marine Corp, Army), officer or enlisted status, or pay grade, or any combination of those characteristics (e.g., female Army officers compared to male Army officers). This type of analysis can possibly be done without a computer, but it would be tedious and time-consuming, at best.

Finally, and importantly, qualitative analysis software supports the development of an "audit trail," a time- and date-stamped description of the decisions and actions of the analytic team.

This is important documentation for clients, some of whom may be uncertain about the rigor with which the qualitative analysis is being done. It is also an invaluable check for the analysts, ensuring that both new and seasoned researchers are able to support both their decisions and their findings with data.

---

## Summary

Qualitative health researchers have helped to shed light on how both patients and clinicians understand states of health, disease, and what constitutes appropriate treatment. The insights generated from their work have contributed to reduced disease transmission, understanding of patients' lived experiences, improved communication between clinicians and the people they treat, and better patient health care experiences. There has always been the potential for misunderstandings to emerge between patients and clinicians, who have very different funds of knowledge and assumptions about the world. The rise in globalization only exacerbates the potential for conflict in the midst of a medical crisis, resulting in increased disease burden on patients and the systems trying to treat them. Health researchers interested in contributing to the development of constructive dialogues in the clinical encounter may well find that qualitative research methods are the right tool for the job.

---

## References

- Alvesson M, Skoldberg K. *Reflexive methodology: new vistas in qualitative research*. Los Angeles: Sage; 2009.
- Bernard HR, Ryan GW. *Analyzing qualitative data: systematic approaches*. Los Angeles: Sage; 2010.
- Boeije H. *Analysis in qualitative research*. Los Angeles: Sage; 2010.
- Cassell J. *Expected miracles: surgeons at work*. Philadelphia: Temple University Press; 1991.
- Chamberlain J. *On our own: patient-controlled alternatives to the mental health system*. New York: Haworth Press; 1978.
- Donohue PJ, Siegel ME. *Sick and tired of feeling sick and tired: living with invisible chronic illness*. New York: WW Norton & Company; 2000.
- Ebola Anthropology Response Platform. <http://www.ebola-anthropology.net/>
- Estroff SE. *Making it crazy: an ethnography of psychiatric clients in an American community*. Berkeley: University of California Press; 1981.
- Farmer P. *AIDS and accusation: Haiti and the geography of blame*. Berkeley: University of California Press; 1993.
- Forum: Qualitative Social Research. Accessible at: <http://www.qualitative-research.net/index.php/fqs/index>.
- Geertz C. *The interpretation of cultures*. New York: Basic Books; 1973.
- Hopper K. Qualitative and quantitative research: two cultures. *Psychiatr Serv*. 2008;59(7):711.
- International Congress of Qualitative Inquiry. Information available at: <http://icqi.org/qualitative-health-townhall-meeting/>
- International Institute for Qualitative Methodology. Information available at: <https://www.ualberta.ca/international-institute-for-qualitative-methodology>
- International Journal of Qualitative Methods. Accessible at: <https://us.sagepub.com/en-us/nam/international-journal-of-qualitative-methods/journal202499#description>
- Ives ED. *The tape-recorded interview: a manual for fieldworkers in folklore and oral history*. Knoxville: University of Tennessee Press; 1995.
- Kuhn TS. *The structure of scientific revolutions*. Chicago: University of Chicago Press; 1962.
- Martin E. *The woman in the body: a cultural analysis of reproduction*. Boston: Beacon Press; 1987.
- Miles MB, Huberman AM. *Qualitative data analysis*. 2nd ed. Newbury Park: Sage; 1994.
- Morgan DL, Krueger RA. *The focus group kit*. Los Angeles: Sage; 1997.
- Morse JM, editor. *Recent advances in cross-cultural nursing*. Edinburgh: Churchill Livingstone; 1988.
- Morse JM, editor. *Cross-cultural nursing: anthropological approaches to nursing research*. New York: Gordon & Breach; 1989a.
- Morse JM, editor. *Qualitative nursing research: a contemporary dialogue*. Rockville: Aspen Press. Rev ed., Newbury Park: Sage; 1989b.
- Morse, JM. Getting Started: Labels, Camps, and Teams. *Qualitative Health Research*, Volume 1991;1(1):3–5.
- Morse JM, Field PA. *Nursing research: the application of qualitative approaches*. Cheltenham: Stanley Thornes Ltd; 1996.
- Murchison JM. *Ethnography essentials: designing, conducting, and presenting your research*. San Francisco: Josey-Bass; 2010.
- Ponterotto JG. Qualitative Research in counseling psychology: a primer on research paradigms and philosophy of science. *J Couns Psychol*. 2005;52(2):126–36.
- Popper K. *The logic of scientific discovery*. London: Routledge; 1934.

- Qualitative Health Research. Accessible at <https://us.sagepub.com/en-us/nam/journal/qualitative-health-research#description>
- Ramin, B.M. & McMichael, A.J. Climate Change and Health in Sub-Saharan Africa: A Case-Based Perspective. *EcoHealth*. 2009;6:52. <https://doi.org/10.1007/s10393-009-0222-4>
- Robins CS, Eisen K. Strategies for the effective use of NVivo in a Large-Scale Study: qualitative analysis and the repeal of don't ask, don't tell. *Qual Inq*. Volume 2017;23(10):768–778.
- Roller MR, Lavrakas PJ. *Applied qualitative research design: a Total quality framework*. New York: Guilford Press; 2015.
- Sokal AD. Transgressing the boundaries: towards a transformative hermeneutics of quantum gravity. *Social Text*. 1996;#46/47:217–52.
- Thorne S, Reimer Kirkham S, O'Flynn-Magee K. The analytic challenge in interpretive description. *Int J Qual Methods*. 2004;3(1):1–11.



---

**Part III**

**Health Care Systems and Policies**



Irene Papanicolas and Peter C. Smith

## Contents

<b>Introduction</b> .....	755
<b>What Is Performance Measurement for?</b> .....	756
<b>Defining and Measuring Performance</b> .....	758
Defining the Unit of Analysis .....	758
Defining Key Performance Objectives .....	759
<b>Methodological Issues</b> .....	760
<b>Conclusions</b> .....	764
<b>References</b> .....	766

## Abstract

The provision of performance information can play a key role in health system evaluation and performance improvement. In this chapter we review the key debates around the conceptualisation of the health system and the domains of performance commonly measured. The chapter outlines the key challenges to data measurement such as data availability

and methodological concerns. Finally the chapter considers issues related to data presentation. The chapter concludes by summarising progress made in performance assessment and outlining new directions for future work.

## Introduction

The provision of relevant, accurate, and timely performance information can play a pivotal role in ensuring the health system is able to deliver effective and efficient health services. Through its capacity to secure accountability in the health system, to determine appropriate treatment paths for patients, and to plan for future service patterns and structures, information can be used to identify and implement potential improvements in service delivery. Performance information thus plays an

---

I. Papanicolas (✉)  
The London School of Economics and Political Science,  
London, UK

Harvard T.H. Chan School of Public Health, Cambridge,  
MA, USA  
e-mail: [i.n.papanicolas@lse.ac.uk](mailto:i.n.papanicolas@lse.ac.uk)

P. C. Smith  
Imperial College, London, UK  
University of York, York, UK  
e-mail: [peter.smith@imperial.ac.uk](mailto:peter.smith@imperial.ac.uk)

important role not only as an intrinsic element of the health system but also as a key component of a great deal of health services research. Underlying all of these efforts is the role it plays in enhancing the decisions that various stakeholders, such as patients, clinicians, managers, governments, and citizens, take in identifying performance improvements and steering the health system toward better outcomes overall.

The use of performance measurement for health system improvement has been strongly advocated by pioneers in the field such as Florence Nightingale and Ernest Codman since the late 1800s. Yet only in the past decades have health systems seen a substantial growth in health system performance measurement and reporting to this end. The new growth in performance information and its use for improvement have been the result of multiple factors on both the demand and supply side. On the demand side, increasing demands of accountability and transparency by the public have created a growing culture requiring proofs and accountability. While on the supply side, great advances in technology have made it possible to develop and store increasing amounts of information, allowing stakeholders instant access to large volumes of data (Smith et al. 2009).

While these factors give major impetus to the use of information for performance improvement, a large number of key debates and barriers remain. Health systems are still experimenting with performance measurement, and large steps are still needed to coordinate efforts and identify what works. The policy agenda has moved from concerns with whether data collection should be undertaken, and in what areas, to concerns of how to summarize and present data and how to coordinate key interests in order to develop firmly based policies and tangible improvements.

This chapter seeks to summarize some of the main issues emerging in the performance measurement debate. The chapter will begin by considering what the key aims of performance measurement are and what performance measurement seeks to evaluate. This section will draw upon some of the debates which have arisen as a

result of the different ways in which the health system and its objectives are conceptualized by different stakeholders and frameworks. The chapter will then consider some of the methodological considerations which have arisen the use and evaluation of performance information. And finally it will conclude by discussing the major challenges found in presenting and using performance measures but also by considering the presenting key lessons and future priorities.

---

## **What Is Performance Measurement for?**

Health systems are complex entities with many different stakeholders, including patients, health-care professionals, health-care providers, purchaser organizations, regulators, the government, and the broader citizenry. As outlined by an early report in the area of health information (Rigby et al. 1999), information can be identified as having five key roles in health care (Table 1) relating to the different accountability relationships that exist between the many stakeholders in the system. Through the collection and use of information for decision-making in health systems, stakeholders can hold each other to account, thereby facilitating improvements in effectiveness and efficiency. Thus, the fundamental role of performance measurement is to help enable accountability relationships to function, by enabling stakeholders to make informed decisions. It is therefore noteworthy that, if the accountability relationships are to function properly, no system of performance information should be viewed in isolation from the broader system design within which the measurement is embedded.

Each of the key roles of information described in Table 1 relates to a separate function or role of the health-care system, such as providing patient care or planning and developing health services. Each entails different information needs in terms of the nature of information, the level of detail and timeliness, and the level of aggregation required, in order to function effectively. For example, in choosing which provider to use, a patient may

**Table 1** The role and significance of information in health care

Role of health care	Type of information needed
Patient care	<p>Information to enable patients to make decisions among providers or treatment options, such as information on:</p> <ul style="list-style-type: none"> <li>Location and quality of nearby emergency health services</li> <li>Quality of options for elective care facilities and physicians</li> <li>Cost of different services/insurance plans</li> <li>Reviews of providers by families, friends, or third parties</li> <li>Information on symptoms and treatment options</li> </ul> <p>Information for patients about how to navigate the health system, such as information on:</p> <ul style="list-style-type: none"> <li>Which services they are entitled to</li> </ul> <p>Information for physicians on patient's health-care needs/problems and clinical history, such as information on:</p> <ul style="list-style-type: none"> <li>Patient diagnosis</li> <li>Past medical history and family medical history</li> <li>Patient lifestyle factors</li> </ul>
Professional practice	<p>Information to compare relative performance to other professionals and communicate with one another, such as:</p> <ul style="list-style-type: none"> <li>Patient information including clinical processes and outcomes of care for all patients with a similar diagnosis or procedure (i.e., registry information)</li> </ul> <p>Information to enable self-review, such as information on:</p> <ul style="list-style-type: none"> <li>Processes of care, clinical processes and outcomes for patients treated as compared to peers, national averages, or best practice</li> <li>Errors and adverse events for patients treated</li> <li>Patient experience for patients treated</li> </ul> <p>Information to defend actions taken when necessary, such as information on:</p> <ul style="list-style-type: none"> <li>Historical patient files</li> <li>Best practice guidelines</li> </ul> <p>Information to provide the foundation for evidence-based practice, such as information from:</p> <ul style="list-style-type: none"> <li>Clinical trials</li> <li>Observational studies</li> </ul>
Management	<p>Information to enable operational management, such as Information on organizational costs and quality</p> <p>Information to optimize resource management and deployment, such as Information on resource availability and needs</p> <p>Information to enable service improvement, such as Information on processes and outcomes of care</p>
Service development	<p>Information to evaluate treatments and services, such as information on:</p> <ul style="list-style-type: none"> <li>Comparative effectiveness of different treatments and services</li> <li>Comparative costs of different treatments and services</li> </ul> <p>Measuring outcomes and thus developing knowledge, such as information from:</p> <ul style="list-style-type: none"> <li>Variation in regional and national population health</li> </ul> <p>Information to plan for future service patterns and structures:</p> <ul style="list-style-type: none"> <li>Variation in supply and demand of health services</li> <li>Projections of changes in supply and demand of health services</li> </ul>
Policy development	<p>Information to provide intelligence for policy formulation, such as information on:</p> <ul style="list-style-type: none"> <li>Variation in outcomes, clinical processes, and patient experience across different geographical regions or organizations</li> <li>Costs and effects of different medical interventions and treatments</li> <li>Prevalence of health needs in the population</li> <li>Aggregate information on preventable or treatable mortality and morbidity</li> </ul> <p>Information to enable inter-sectoral action, such as information on:</p> <ul style="list-style-type: none"> <li>Prevalence of particular lifestyle choices or behaviors</li> <li>Evidence of association between particular lifestyle choices or behaviors to health outcomes</li> </ul>

need detailed comparative data on health outcomes for a specific intervention. In contrast, in holding a government to account, and deciding for whom to vote, a citizen may seek out highly aggregate summaries and trends. Many intermediate needs arise. In order to contribute to operational management, more aggregate information and detailed assurance on safety aspects may be necessary. This variety of uses highlights greatly different information needs in terms of the nature, detail, timeliness, and level of aggregation information users require. A fundamental challenge in performance measurement is to create information systems that are able to cater efficiently for these diverse needs, both in terms of data collection and data presentation and interpretation.

In practice the development of performance measurement has rarely been pursued with a clear picture of who the information users are or what their information needs might be. Instead performance measurement systems have often developed opportunistically, usually seeking to inform a variety of users and presenting a wide range of data in the hope that some of the information collected will be useful to various parties. Yet, given the diverse information needs of the different stakeholders in health systems, it is unlikely that a single method of performance reporting will be useful for everybody. Instead data sources should be designed and exploited with the needs of different users clearly in mind. This may often involve using data from the same sources in different forms. A major challenge for health systems is therefore to develop more nuanced ways of collecting and presenting performance measures for the different stakeholders without imposing a huge burden of new data collection and analysis.

The starting point of most performance assessments is the creation of a conceptual framework on which to base the collection of information and to use as a heuristic for the understanding of the entity being assessed (whether it be the entire health system, a provider organization, or an individual practitioner). A theoretical framework is necessary to help define a set of measures that reflect key organizational objectives and in turn

allow for an appropriate assessment of its performance. In the past decade, numerous conceptual frameworks have been created for health system performance assessment at the international level and national level. In many cases countries have developed more than one performance framework, reflecting variations in national and/or local priorities or the performance of different areas of the health system. While existing frameworks have varied purposes, they all aim to provide a better understanding as to what constitutes “good” performance by identifying the entity whose performance is being assessed, its key objectives, and the underlying structures and factors that drive performance (Papanicolas and Smith 2014).

---

## Defining and Measuring Performance

The role of performance measurement is to measure, analyze, and report the extent to which the health system is achieving its key objectives. In order to assess performance successfully, it is important to be able to unambiguously define the entity being assessed (whether this be the health system, an organization, or an individual), as well as the key performance objectives of this entity.

## Defining the Unit of Analysis

One of the main areas of debate across this field of study involves clearly defining the unit under scrutiny, whatever the level of analysis. At the system level, differences exist between national and international stakeholders in determining where the health system boundaries lie and what responsibilities lie within the jurisdiction of the health system. In particular, there is no consensus as to whether a definition of the “health system” should encompass the wider determinants of health outcomes and whether it should include activities which impact health outcomes such as public health, health promotion, and targeting social determinants of health (Papanicolas et al.

2013). There can be no right answer to this question, as institutional arrangements differ between countries, and there are arguments for promoting the use of both wider and narrower boundaries depending on the purpose of the analysis. However lack of consensus on this issue makes international comparison of performance assessment difficult (Papanicolas and Smith 2014).

At the organizational level, boundaries between different sectors of care such as primary care, hospital care, and long-term care are rarely clearly defined. Part of the difficulty in producing a coherent definition of these services and organizations emerges from differences in remits within and across systems. For example, something like rehabilitation after surgery may be provided in the hospital sector in some systems and in long-term care facilities in others. However it would be misleading to compare the performance of the two hospital sectors without considering this difference. Whatever the chosen definition, in any evaluation of performance, the crucial objective is to ensure that the achievements being assessed accurately represent the contribution attributable to the entities under scrutiny. For example, in the performance assessment of a hospital, it is essential to isolate the contribution of hospital care to the attainment of performance objectives (e.g., health improvement) and where necessary to adjust for any contribution of other activities such as primary care provision, public health, and contextual factors such as the economic, political, and demographic environment. It is thus necessary for one to consider what range of services falls within the accountability of the hospital – and how the contribution of these services can be assessed controlling for other factors external to the responsibilities of the hospital.

## Defining Key Performance Objectives

Section “[What Is Performance Measurement for?](#)” above outlines the main objectives of performance assessment and the potential that information holds to ensure that the accountability relationships within the health system can operate in a manner that enables the health-care system

to achieve its overarching goals. Thus, to be able to assess the performance of health system, it is important to articulate clearly its key objectives. There exists a substantial literature which outlines the main goals of the health system (Aday et al. 2004; Atun 2008; Commonwealth Fund 2006; Hurst and Jee-Hughes 2001; IHP 2008; Jee and Or 1999; Kelley and Hurst 2006; Klassen et al. 2009; Murray and Frenk 2000; Roberts et al. 2008; Sicotte et al. 1998), and while there are differences related to the definitions of what particular objectives entail, there seems to be relative consensus on the objectives themselves. These objectives can usually be summarized under a limited number of headings broadly summarized as:

- The health conferred on citizens by the health system
- The extent to which the health system is equitable
- The extent to which patients and their families are protected from the direct costs of needed health care
- The patient experience offered by the health system
- The efficiency and productivity with which health resources are utilized

The fundamental goal of all health systems is to improve the health of patients and the general public. However, aside from being concerned with the absolute level of health improvement in each system, a number of performance frameworks highlight the importance of distributional (or equity) issues, expressed in terms of inequity in health outcomes. Most health systems today are concerned not only with the ability of the health systems to improve health but to do so across all groups in the population. Related to this concept is the issue of equity of access to health care or equity of access to and financing of health care; most health systems also seek to protect citizens from the impoverishment that can arise from health-care expenditure and to ensure all groups of the population have access to at least a basic package of health services (Papanicolas et al. 2013).

In 2000, the *World Health Report* “Health Systems: Improving Performance” highlighted health system “responsiveness” as an intrinsic goal of the health system (Murray and Frenk 2000; WHO 2000). The WHO definition refers to “responsiveness to the legitimate expectations of the population for their interaction with the health system,” and it captures dimensions unrelated to health outcomes such as dignity, communications, autonomy, prompt services, access to social support during care, quality of basic services, and choice of provider. Often this goal is also referred to as patient or population satisfaction or patient experience, yet while there is overlap across these three concepts, they do not all encompass the same characteristics but almost always relate to the underlying expectations of patients and the population. As with health outcomes, it is not only the absolute level of responsiveness/satisfaction or good experience in a system that is of interest but how this is distributed among different groups in the population.

Finally, efficiency and productivity, or the extent to which health resources are used to produce valued outcomes, is also a key objective of health systems. Reflecting the wide range of potential perspectives, economists and policy makers have adopted different conceptualizations of efficiency when analyzing different levels of the health system. Systems-level efficiency is concerned with understanding how well a specific system is using the resources at its disposal to improve health and secure related objectives (Papanicolas and Smith forthcoming). At the organizational level, definitions usually refer to the extent to which health service objectives – such as hospital objectives – have been achieved compared to the maximum that could be attained, given the resources available and the external constraints on attainment. While, at the very micro level, efficiency can be related to decisions of individual clinicians on how to distribute health-care resources across treatment options in order to maximize valued outputs. The study of this type of efficiency often takes the form of a systematic analysis of the effects and costs of alternative methods or programs for achieving

the same objective (e.g., improving quality of life, extending years of life lived, or providing services).

As stated above, the overall aim of performance measurement is to measure, analyze, and report the extent to which the health system is achieving its key objectives. However, we have also seen that information requirements necessary to measure performance vary across the key roles of the health system, the different stakeholders, and the different levels of analysis. Table 2 considers some of the key types of measures relating to the objectives discussed above at different units of analysis, in particular relating to (1) the system level, (2) the organizational level, and (3) the individual level. Information at the systems level is aggregated information that allows stakeholders to consider how performance objectives are being met at the population level. This information can be useful for national or regional benchmarking exercises or to gauge overall performance on particular goals or to assess the impact of system-level reforms. Organizational-level performance can be crucial for many of the key roles of the health system, such as allocating resources, patient choice, treatment, and policy evaluation. Finally, information at the individual level can be very important for physicians and managers to ensure that safe and effective services are delivered to patients.

---

## Methodological Issues

The diverse set of users and information needs in a health system call for a wide variety of measurement techniques and indicators. Various approaches toward data collection are needed to assemble the necessary information, such as national surveys, patient surveys, administrative databases, and routinely collected clinical information. The domain of performance being examined will in part determine the most appropriate data collection technique (Table 3). For example, when measuring responsiveness, household or individual surveys are likely to be the best sources of patient’s experiences and perspectives, whereas when looking at specific clinical outcomes,

**Table 2** Measures of key performance objectives at different levels of analysis

Performance objective	Types of measures and their uses
Health improvement	<i>System level:</i> measures of aggregated data on the health of the population (e.g., life expectancy, disability-adjusted life years, avoidable mortality, survival rates)
	<i>Organization level:</i> measures of aggregated data on the contribution to health of particular health sectors or services (e.g., avoidable hospitalizations, hospital standardized mortality rates, emergency readmission rates for different organizations/conditions)
	<i>Individual level:</i> measures of health status/health gain for individuals (e.g., QALY, survival, patient-reported outcome measures)
Patient experience	<i>System level:</i> measures of aggregated data population experiences/satisfaction with the health system (e.g., population satisfaction, population experiences, average waiting times)
	<i>Organization level:</i> measures of aggregated data on satisfaction/experience for particular health sectors or services (e.g., rates of patient satisfaction, aggregated patient experiences, number of patient recommendations)
	<i>Individual level:</i> measures of satisfaction/experience/responsiveness of individuals (e.g., overall physician rating)
Equity and fair financing	<i>System level:</i> measures of the extent to which there is equity in health, access to health care, responsiveness, and financing (e.g., rates of access of the population, indices of equity in health and access, out-of-pocket payments as a % of total health expenditure, catastrophic spending, impoverishing spending)
	<i>Organization level:</i> measures of the extent to which there is equity of access, responsiveness, and financing of particular health-care services (e.g., utilization rates, unmet need of medical care and dental care)
	<i>Individual level:</i> n/a
Efficiency/productivity	<i>System level:</i> the extent to which health system objectives are maximized given existing resources (e.g., ratio of health system outputs to inputs)
	<i>Organization level:</i> the extent to which health sector or health service outputs are maximized given resources available (e.g., unit costs, average length of stay)
	<i>Individual level:</i> identifying the treatment option which yields the maximum effectiveness per unit cost (e.g., QALY/cost)

clinical registries may be a more informative and cost-effective source of information. In practice, although performance measurement efforts have progressed over recent years, many health systems still rely on readily available data as a basis for performance measurement. An important research agenda is to determine where new or revised data collection initiatives would be most valuable.

Regardless of the data sources used, a fundamental issue that arises when seeking to interpret performance data is: What has caused the observed performance and to what practitioners, organizations, or agencies should variations in performance be attributed? The key performance

objectives outlined in section “[Defining and Measuring Performance](#)” are often the product of numerous determinants. An individual’s health status, for example, can be directly influenced in the short term by actors in the health services (e.g., improving medical care), others that require long-term action of actors not directly associated with health services (e.g., environmental policy), and yet others that depend primarily on the actions of individuals and their families (e.g., diet).

Various statistical methods can be used to adjust information for different risk factors, such as differences in resources, case mix, and environmental factors, to make performance more comparable across organizations or practitioners.



**Table 3** Data sources – strengths and weaknesses

Data type	Advantages	Disadvantages
Administrative data	<ul style="list-style-type: none"> <li>Readily available</li> <li>Ease of access</li> <li>Relatively low acquisition costs</li> <li>Clear and comparable data</li> <li>Typically cover large populations</li> <li>Provide a wealth of information on services provided and potential costs</li> </ul>	<ul style="list-style-type: none"> <li>Payment-related incentives may influence data content</li> <li>Structure of system will influence degree of data available</li> <li>Coding of diagnosis may be problematic</li> <li>May not capture crucial clinical parameters</li> <li>Timing of data entry may not be clear</li> </ul>
Survey data	<ul style="list-style-type: none"> <li>No strong incentives for gaming</li> <li>Provides the only source of information on experiences, views, and opinions</li> <li>Subjective measures are often shown to be good measures of objective measures</li> </ul>	<ul style="list-style-type: none"> <li>May be subject to survey bias if response rates are not sufficiently high</li> <li>Responses can be very sensitive to conditioning effects related to survey length or question wording</li> <li>May be sensitive to cultural, ethnic, and even gender bias</li> <li>Longitudinal surveys may be subject to bias related to attrition</li> </ul>
Medical records	<ul style="list-style-type: none"> <li>Provide a rich source of clinical information</li> <li>Track data over time</li> </ul>	<ul style="list-style-type: none"> <li>May contain contradictory information</li> <li>Susceptible to manipulation</li> <li>Requires trained and skilled staff</li> <li>Reports may be variable and not directly comparable</li> </ul>
Clinical registries	<ul style="list-style-type: none"> <li>Provides a rich source of data for large numbers of patients suffering a particular health condition</li> <li>Uniformity in data collection methods and the frequency of data collection</li> <li>Includes important clinical information and patient information</li> </ul>	<ul style="list-style-type: none"> <li>Often limited to particular health conditions</li> <li>Subject to bias in terms of who is included in the registry</li> </ul>

These methods are known as “risk adjustment” techniques. Where variations in performance measures are known to be influenced by factors beyond the control of the entities under scrutiny, it becomes essential to employ methods of risk adjustment when using and comparing indicators to help account for these variations. For example, when measuring hospital outcomes as an indication of quality, it may become crucial to adjust for patient attributes such as their age, comorbidities, or socioeconomic class. Failure to risk-adjust outcome measures before comparing performance may result in drawing misleading conclusions and can have serious implications for policy and quality improvement (Iezzoni 2013). However, many methods of risk adjustment remain highly contested. Therefore, whenever risk adjustment is undertaken, it should be presented in a clear transparent manner together with the final performance data.

Furthermore, when performance assessment is used for health service improvement, it is essential that causality for observed measures is attributed to the correct sources or parties (Terris and Aron 2009).

When collecting and assessing performance information, two types of error should be recognized and controlled for to the extent possible. The first of these is random error, which emerges with no systematic pattern and is always present in quantitative data. Random error can give rise to two types of false inference, commonly known as type 1 errors (false positive) and type 2 errors (false negative). The traditional way of controlling for these errors is to apply statistical tests to data at a high significance level (usually 0.05 or 0.01). Although well understood, this statistical approach is essentially arbitrary and ignores the relative cost of making either type of error. The second type of error is systematic error which may

**Table 4** Usefulness of outcome and process indicators

Type of indicator	Advantages	Disadvantages	Areas best used
Outcome indicators	Stakeholders often find outcome measures more meaningful Directs attention to and focuses health goals on the patient Encourage long-term health promotion strategies Increasing use of PROMs Not easily manipulated	May be ambiguous and difficult to interpret as they are the result of many factors, which are difficult to disentangle Take time to collect and for outcome to materialize Require a large sample size to detect statistically significant effects Can be difficult to measure (i.e., wound infection)	To measure quality of homogenous procedures To measure quality of homogenous diagnosis with strong links between interventions and outcomes To measure quality of interventions done to heterogeneous populations suffering a common condition
Process indicators	Easily measured without major bias or error More sensitive to quality of care Easier to interpret Require a smaller sample size to detect statistically significant effects Can often be observed unobtrusively Provide clear pathways for action Capture aspects of care that are valued by patients aside from outcomes	Often too specific, focusing on a particular intervention or condition May quickly become dated as models of care and technology develop May have little value for patients unless they understand how they relate to outcomes May be easily manipulated	To measure quality of care, especially for treatments where technical skill is relatively unimportant to measure quality of care of the homogenous conditions in different settings

Source: Adapted from Davies (2005) and Mant (2001)

occur if there have been errors in measurement approaches, such as flawed sampling methods. Systematic errors of this sort will lead to erroneous conclusions concerning a variable's true value. In order to avoid systematic errors, it is critical that data collection methods are carefully designed, implemented, and audited.

Traditionally, performance measures have been classified as structure, outcome, or process measures. Outcome reflects the eventual objective of the system. However, certain process measures may be more realistic indicators of quality if they are known to be associated with good future outcomes. Different types of indicators will be appropriate depending on the setting. For example, outcome measures such as mortality may be more useful when looking at population health,

while process measures will be more indicative of the quality of care for a specific procedure. It is critical that designers of performance measurement schemes are aware of the advantages and disadvantages of different types of indicators when using them to assess performance. Table 4 summarizes the main advantages and disadvantages of using outcome and process indicators and the areas of performance measurement where they are most useful.

Experience indicates that a balanced approach with multiple aggregated and disaggregated indicators is most desirable to cater for the information needs of different stakeholders and to allow more informed policy decisions. For this reason, composite indicators – indicators which combine separate performance indicators into a single

index or measure – are often used to rank or compare the performance of different practitioners, organizations, or systems by providing a “bigger picture” and offering a more rounded view of performance (Goddard and Jacobs 2009). The main virtue of composite indicators is that they capture attention in a way that a mass of separate indicators cannot. However, critics of composite measures argue that reducing the measurement of objectives, or entire dimensions, to one indicator runs the risk of being too simplistic and masks many of the variations in performance that should be studied.

Indeed, if composite indicators are not carefully designed, they may be misleading and could lead to serious failings if used for health system policy making or planning (Smith 2002). One of the main challenges encountered in the creation of composite indicators is selecting which measures to include in the indicator and with what weights, particularly in areas where there is little choice of data, and questionable sources may be used for some components of the indicator. Thus, when using composite indicators, it is prudent to give a full description of all the information that is summarized in the indicator, to provide an insight into the performance of each component and help pinpoint the reasons for variation. In addition, the composite and its inputs should be presented with proper uncertainty measures, which may be more informative than measures of central tendency (Jacobs et al. 2005; Naylor et al. 2002).

It is important to note that rapid progress is being made in all areas of health system data collection, including areas such as the design, collection, governance, linkage, and dissemination of data. These developments have the potential to add further value to the existing data collected, particularly by extending the application of what is already available and by collecting new data in a more coordinated, timely, and reliable fashion. Data linkage is allowing researchers and policy makers to create a more complete record of all factors that contribute to health, facilitating the creation of less noisy indicators and a more holistic picture of health determinants. The adoption of IT systems in health-care

organizations and the systematization of classifications within and across countries (using tools such as diagnostic resource groupings and/or ICD codes) also allow more robust comparisons across organizations. Finally, another very large area of development is that of information and communication technologies (ICT), often described within the EU context in particular as “e-health,” which has the potential to improve greatly the scope, volume, and quality of performance data.

---

## Conclusions

The ultimate aim of performance measurement is to help hold the various agents to account, given the organization and structure of the health system, by enabling these stakeholders to make informed decisions. In order for these accountability relationships to function properly, no performance information system should be viewed outside its broader context within which the measurement is fixed. Where possible the performance measurement should provide information for all the relevant accountability relationships present in the health system.

If undertaken carefully, performance measurement can offer a powerful resource for identifying weaknesses and suggesting relevant reforms. The progress that has been achieved is impressive, both in the scope of areas for which data is now available and in the degree to which comparability across different entities has been improved. Table 5 outlines the key developments that have been made across some health service performance domains and also highlights some of the main challenges that remain.

The data collection techniques and methodological tools used for performance measurement have developed considerably in the past decade. The debates raised by the WHO 2000 report in particular have spurred the development of datasets, which are updated regularly with new surveys, process indicators, or outcome indicators in order to best operationalize theoretical concepts. Considerable progress has also been made in the measurement of patient-reported outcomes, patient satisfaction measures, and patient

**Table 5** Challenges and developments for the measurement of health service performance domains

Performance domain	Challenges for measurement	Developments in measurement
Health improvement	<p>Many aggregate measures fail to distinguish the contribution of the health system</p> <p>Problems of comparability among over time, reflecting changes in and differences between coding rules</p> <p>Large gaps in availability of evidence on the effectiveness of treatments reducing mortality</p> <p>Limited set of dimensions captured by outcome measures with a marked lack of measures on disabilities or discomfort</p> <p>Lack of available, good-quality, and comparative data at the patient level</p>	<p>The development of electronic health records (EHRs) provides more complete information on all factors influencing outcomes</p> <p>Increase in registry data, which identifies individual patients and traces them through the care process</p> <p>Increase in measures of outcomes that are not defined in terms of cure, which are important for the measurement of chronic disease and long-term care</p>
Equity	<p>Lack of existing datasets which provide a longitudinal perspective</p> <p>Limited evidence has been recorded on how sensitive inequalities are to the inclusion of environmental effects</p> <p>Limited understanding of the factors explaining the health production process and sources of inequalities, including the role of mental conditions along with cognitive biases in measuring self-reported health</p> <p>Inadequate identification of what stands behind measures of socioeconomic position, namely, different income sources and measures of wealth and social environmental controls which differ across the life cycle</p>	<p>Better collection of indicators on determinants of health</p> <p>Investing in data linkages to allow desegregation by socioeconomic status and better monitoring of health inequalities</p>
Patient experience	<p>Lack of conceptual clarity as to what is the difference between satisfaction, patient experience, and responsiveness</p> <p>Lack of clarity as to whose experiences/ satisfaction should be measured (population vs. patient vs. general experts)</p> <p>Surveys on satisfaction are very sensitive to question wording, sampling, and demographic factors</p>	<p>Developing more research to understand determinants of satisfaction, patient experience, and responsiveness</p> <p>Developing more precise questions of experience and standardized questionnaires for the evaluation of health services</p>
Efficiency	<p>The production process underlying health systems is intrinsically complex and poorly understood. Most measures make simplifying assumptions that may sometimes result in misleading data</p> <p>Outputs are generally multidimensional, and therefore preference weights are needed if they are aggregated into a single measure of attainment. The choice of such weights is intrinsically political and contentious</p> <p>A fundamental challenge in developing an efficiency measure is ensuring that the output that is being captured is directly and fully dependent on the inputs that are included in the measurement</p> <p>Environmental factors, policy constraints, population characteristics, and other factors may be largely responsible for determining health outcomes, yet it is difficult to</p>	<p>Research to find suitable metrics that measure organizational factors and administrative structures, which influence inputs and outputs</p> <p>Improve clarification on the type of efficiency being measured by different indicators</p> <p>Improve the conceptualization of the production process in order to better harmonize data collection efforts</p> <p>Improve collection of high-quality comparable data on outputs, inputs, and environmental factors necessary for risk adjustments</p> <p>Invest in research to refine methodologies for whole-system efficiency measurement</p> <p>Find a balance between whole-system measures and more fragmented efficiency measures</p> <p>More consideration of how indicators take</p>

*(continued)*

**Table 5** (continued)

Performance domain	Challenges for measurement	Developments in measurement
	<p>incorporate all possible determinants appropriately into an efficiency assessment</p> <p>From an accounting perspective, the assignment of inputs and associated costs to specific health system activities is fundamentally problematic, often relying on arbitrary accounting rules or other questionable assignments</p> <p>Although researchers have developed indicators that seek to measure full production processes, these measures are often not the most informative for policy makers looking to identify and address inefficiencies</p> <p>Many outputs are the results of years of health system endeavor and cannot be attributed to inputs in a single period</p>	<p>static and dynamic elements of inputs and outputs into account</p>

experience measures. Indicators such as avoidable mortality, which seek to measure the contribution of health care to health, are also being better developed and more frequently used. Indeed, indicators are being selected through rigorous selection mechanisms that aim to identify how appropriate they are, rather than how readily available they are. In addition, risk adjustment techniques have become more advanced and allow us to better control for exogenous factors that may lead to changes in performance.

## References

- Aday LA, et al. Evaluating the healthcare system: effectiveness, efficiency, and equity. Chicago: Health Administration Press; 2004.
- Atun R, Mendabde N. Health systems and systems thinking. In: Coker R, Atun R, McKee M, editors. Health systems and the challenge of communicable diseases: experiences from Europe and Latin America. European Observatory on Health Systems and Policies Series; 2008.
- Commonwealth Fund. Framework for a high-performance health system for the United States. New York: The Commonwealth Fund; 2006.
- Davies H. Measuring and reporting the quality of health care. NHS Quality Improvement Scotland; 2005.
- Hurst J, Jee-Hughes M. Performance measurement and performance management in OECD health systems. OECD Labour Market and Social Policy Occasional Papers No. 47. Paris: Organisation for Economic Co-operation and Development; 2001.
- Iezzoni L. Risk adjustment for measuring health outcomes. Arlington: Health Administration Press/AUPHA; 2013.
- IHP. Monitoring performance and evaluating progress in the scale-up for better health: a proposed common framework. Document prepared by the monitoring and evaluating working group of the International Health Partnership and Related Initiatives (IHP+) Led by the WHO and the World Bank; 2008.
- Jacobs R, Goddard M, Smith P. How robust are hospital ranks based on composite performance measures? *Med Care*. 2005;43(12):1177–84.
- Jee M, Or Z. Health outcomes in OECD countries: a framework of health indicators for outcome oriented policymaking. OECD Labour Market and Social Policy Occasional Papers No. 36. Paris: Organisation for Economic Co-operation and Development; 1999.
- Mant J. Process versus outcome indicators in the assessment of quality of health care. *International J Qual Health Care*. 2001;13(6):475–480.
- Naylor DC, Iron K, Handa K. Measuring health system performance: problems and opportunities in the era of assessment and accountability. In: Organization of Economic Co-operation and Development (OECD), editor. Measuring up: improving health system performance in OECD countries. Paris: OECD Publications; 2002.
- Papanicolas I, Kringos D, Klazinga NS, Smith PC. Health system performance comparison: new directions in research and policy. *Health Policy*. 2013;112(1–2):1–3. 2013; ISSN 0168–8510.
- Papanicolas I, Smith PC. Theory of system level efficiency in health care. In: Culyer AJ, editor. Encyclopedia of health economics. Philadelphia: Elsevier; 2014. p. 386–394. ISBN 9780123756787.
- Rigby M, Roberts R, Purves I, Robins S. Realising the fundamental role of information in health care delivery & management: reducing the zone of confusion. Research report. Nuffield Trust; 1999.

- Roberts MJ, et al. *Getting health reform right: a guide to improving performance and equity*. Oxford: Oxford University Press; 2008.
- Sicotte C, et al. A conceptual framework for the analysis of health care organizations' performance. *Health Serv Manage Res*. 1998;11:24–48.
- Smith PC. Developing composite indicators for assessing health system efficiency. In: Smith PC, editor. *Measuring up: improving the performance of health systems in OECD countries*. Paris: Organization for Economic Cooperation and Development; 2002.
- Smith PC, et al., editors. *Performance measurement for health system improvement: experiences, challenges and prospects*. Cambridge: Cambridge University Press; 2009.
- Terris DD, Aron DC. Attribution and causality in health care performance measurement. In Smith PC, Mossialos E, Leatherman S, Papanicolas I, editors. *Performance measurement for health system improvement: Experiences, challenges and prospects*. Cambridge: Cambridge University Press; 2009.
- WHO. *World health report 2000. Health systems: improving performance*. Geneva: World Health Organization, 2000.



Gregory Marchildon

## Contents

<b>Introduction</b> .....	770
<b>Organization and Governance</b> .....	771
<b>Financing</b> .....	773
<b>Physical and Human Resources</b> .....	773
<b>Delivery of Health Services</b> .....	774
<b>Reforms</b> .....	775
<b>Assessment</b> .....	776
<b>References</b> .....	777

### Abstract

With a population of 35 million people spread over a vast area, Canada is a highly decentralized federation. Provincial governments carry much of the responsibility for the governance, organization, and delivery of health services although the federal government plays an important role in maintaining broad standards for universal coverage, direct coverage for specified populations, data collection, health research, and pharmaceutical regulation. Roughly 70% of total health spending is financed from the general tax revenues of federal, provincial, and territorial governments.

Most public revenues are used to provide universal access to acute, diagnostic, and medical care services that are free at the point of service as well as more targeted (nonuniversal) coverage for prescription drugs and long-term care services. In the last decade, there have been no major pan-Canadian health reforms, but individual provincial and territorial governments have focused on reorganizing and fine-tuning their regional health system structure and improving the quality, timeliness, and patient experience of primary, acute, and chronic care services. While Canada's system of universal coverage has been effective in providing citizens with deep financial protection against hospital and physician costs, the narrow scope of coverage has also produced some gaps in coverage and equitable access (Romanow 2002).

---

G. Marchildon (✉)  
Institute of Health Policy, Management and Evaluation,  
University of Toronto, Toronto, ON, Canada  
e-mail: [greg.marchildon@utoronto.ca](mailto:greg.marchildon@utoronto.ca)

### Introduction

Canada is the second largest country in the world as measured by area, with a mainland that spans a distance of 5514 km from east to west and 4634 km from north to south. The climate is northern in nature with a long and cold winter seasons experienced in almost all parts of the country. The country has a population of 35 million with most of the population concentrated in urban centers close to the border with the United States and the remainder scattered over vast rural and remote areas (Fig. 1).

Canada is a high-income country with an advanced industrial economy and one of the world's highest Human Development Index rankings. Relative to other OECD countries, Canada's economic performance has been solid despite the recession triggered by the financial crisis of 2008.

The burden of disease is among the lowest in the OECD even though Canada's ranking, based on health-adjusted life expectancy (HALE), slipped from second in 1990 to fifth by 2010 (Murray et al. 2013). The two main causes of death in Canada are cancer and cardiovascular disease.

Canada is a constitutional monarchy, based on a British-style parliamentary system, and a federation with two constitutionally recognized orders of government. The federal government is responsible for certain aspects of health and pharmaceutical regulation and safety, data collection, research funding, and some health services and coverage for specific populations, including First Nations and Inuit. The second order of governments consists of ten provincial governments which bear the principal responsibility for a broad range of social policy programs and services (Marchildon 2013).



Fig. 1 Map of Canada



## Organization and Governance

In Canada, the governance, organization, and delivery of health services are highly decentralized for at least three reasons: (1) the constitutional responsibility of provinces for the funding, administration, and delivery of most health services, (2) the status of physicians as independent contractors, and (3) the existence of multiple organizations, from regional health authorities to privately owned and governed hospitals and clinics, all of which operate at arm's length or independently from provincial governments.

Provincial and territorial governments are responsible for administering their own tax-funded universal, first-dollar coverage programs. Historically, the federal government used its spending power to encourage the introduction of these programs based on high-level national principles, including the portability of coverage among provinces and territories. In most provinces, health services are organized and delivered by regional health authorities (RHAs) which have been legislatively delegated to provide hospital, long-term, and community care as well as improve population health within defined geographical areas.

Provincial ministries of health retain the responsibility to provide targeted coverage for pharmaceuticals and for remunerating physicians. Most physicians work on fee-for-service with fee schedules determined through negotiations between the medical associations and ministries of health at the provincial level of government. As independent professionals as opposed to salaried employees, physicians have considerable autonomy in terms of the managerial control of provincial health ministries or RHAs.

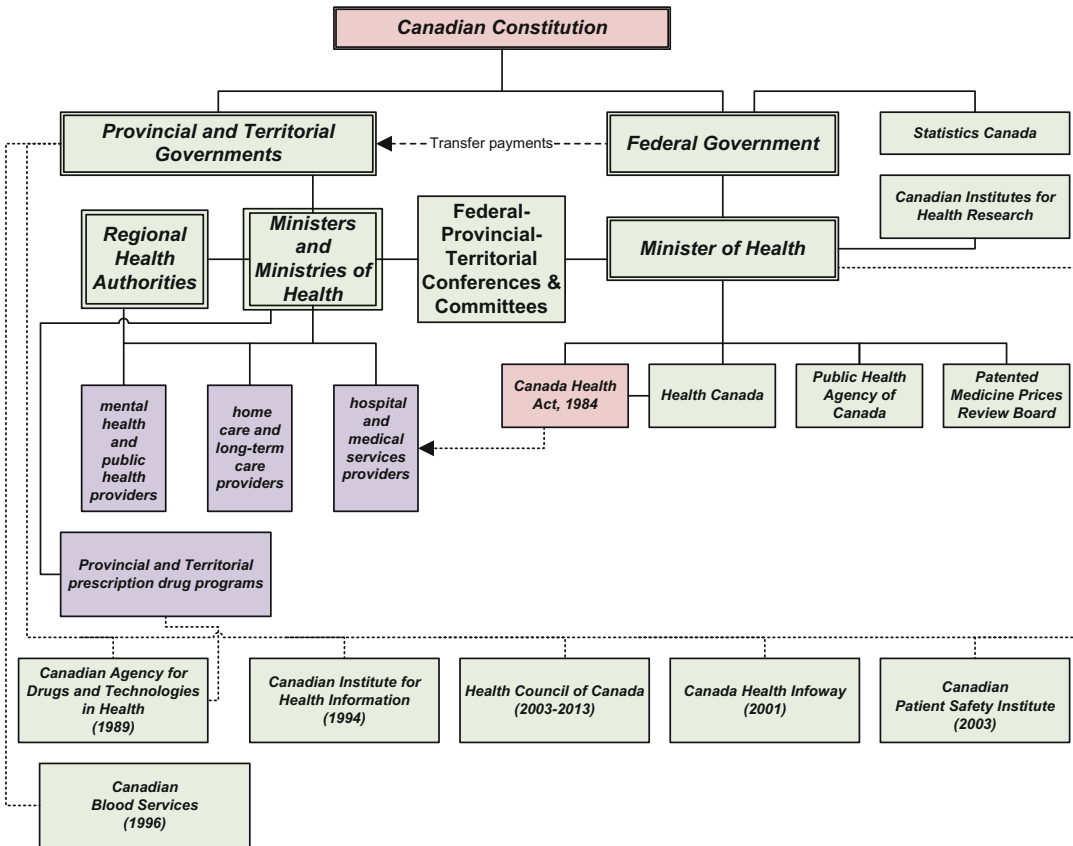
Despite this apparent decentralization, the federal government retains significant steering responsibilities. Through its cash transfers to the provincial governments and the threat of their withdrawal, the federal government sets pan-Canadian standards for hospital and medical care services through the Canada Health Act. The federal department of health – Health Canada – is responsible for ensuring that provincial governments are adhering to the five criteria in the Act:

public administration, comprehensiveness, universality, portability, and accessibility. Established in 2004 in response to the lack of national direction during the severe acute respiratory syndrome (SARS) epidemic the year before, the Public Health Agency of Canada performs a broad array of public health functions including infectious disease control, surveillance, and emergency preparedness and, through community partners, facilitates various health promotion and illness prevention initiatives (Fig. 2).

Due to the constitutional division of powers in Canada, there is no single ministry or agency responsible for system-wide national planning. Provincial ministries of health are responsible for planning and regulating their respective health systems, but they collaborate through mechanisms such as federal-provincial-territorial councils and working groups of Ministers and Deputy Ministers of Health. The provincial and federal governments have also established a number of specialized intergovernmental agencies to pursue more specialized objectives including health data collection and dissemination (the Canadian Institute for Health Information or CIHI), health technology assessment (the Canadian Agency on Drugs and Technologies in Health or CADTH), electronic health records (Canada Health Infoway), and patient safety.

Provincial and territorial governments regulate health facilities and organizations since RHAs are delegated authorities without a law-making or regulatory capacity. These governments are also responsible for managing blood products and services through Canadian Blood Services in most of the country and Héma-Québec in the province of Quebec. Provincial or other governments are not directly involved in facility accreditation, and health organizations are accredited on a voluntary basis through Accreditation Canada, a membership-based nongovernmental body. Most health professions, including physicians and nurses, are self-regulating within each province and territory based on framework laws established by the relevant governments.

Six provincial governments have established health quality councils to work with health providers and organizations to improve quality and



**Fig. 2** Organization of the Canadian health system

safety, as well as report outcomes to the general public. However, no government has given a provincial quality council the power to regulate quality or set enforceable standards.

The federal government through Health Canada regulates medical devices; determines the initial approval and labeling of all prescription drug therapies, herbal medicines, and homeopathic preparations; and prohibits direct-to-consumer advertising of pharmaceuticals. Pharmaceutical advertising targeting physicians is subject to federal law as well as to codes established by industry associations. The federal government has exclusive jurisdiction over the patenting of new inventions, including pharmaceuticals, and patent protection is set at the 20-year OECD norm. Provincial governments use a number of regulatory tools, including reference pricing, licensing of generics, bulk purchasing, tendering, and discounting, to contain the

cost of their respective prescription drug plans (Paris and Docteur 2006).

Due to a high degree of health system decentralization, physician autonomy, and onerous privacy laws, Canada has been slower than other countries in integrating information and communications technology (ICT) into health delivery. In a 2009 survey of 11 OECD countries by the Commonwealth Fund, Canadian family doctors scored the lowest in terms of using electronic health records (EHRs) and had the lowest electronic functionality (Schoen et al. 2009). Although the evidence is limited and now somewhat dated, it does appear that hospitals in Canada are also behind in their adoption and use of ICT (Urowitz et al. 2008).

Three provincial organizations and one national-level organization provide health technology assessments (HTA) to provincial and

federal ministries of health and delegated health authorities. As the sole pan-Canadian HTA agency, CADTH's mandate is to provide evaluations of new prescription drugs, as well as medical devices, procedures, and systems, to federal, provincial, and territorial governments. CADTH's recommendations are advisory in nature, and it is up to the governments in question to decide whether or not to introduce these technologies.

The patient rights movement is relatively underdeveloped in Canada compared to similar movements in the United States and Western Europe. While there are patient-based organizations focusing on particular diseases, there are only a handful of more broadly based, rights-oriented patient groups. In recent years, patient rights have been exercised through the courts, relying on the constitutional "right to life, liberty and security of the person" in the Canadian Charter of Rights and Freedom, although most attempts to extend this to a right of access to quality health care within a reasonable time have failed (Jackman 2010).

Patients and their respective physicians have been more successful in using such Charter rights to create a right to private health care and private health insurance. In 2005, the Supreme Court of Canada provided a limited form of this right in a situation where the majority of the court interpreted public waiting lists for certain types of elective surgery as unreasonable (Flood et al. 2005).

---

## Financing

Every provincial and territorial government provides universal coverage to medically necessary hospital, diagnostic, and medical care services (Taylor 1987). These 13 governments act as single payers in providing full coverage for their respective provincial and territorial residents. In return for receiving federal transfers, provincial and territorial benefits are provided on a first-dollar basis and on the same terms and conditions to all residents as stipulated in the Canada Health Act. Moreover, these

benefits are portable among the provinces and territories. Beyond this so-called Medicare coverage, federal, provincial, and territorial governments offer their own categorical programs in, and targeted benefits for, long-term care and prescription drugs.

Based on 2011 data, federal, provincial, and territorial governments were responsible for funding 70.4% of all health spending in Canada, the majority of which is raised through general taxation. Three provinces supplement their revenues through annual health-care premiums, but these too flow into provincial general revenue funds. The remaining health financing comes from out-of-pocket payments (14.7%), private health insurance (11.8%), and other sources (3.1%) (CIHI 2013).

Since the Canada Health Transfer constitutes roughly 20% of total provincial government health expenditures, the provincial governments are responsible for raising the lion's share of revenues for health (CIHI 2013). Provincial tax revenues come from a number of sources, including (in rough order of importance) individual income taxes, consumption taxes (including "sin" taxes on alcohol and gambling), and corporation taxes. In those provinces benefitting from an abundance of natural resources, resource royalties and taxes are significant sources of revenue (Marchildon 2013).

Consistent with being a tax-based Beveridge-style health system, there is limited pooling of funds in the Canadian system. However, there is a type of pooling through cash transfers – from the federal government (which collects tax at the national level) to the provincial and territorial governments and from provincial governments (which pool federal transfers with own-source revenues) to RHAs – which, as public non-governmental bodies, have no autonomous powers of taxation.

---

## Physical and Human Resources

From the 1940s until the 1960s, Canada experienced a boom in hospital building encouraged by the introduction and expansion of universal hospital

coverage and federal hospital construction grants. By the 1990s, much of this hospital infrastructure was outdated. Some provincial governments also felt burdened with too many small and inefficient hospitals in rural and remote areas. As a result, hospitals were closed, consolidated or converted, and, in some provinces, put under the governance and ownership of newly created RHAs (Ostry 2006).

Despite recent reinvestments in hospital capital, less in bricks and mortar and more in medical equipment, imaging technologies, and ICT, the number of acute care beds per capita has continued to decline. This is in part the result of improvements in clinical procedures and the expansion of non-hospital-based surgical clinics that specialize in day surgeries. Although in the past Canada had fallen behind other OECD countries in terms of the supply and use of advanced imaging equipment, the supply of computed tomography (CT) scans, magnetic resonance imaging (MRI), and positron emission (PET) scans is now closer to the OECD average.

After a lengthy period in the 1990s when the supply of physicians and nurses was reduced because of the concerted efforts of governments to reduce spending and pay down accumulated public debt, spending on the health workforce has climbed steadily since the turn of the century. Medical, nursing, and other health profession faculties have expanded their seats to produce more graduates, even while an increasing number of foreign-educated doctors and nurses have immigrated to Canada.

With the exception of physicians, most health workers are employees of health-care organizations, RHAs, and health ministries and are remunerated through salary and wage income. The majority of health workers in the public sector are unionized, and their remuneration is set through collective bargaining agreements. The majority of physician remuneration is through fee-for-service. However, alternative payment contracts – particularly for general practitioners (GPs) – are becoming more common in part as a result of primary care reforms.

## Delivery of Health Services

All provincial and territorial governments have public health programs. They also conduct health surveillance and manage epidemic response. While the Public Health Agency of Canada develops and manages programs supporting public health programs at the provincial, regional, and local community levels, the stewardship for most day-to-day public health activities and supporting infrastructure remains with the provincial and territorial governments.

Most primary care is provided by GPs and family physicians, with family medicine recently recognized as a specialization by the Royal College of Physicians and Surgeons of Canada. Although mandated through policy and practice rather than law, GPs and family physicians act as gatekeepers, deciding whether patients should obtain diagnostic tests and prescription drugs or be referred to medical specialists.

Provincial ministries have renewed efforts to reform primary care in the last decade. Many of these reforms focus on transitioning from the traditional physician-only practice to interprofessional primary teams capable of providing a broad range of primary, health promotion, and illness prevention services.

Almost all acute care is provided in public or private nonprofit hospitals, although specialized ambulatory and advanced diagnostic services are sometimes provided in private for-profit clinics, particularly in larger urban centers. Most hospitals have an emergency department that is fed by independent emergency medical service units providing first response care to patients while being transported to the hospital. Due to the scattered nature of remote communities without secondary and tertiary care, provincial and territorial governments provide air-based medical evacuation, a major expenditure item for the most northern jurisdictions (Marchildon and Torgerson 2013).

Long-term care services, including supportive home and community care, are not classified as insured services requiring universal access under the five national criteria set out in the Canada Health Act. As a consequence, public policies, subsidies, programs, and regulatory regimes for

long-term care vary considerably among the provinces and territories. Facility-based long-term care (LTC) ranges from residential care with some assisted living services to chronic care facilities (originally known as nursing homes) with 24-hour-a-day nursing supervision. Most residential care is privately funded, whereas high-acuity LTC (requiring 24-hour-a-day nursing supervision) is heavily subsidized by provincial and territorial governments (Canadian Healthcare Association 2009).

Until the 1960s, the locus of most mental health care was in large, provincially run psychiatric hospitals which in turn had evolved out of the nineteenth century asylum and the twentieth century mental hospital. With the introduction of pharmaceutical therapies and a greater focus on reintegration into the community, mental health conditions have since been mainly treated on an outpatient basis or, in the case of severe episodes, in the psychiatric wards of hospitals. GPs provide the majority of primary mental health care, in part because medical care is an insured service with first-dollar coverage, whereas psychological services are provided largely on a private basis.

While drugs administered in hospitals are fully covered as an insured service under the Canada Health Act, every provincial and territorial government has a prescription drug plan that covers a portion of the cost for outpatient prescription drugs. The majority of these drug plans target low-income or retired residents. The federal government provides pharmaceutical coverage for eligible First Nations and Inuit. These public insurers depend heavily on health technology assessment to determine which drugs should be included in their respective formularies.

Almost all dental care is delivered by independent practitioners, and 95% of these services are paid privately. Dental services are paid for through private health insurance – provided mainly through employment-based benefit plans – or out of pocket. As a consequence of access being largely based on income, outcomes are highly inequitable.

For historical reasons, the federal government finances a host of health service programs

targeting Aboriginal Canadians, in particular eligible First Nation and Inuit citizens. These services include health promotion, disease prevention, and public health programs as well as coverage for medical transportation, dental services, and prescription drug therapies. Despite these targeted efforts, the gap in health disparity between these Aboriginal citizens and the majority of society remains large. Since the 1990s, there have been a series of health-funding transfer agreements between the federal government and First Nation governments – largely based on reserves in rural and remote regions of Canada. At the same time, there has been an Aboriginal health movement advocating for a more uniquely Aboriginal approach to health and health care (Marchildon 2013).

---

## Reforms

There have been no major pan-Canadian health reforms in the past decade. However, individual provincial governments have concentrated on two categories of reforms: (1) structural change involving the governance and management of health services as a more integrated health system, mainly through the reorganization and fine-tuning of their regional health systems, and (2) process-type reforms, aimed at addressing bottlenecks in delivery, improving patient responsiveness and elevating both quality and safety.

The introduction of RHAs allowed provincial governments to directly manage the health system through arm's-length delegated bodies. RHAs manage services as purchaser-providers except in Ontario when the local health integration networks (LHINs) fund (purchase) but do not deliver services directly. The purpose of the reform was to gain the benefits of vertical integration by managing facilities and providers across a broad continuum of health services and to improve the coordination of “downstream” curative services with more “upstream” public and population health services and interventions. In the last decade, there has been a trend to reduce the number of RHAs, thereby increasing the geographic and population size of RHAs in each province,

in order to capture greater economies of scale and scope.

Influenced chiefly by quality improvement initiatives in the United States and the United Kingdom, provincial ministries of health have established new institutions, mechanisms, and tools to improve the quality, safety, timeliness, and responsiveness of health services. Six provinces have established health quality councils to accelerate quality improvement initiatives at the provincial, regional, and clinical levels. Some provinces have also launched patient-centered care initiatives aimed at improving the experience of patients and informal caregivers. Patient dissatisfactions with long wait times for elective surgery as well as specialist and diagnostic services have triggered efforts in all provinces to better manage and reduce wait times.

In contrast, the federal government has largely removed itself from engaging the provinces in any pan-Canadian reform efforts. This is in part the consequence of the perceived failure of the “10-Year Plan to Strengthen Health Care,” signed by the Prime Minister and the Premiers of all provinces and territories in 2004.

The “10-Year Plan” ends in the fiscal year 2013–2014. In December 2011, the federal government announced its reconfiguration of the Canada Health Transfer for the decade following the 10-Year Plan. After 2014, increases in the transfer to the provinces, originally 6% per annum, will be held to the rate of economic growth with a minimum floor of 3%, and all transfers will be made on a pure per capita basis, without taking into consideration the tax capacity of the provinces. The removal of any equalization component in the transfer will make it more difficult for lower-income provinces to continue to ensure coverage is maintained at the standard enjoyed in higher-income provinces.

---

## Assessment

The model of universal Medicare has been effective in protecting Canadians against high-cost hospital and medical care. At the same time, the

narrow scope of the benefit package has resulted in larger gaps in coverage, as pharmaceutical therapies and LTC have grown in importance over time. Since 70% of financing for health care in Canada comes from general taxation, there is more equity in financing, but there is less equity in financing for the remaining 30%, which comes from out-of-pocket sources and employment-based insurance benefits associated with better-paid jobs.

There are disparities in terms of access to health care, but outside of a few areas such as dental care and pharmaceuticals, they do not appear to be large. For example, there appears to be a pro-poor bias in terms of primary care but a pro-rich bias in the use of specialist physician services, but the gap in either case is not large.

There is also an historic east-west economic gradient dividing the less wealthy provinces in eastern Canada and the wealthier provinces in the more western parts of the country from Ontario to British Columbia. In the present, the economic division is more between those provinces rich in natural resources – particularly petroleum-producing provinces such as Alberta, Saskatchewan and Newfoundland – and those provinces without such resources. These differences are addressed through equalization payments from federal revenue sources to “have-not” provinces that ensure the latter have the revenues necessary to provide comparable levels of public services, including health care, without resorting to prohibitively higher tax rates.

While Canadians are generally satisfied with the financial protection offered by Medicare, they are less satisfied with their access to particular services. Beginning with the budget cuts to health care in the 1990s, emergency rooms became overcrowded and waiting times for non-urgent care became lengthier (Tuohy 2002). Based on a survey of patients in selected OECD countries conducted in 2010, Canada ranked poorly in terms of waiting times for physician care and nonurgent surgery (Schoen et al. 2010). However, based on relevant mortality and morbidity indicators of health system performance, such as amenable mortality, Canada fares considerably better, posting better

outcomes than those in the United Kingdom and the United States (Nolte and McKee 2008).

Canadian performance in terms of the quality of health care has also improved in recent years. This may be a result of the policy focus of provincial governments on quality, assisted by health quality councils and the comparative indicators collected and disseminated by the Canadian Institute for Health Information. This improvement is now being extended to patient responsiveness in the hope that this will improve the quality of the patient experience.

There have been few studies of technical efficiency of health systems in Canada (CIHI 2011). However, some provincial governments are beginning to arrange for external evaluations of recent reforms. In the particular, the recent application of “lean production” methodologies in some provincial health systems can be interpreted as an effort to achieve greater efficiency. First developed by Toyota to achieve greater technical efficiency and higher quality in automobile productions, lean techniques have been applied to hospitals and other health settings in a number of provinces. The objectives of the lean projects have ranged from reducing surgical wait times to improving patient safety (Fine et al. 2009).

Due to the number of trends and institutional changes, health systems in Canada are more transparent today than in the past. Whether in their roles as citizens, taxpayers, patients, or caregivers, Canadians have been demanding greater transparency on the part of their governments and publicly funded health-care organizations and providers. They now receive a range of health information and analysis from a number of new provincial and intergovernmental organizations, including the Health Council of Canada which provides accessible reports on the state of Canadian health care. In addition, a number of advocacy organizations and think tanks also provide regular reports on health system issues of concern and interest to the general public.

## References

- Canadian Healthcare Association. *New directions for facility-based long term care*. Ottawa: Canadian Healthcare Association; 2009.
- CIHI. *Health care cost drivers: The facts*. Ottawa: Canadian Institute for Health Information; 2011.
- CIHI. *National health expenditure trends, 1975–2013*. Ottawa: Canadian Institute for Health Information; 2013.
- Fine BA, et al. *Leading lean: a Canadian healthcare leader's guide*. *Healthc Q*. 2009;12(3):32–41.
- Flood CM, Roach K, Sossin L, editors. *Access to care, access to justice: the legal debate over private health insurance in Canada*. Toronto: University of Toronto Press; 2005.
- Jackman M. *Charter review as a health care accountability mechanism in Canada*. *Health Law J*. 2010;18:1–29.
- Marchildon GP. *Canada: health system review*. *Health Syst Transit*. 2013;15(1):1–179. Copenhagen: WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies.
- Marchildon GP, Torgerson R. *Nunavut: a health system profile*. Montreal/Kingston: McGill-Queen's University Press; 2013.
- Murray CJL, Richards MA, Newton JN, et al. *UK health performance: findings of the Global Burden of Disease Study*. *Lancet*. 2013;381:997–1021.
- Nolte E, McKee M. *Measuring the health of nations: updating an earlier analysis*. *Health Aff*. 2008;27(1):58–71.
- Ostry A. *Change and continuity in Canada's health care system*. Ottawa: CHA Press; 2006.
- Paris V, Docteur E. *Pharmaceutical pricing and reimbursement policies in Canada*. Paris: Organisation of Economic Co-operation and Development, Health Work Group; 2006.
- Romanow RJ. *Building on values: the future of health care in Canada*. Saskatoon: Commission on the Future of Health Care in Canada; 2002.
- Schoen C, et al. *A survey of primary care physicians in eleven countries*. *Health Aff*. 2009;28(6):w1171–83.
- Schoen C, Osborn R, Squires D. *How health insurance design affects access to care and costs, by income, in eleven countries*. *Health Aff*. 2010;29:w2323–34.
- Taylor MG. *Health insurance and Canadian public policy: the seven decision that created the Canadian healthcare system*. 2nd ed. Montreal: McGill-Queen's University Press; 1987.
- Tuohy CH. *The cost of constraint and prospects for health care reforms in Canada*. *Health Aff*. 2002;21(3):32–46.
- Urowitz S, et al. *Is Canada ready for patient accessible electronic health records? A national scan*. *BCM Med Inform Decis Making*. 2008;8:33. <http://www.biomedcentral.com/1472-6947/8/33>. Accessed 25 Sept 2012.



David Hipgrave and Yan Mu

## Contents

<b>Introduction</b> .....	780
China's Current Health System Reform .....	781
<b>Organization, Governance, and Accountability</b> .....	783
Organization of the Health System .....	783
Accountability Within Government and to the Population .....	784
<b>Planning, Regulation, and Monitoring</b> .....	786
Monitoring Progress: China's Health Information Systems and Technology .....	786
<b>Financing</b> .....	788
Sources of Funding and Accountability for Its Use .....	788
Difficulties Using Available Health Financing for Policy Implementation .....	789
Health Expenditure and Sources of Revenue .....	790
Collection and Pooling of Funds .....	790
Coverage, Benefit, and Cost Sharing .....	792
Payment Methods for Health Services .....	794
<b>Physical and Human Resources</b> .....	794
Infrastructure and Its Funding .....	794
Health Workforce and Trends .....	795
Remuneration of Health Workers .....	797
<b>Health Services Delivery and Outcomes</b> .....	797
Primary Care and Public Health .....	797
Clinical Services .....	798

---

D. Hipgrave (✉)  
UNICEF, New York, NY, USA

Nossal Institute for Global Health, University of  
Melbourne, Melbourne, VIC, Australia  
e-mail: [dhipgrave@gmail.com](mailto:dhipgrave@gmail.com)

Y. Mu  
UNICEF China, Beijing, China  
e-mail: [ymu@unicef.org](mailto:ymu@unicef.org)



Pharmaceutical Care .....	800
Private Healthcare .....	801
Health Outcomes .....	801
<b>Assessment</b> .....	803
<b>References</b> .....	804

### Abstract

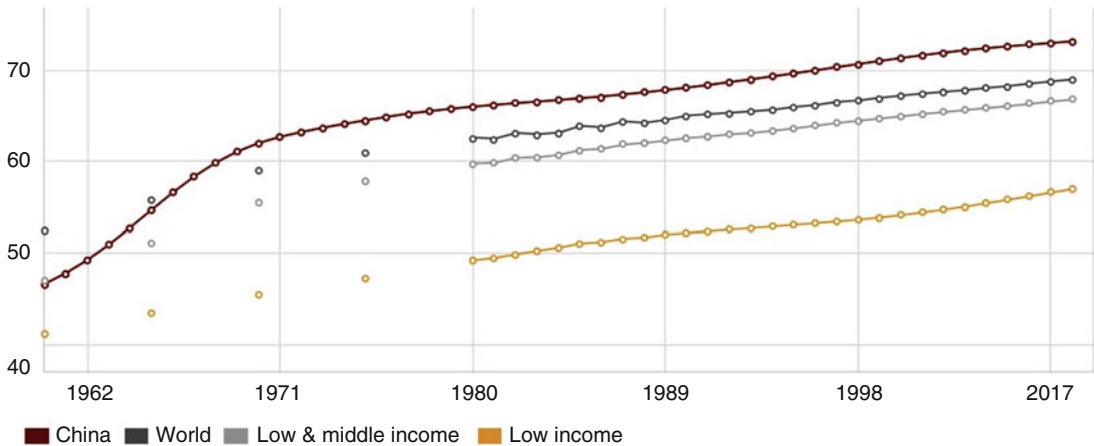
The health of China's population improved dramatically during the first 30 years of the People's Republic, established in 1949. By the mid-1970s, China was already undergoing the epidemiologic transition, years ahead of other nations of similar economic status, and by 1980, life expectancy (67 years) exceeded that of most similarly low-income nations by 7 years. Almost 30 years later, China's 2009 health reforms were a response to deep inequity in access to affordable, quality healthcare resulting from three decades of marketization, including de facto privatization of the health sector, along with decentralized accountability and, to a large degree, financing of public health services. The reforms are built on earlier, equity-enhancing initiatives, particularly the reintroduction of social health insurance since 2003, and are planned to continue until 2020, with gradual achievement of overarching objectives on universal and equitable access to health services. The second phase of reform commenced in early 2012. China's health reforms remain encouragingly specific but not prescriptive on strategy; set in the decentralized governance structure, they avoid the issue of reliance on local government support for the national equity objective, leaving the detailed design of health service financing, human resource distribution and accountability, essential drug lists and application of clinical care pathways, etc. to local health authorities answerable to local government, not the Ministry of Health. Community engagement in government processes, including in provision of healthcare, remains limited. This chapter uses the documentation and literature on health reform in China to provide a comprehensive overview of the current situation of the health sector and its reform in the People's Republic.

### List of Abbreviations

CDC	Communicable disease control
GDP	Gross domestic product
HMIS	Health MIS
HSR	Health system reform
LMIC	Low- and middle-income countries
MCH	Maternal and child health
MDGs	Millennium Development Goals
MIS	Management information system
MoH	Ministry of Health
NCDs	Noncommunicable diseases
NDRC	National Development and Reform Commission
NEDL	National Essential Drugs List
NEMS	National Essential Medicines Scheme
NHFPC	National Health and Family Planning Commission
PRC	People's Republic of China
RCMS	Rural cooperative medical (insurance) scheme
RMB	Renminbi (unit of currency)
TCM	Traditional Chinese medicine
THE	Total health expenditure
UEBMI	Urban employees basic medical insurance
URBMI	Urban residents' basic medical insurance

### Introduction

Most people are familiar with two things about modern China. The first is its physical size and enormous population. In land area, China is the world's third largest nation, theoretically spanning 4 h of time difference from west to east (while officially operating on one time zone). Its 2010 census revealed a population approaching 1.34 billion, the world's largest. China's population grew most rapidly from the late 1950s to the



**Fig. 1** Life expectancy in years: China, the world, and low-income and middle-income nations (Source: World Bank data available at <http://data.worldbank.org/>)

early 1970s, due to the formerly high fecundity of its women alongside a rapid fall in the crude death rate due to communicable disease control (CDC) and basic public health measures. Life expectancy also rose rapidly during this period (Fig. 1) (Hipgrave 2011a).

The second familiar aspect is China's meteoric economic development, with an average annual growth rate of around 10% for most of the last 30 years, only falling to 7–8% since the global financial crisis.

These familiar aspects of China have depended on the health of its population improving dramatically during the first 30 years of the People's Republic of China (PRC) since its establishment in 1949. By the mid-1970s, China was already undergoing the epidemiologic transition, years ahead of other nations of similar economic status, and by 1980, life expectancy in low-income China (67 years) exceeded that of most similarly low-income nations by 7 years (Jamison et al. 1984).

However, with CDC (Hipgrave 2011a), economic development, rapid urbanization, and a dramatically ageing population, China's health system now faces a vastly different range of issues. China will soon become the first large nation to age before achieving developed nation status. Noncommunicable diseases (NCDs) now account for over 80% of deaths in China and almost 70% of its total disease burden (The World Bank Human Development Unit 2011). A World

Bank analysis of NCDs in China (The World Bank Human Development Unit 2011) concluded that “a reduced ratio of healthy workers to sicker, older dependents will certainly increase the odds of a future economic slowdown and pose a significant social challenge in China” (page 2). Equally challenging is the provision of new services for the prevention and management of chronic illness and the government's averred commitment to equity and universal health coverage. These challenges and commitments were among the stimuli to the major health system reform (HSR) that China commenced in 2009 (State Council 2009).

### China's Current Health System Reform

China's most recent HSR was a response to deep inequity resulting from three decades of marketization and de facto privatization of the health sector. It was the culmination of many years of debate (Tang et al. 2014a) after acknowledged inaction on the heavy burden of healthcare on household expenditure (Blumenthal and Hsiao 2005; Huang 2011; Liu 2004; Liu et al. 2003; Tang et al. 2008). It comprises initiatives in five main areas:

1. Expanding the coverage and benefit of health insurance schemes in urban and rural areas
2. Establishing a national essential medicines scheme to ensure the availability of affordable

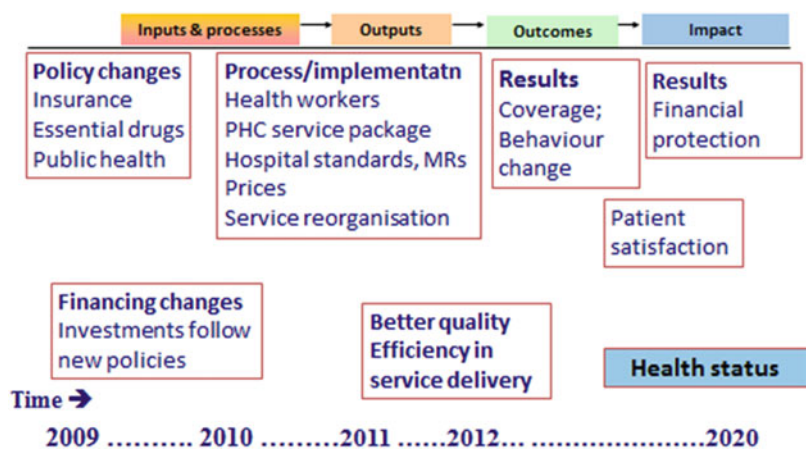
- medicines and reduce the ability of health providers to profit from the sale of drugs
- 3. Improving basic service availability and quality while also reducing referrals to specialist care and hospitals
- 4. Ensuring the availability of basic public health services for all populations
- 5. Piloting public hospital reform, particularly in order to separate hospital management and clinical service provision

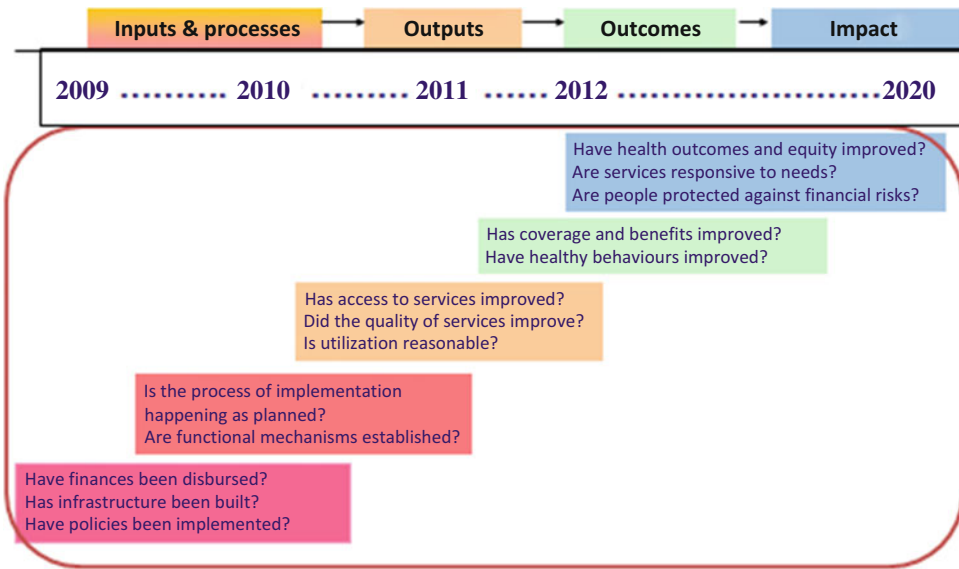
The current HSR builds on earlier, equity-enhancing initiatives including the reestablishment of rural health insurance (Meng et al. 2012) and subsidized hospital maternity services (Feng et al. 2010a). Early progress on the first phase of China’s current HSR (2009–2012) was extensively reviewed, both internally by domestically commissioned teams of international (unpublished) and national experts (Wu and Yang 2013; Li and Chen 2012) and externally (Yip et al. 2012). The Reform is planned to continue to 2020, with gradual achievement of its overarching objectives on universal and equitable access to health services; the second phase (2012–2015) was announced in early 2012 (Ministry of Health 2012a), and a major additional pronouncement on county hospital reform was made in early 2014 (State Council 2014). Monitoring and evaluation of the Reform is slated to prioritize its different hierarchical elements (Figs. 2 and 3), although detailed plans for such evaluation have not been released.

China’s commitment to HSR indicates its ongoing priority for the highest echelons of government (Ministry of Health 2012a). The four-year plan for phase 2 reiterates the goal of universal access to basic health services and seeks to resolve constraints to the supply of China’s increasing and diverse health needs. It again commits to expanding insurance benefits and introduces priority to unifying China’s several health insurance schemes; it encourages development of commercial insurance and the introduction of capitation and other payment reforms to separate doctors from the financial management of hospitals; it suggests that the private sector should manage 20% of health services by 2015; family general practice is promoted alongside expanding community and public health services, and the drug production, prescription, and pricing will be further consolidated and regulated; performance-based funding of health staff is also mentioned. These individual areas are discussed further below.

The plan is encouragingly specific but not prescriptive on strategy and avoids the issue of local accountability for financing various health programs, stipulating only that government spending on health should gradually increase as a proportion of total government expenditure. This vagueness hints at a major problem for China’s health sector, the reliance on local government support for the national equity objective (Hipgrave et al. 2012). Another major problem remains the difficulty of reforming hospital management, effectively undoing the private, for-profit system that evolved over recent decades. As a result, China’s

**Fig. 2** Mapping China’s health reform priorities over 2009–2020 (Source: WHO China)





**Fig. 3** Focus of monitoring and evaluation of health reforms in China (Source: WHO China)

HSR has not yet reduced the proportional financial burden of healthcare on households or their risk of catastrophic expenditure on health (Meng et al. 2012).

## Organization, Governance, and Accountability

### Organization of the Health System

China’s former Ministry of Health (MoH) recently merged with the body previously responsible for family planning to form the National Health and Family Planning Commission (NHFPC). The Commission contains 23 different departments, offices, and bureaux responsible for setting standards and for the planning, administration, oversight, and reporting on China’s health sector. However, as with most of China’s social sectors, there is a heavy decentralization of responsibility for local planning, financing, and implementation of health services in China (Wong 2010; Zhou 2010a). In China’s decentralized system, policies and reform guidelines are set at national level but implementation is delegated to local authorities at provincial and

lower levels. A hierarchy of health authorities oversees these issues at province, prefecture, county, and township levels.

In China’s political economy and governance structure, local health authorities are more responsive to local government than to higher-level cadres within the health sector, meaning that uptake of national policies and recommendations is only guaranteed if there is broad agreement across all sectors of government and at local government level. In the past, when the health sector was of low priority, this severely limited the implementation of national laws relevant to the health sector. For example, the 1989 Law on Control of Infectious Diseases conferred on local government’s responsibility for various forms of reporting and action, but was weakly implemented, culminating in the wake-up call of SARS in 2003, redrafting of the law and major reform of CDC (Hipgrave 2011a; Wang et al. 2008a). Initiatives depending on countrywide uptake such as the 2010 national measles vaccination campaign still rely heavily on local funding and prioritization; recent environmental degradation and food and drug safety scandals are further evidence of the lack of cross-sectoral priority given to the health sector in China. The partial

rollback of the one-child policy announced at national level in 2013 remains subject to interpretation and optional implementation by provincial governments. Despite its evident high priority (Tang et al. 2014a), many aspects of the HSR itself are dependent on the same support and follow-up by provincial and even county governments (Hipgrave et al. 2012; Brixi et al. 2012).

To ensure that HSR would receive adequate local priority despite this structure and accountability, in early 2010 the HSR Leading Group in the State Council signed “accountability contracts” with provinces on key reform areas, for subsequent delegation and implementation at lower levels (China News Network 2010). In some provinces, a few key HSR targets such as health insurance coverage were incorporated into subnational officials’ performance evaluation criteria, which has been effective in ensuring progress. However, in other, more complicated reform areas, such as strengthening primary healthcare, public hospital reforms, and others, ensuring progress has been more difficult. Indeed, the reform of public hospitals suffers from a lack of consensus or clear national guidance on direction, limiting its prioritization and implementation outside pilot areas, particularly at low levels.

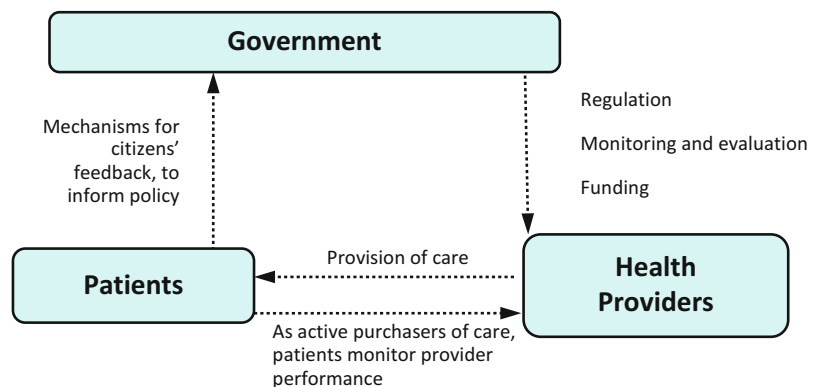
### Accountability Within Government and to the Population

Figure 4 illustrates the ideal accountability relationships among government, healthcare providers, and citizens (society) in the delivery of healthcare.

However in China, such relationships have not yet been forged. While there are promising moves to make local government generally more accountable to the public (such as measurement of “green gross domestic product (GDP)” and independent surveys of public opinion on local government performance in some provinces), the main motivation for subnational authorities remains economic development and revenue generation (Zhou 2010b). Moreover, while banking, communications, etc. are carefully regulated and monitored from above, like most social sectors, health services are largely organized and monitored at the local level. It is too costly for China’s undermanned central government to independently monitor and evaluate subnational health performance (Wong 2010; Zhou 2010b). These circumstances explain the limited ability of national health officials to ensure the HSR is fully pursued at grassroots level.

In theory, all government plans represent the will of the people as they are ratified by the National People’s Congress. However, many Congress members are unelected (in the western democratic sense) appointees, and the People’s Congress generally rubber-stamps the documents presented. However, with the increasing attention of the Party and government in China to public comment through social media, albeit increasingly censored (Osnos 2014), and local protests, there is growing acknowledgment of their answerability to the general public. Therefore, while during local planning there is almost no formal process for the public to make input, there are opportunities for the general population to voice

**Fig. 4** Accountability relationships for healthcare (Source: Adapted from The World Development Report 2004 (The World Bank 2003))

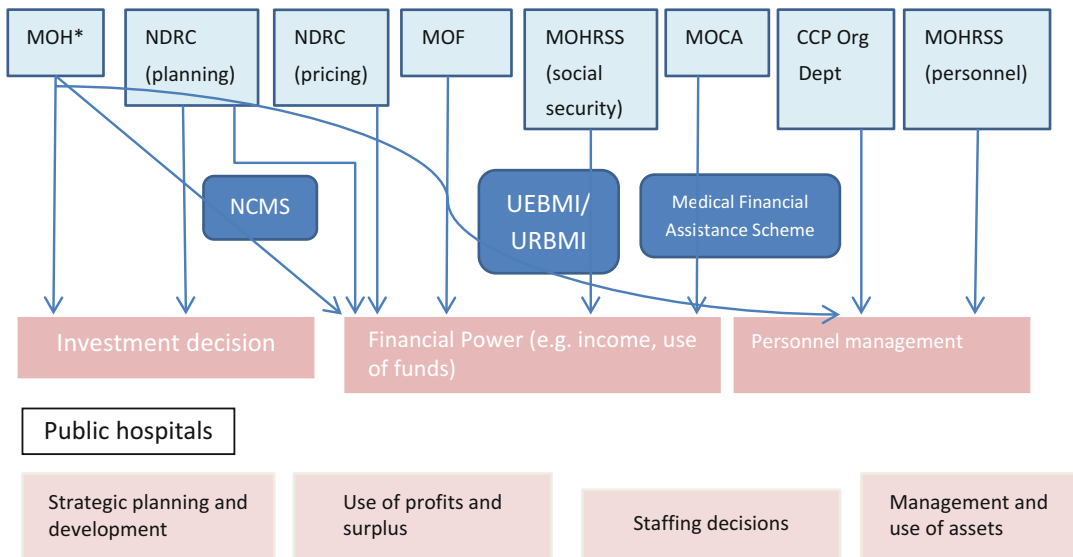


concerns through the courts, social media, petitions, protests, etc., especially when issues affect a significant proportion of a community. Although the process is usually slow (the HSR took many years to be formalized (Tang et al. 2014a)), there is usually gradual recognition and acknowledgment of the need to act. On the other hand, implementation of plans usually requires higher-level pressure on the various lower tiers of government, and this pressure progressively dissipates further down the hierarchy; it may be ignored for issues that don't have high-level and cross-sectoral support and the support of local government. Hence, targets for insurance coverage and drug price control are accepted, but controlling the environmental impact of local industry is often ignored (Human Rights Watch 2011). In this process, public influence is rather indirect and can be ignored if local economic, political, or vested interests override it.

Patients' concerns in healthcare delivery may be channeled formally through the National People's Congress at different levels (although usually only major complaints reach this level) or informally through social media. However,

mechanisms to tap the feedback of patients, as the end users of health services, have not been established. There is no ombudsman or independent regulator in China's health system, and senior appointments are normally approved by the ruling Party organization. However, since launching the HSR, government is learning that empowering patients and regularly collecting their feedback on key parameters such as service prices and quality strengthens accountability across the government levels and can help achieve the overall goals of the reform (State Council 2014). Patient satisfaction and feedback is increasingly incorporated into the performance evaluation framework for HSR implementation (Ma 2013). However, this practice has not yet been standardized, systematized, and regularized throughout China.

An example of the problem China is having in effecting the most difficult aspect of the HSR, the reform of public hospitals, was recently summarized by eminent researchers on China (Yip et al. 2012), who noted the complex web of relationships that govern this endeavor (Fig. 5). It seems likely that China will need all the years up to 2020



**Fig. 5** Dispersion of power between ministries and public hospitals in China. \*MOH Ministry of Health, NDRC National Development and Reform Commission, MOF Ministry of Finance, MOHRSS Ministry of Human Resources and Social Security, MOCA Ministry of Civil

Affairs, CCP Org Dept Organizational Department of Chinese Communist Party, NCMS New Cooperative Medical Scheme, UEBMI Urban Employee Basic Medical Insurance, and URBMI Urban Residents Basic Medical Insurance (Based on Yip et al. 2012)

to make progress in this area of reform, although some commentators doubt this will be achieved in the current context (Zhang and Navarro 2014).

## Planning, Regulation, and Monitoring

The normal sector-planning practice in China follows the National Five-Year Plan for Social and Economic Development, with different social sectors (including health) developing their respective plans at five-yearly intervals with annual updates. However, the special need for health reform did not allow China's HSR to fall neatly in line with regular national development planning, which covers two five-yearly periods per calendar decade: the first three-year phase of the HSR covered 2009–2011, while the second overlaps with the latter part of the government's 12th Five-Year Plan period: 2012–2015. Moreover, the HSR was developed as a cross-sectoral endeavor led by the national planning ministry (the National Development and Reform Commission or NDRC) to address long-accumulated concerns of the population (State Council 2009; Tang et al. 2014a). While it overlapped with a MoH planning and development activity, Healthy China 2020, the HSR was not only a MoH initiative.

As part of the government's regular planning, the new NPFPC drafts annual national health work plans with annual targets and submits annual budget proposals for approval by the Ministry of Finance and the NDRC, which approves major construction initiatives such as health infrastructure development. With major events as the HSR, new changes and innovations are often seen in the plans year on year. At subnational levels, health-related authorities (not only health bureaux) in provinces, prefectures, and counties submit annual planning and budget proposals in line with health service delivery needs and stewardship to the development planning and finance authorities at the corresponding tier. Implementation is financed by local budget supplemented by transfers from higher tiers of government (explained below). Local data should be used in formulating plans, but as there is little tradition of regular, independent, or audited data gathering in

China, desensitization of administrative and economic data is suspected (Cai 2008; Hu et al. 2011; Walter and Howie 2011; Kaiman 2013; Anonymous 2012).

Regulation of the health sector follows the accountability structure outlined above and appraises progress and achievement against high-level targets set at national and local levels. Performance assessment tends to be quantitative (relating to coverage or throughput of health services), although assessment on more subtle measures such as patient satisfaction, service quality, and disease management has commenced (as outlined in a Guidance on Performance Assessment of Basic Public Health Services Delivery, jointly promulgated by Ministry of Health and Ministry of Finance in January 2011). At management level, government officials are also increasingly being appraised according to efficiency and innovations in rolling out reform initiatives at local level.

## Monitoring Progress: China's Health Information Systems and Technology

With around 20% of the world's people, population-level changes in China's health status or indeed any globally important indicator have a major influence on corresponding global progress. For example, China's progress toward regional and global achievement of the Millennium Development Goal (MDG) targets will impact any final evaluation of the MDGs in 2015.

However, global statistics in any of the biological, physical, and social sciences can only be calculated if China's data is included and considered to be reasonably accurate, and data from China is not always available. Many lists of global indicators lack an entry from China, and the accuracy of what is released has been questioned (Cai 2008; Mulholland and Temple 2010). Usually, this is simply because China itself does not collect national statistics on the relevant indicators or not in ways comparable with other nations (e.g., see [http://www.countdown2015mnch.org/documents/2012Report/2012/2012\\_China.pdf](http://www.countdown2015mnch.org/documents/2012Report/2012/2012_China.pdf)). However, as long ago as 2000, perspectives on China's

mortality data were quite positive (Banister and Hill 2000).

The overall lack of data from China rouses suspicion. But while China's official statistics often lack breakdowns on key indicators (e.g., until recently, child mortality by gender or cause of death; nutrition status by province) or vary widely from one official source to the next (such as the annual birth cohort (Cai 2008) or number of road deaths (Hu et al. 2011)), these issues distract from China's efforts to improve the content, frequency, quality, and public availability of official data in recent decades (Banister and Hill 2000). Indeed, UNICEF's "Atlas on Children in China" publishes a wide range of official and recent data (<http://www.unicefchina.org/en/index.php?m=content&c=index&a=lists&catid=60>), and health statistics and other yearbooks are published annually (Ministry of Health 2012b; National Bureau of Statistics 2012, 2016) with a great degree of detail and disaggregation.

An increasing number of official and peer-reviewed publications on maternal and child health (MCH) in China report official government data (Wang et al. 2011, 2012; Rudan et al. 2010; Ministry of Health 2011a; Feng et al. 2010b, 2011), and this is contributing to summaries of global progress on the world's health status and MDGs 4 and 5. China relies on several different sources to provide health administrators, the public and academia with information on the health sector. While it has never conducted a demographic and health survey, and its last multi-indicator cluster survey was in 1995, China's national health services survey has been undertaken with a reasonably consistent methodology on a five-yearly basis since 1993. Many publications have used this source to assess progress in aspects of China's health system (Meng et al. 2012) and on its health indicators (Wang et al. 2012).

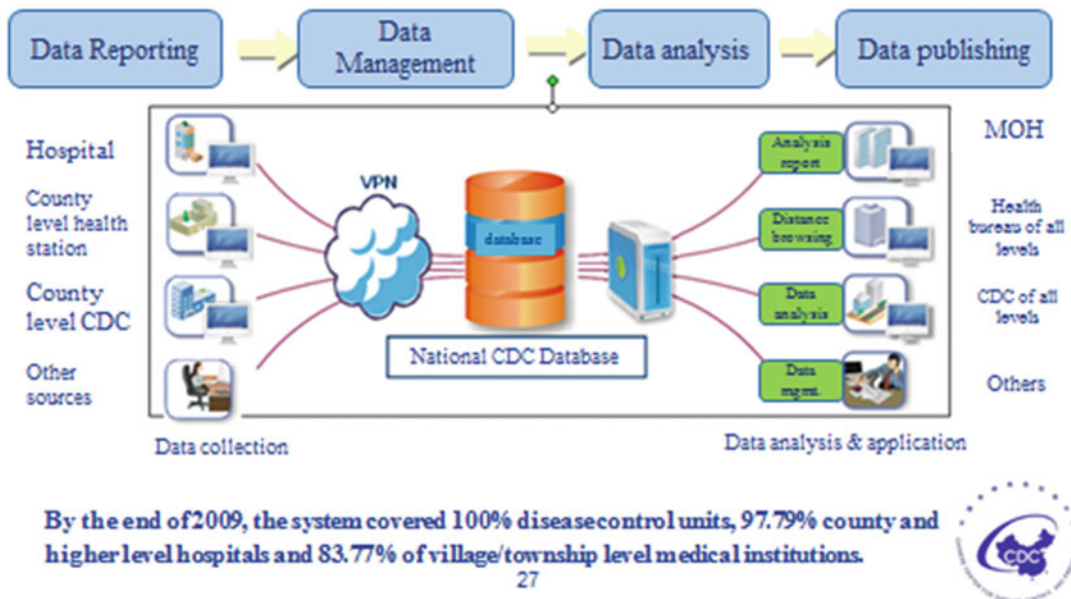
As an example of the other sources used, China's official MCH management information system (MIS) and the China Health Statistics Yearbook (Ministry of Health 2012b) rely on data from the following:

1. MCH Annual Reports: administrative reports submitted by ~3000 counties and districts across the nation (Ministry of Health 2007).

2. Maternal and Child Mortality Surveillance network, which has been summarized elsewhere (Wang et al. 2011).
3. The China Food and Nutrition Surveillance System, which surveys 40 surveillance sites on a five-yearly basis, most recently in 2010.
4. The ten-yearly National Nutrition Survey, a comprehensive, age-stratified, sex-stratified, and geographically stratified survey with a sample size of almost 200,000 (last completed in 2012).
5. The China Immunization Registration and Information System, a newly computerized administrative system that reports vaccination coverage to the NHFPC.
6. Data gathered on health facilities, human resources, equipment, and services provided to outpatients and inpatients at various sub-national levels and collected by the MoH Center for Health Statistics and Information.
7. China's National Notifiable Disease Reporting System, through which each county reports on 35 notifiable diseases. After SARS, this reporting system was massively upgraded to become web-based with reporting in real time (Fig. 6).
8. Disease Surveillance Points on births, deaths, and on cases of 35 notifiable diseases at 145 selected points around the nation.
9. China's Vital Registration System, which covers around 8% of the nation's population but is biased toward urban and eastern locations.
10. National Health Services Survey, which focuses on health status, service uptake, and health financing (Meng et al. 2012); it was last conducted in 2013.
11. National Census, last conducted in 2010 (National Bureau of Statistics 2012), including substantive demographic information.
12. National one percent (inter-census) Household Survey, conducted between the ten-yearly national censuses, last conducted in 2005.

Notwithstanding recent attempts to improve the health MIS (HMIS), monitoring China's





**Fig. 6** Web-based national notifiable disease reporting since 2004 (Source: China Centre for Disease Control, Beijing (with permission))

HSR and health status relies largely on output-based reporting or describes numeric improvements emanating from high-profile national initiatives (Meng et al. 2012), often lacking denominators (Huang 2011; Yip et al. 2012; Ministry of Health 2012c). China does not have a tradition of locally representative, population-based surveys on health outcomes; those which are undertaken are almost never independent. The disaggregated impact of health initiatives and local health status remains unknown except at crude (regional and urban-rural) levels (Meng et al. 2012; Ministry of Health Centre for Health Statistics and Information 2009). This lack of data reduces the ability of governments to allocate resources according to local demography and disease epidemiology (which are changing rapidly with urbanization). In this context, quality implementation of new HMIS initiatives (Hipgrave 2011b) will be critical; however, again these are national initiatives reliant on local funding. The HMIS is mentioned as a priority for the second phase of the HSR (Ministry of Health 2012a), but in general the monitoring and evaluation of China's health sector remains weak and non-independent and is not prioritized at subnational level.

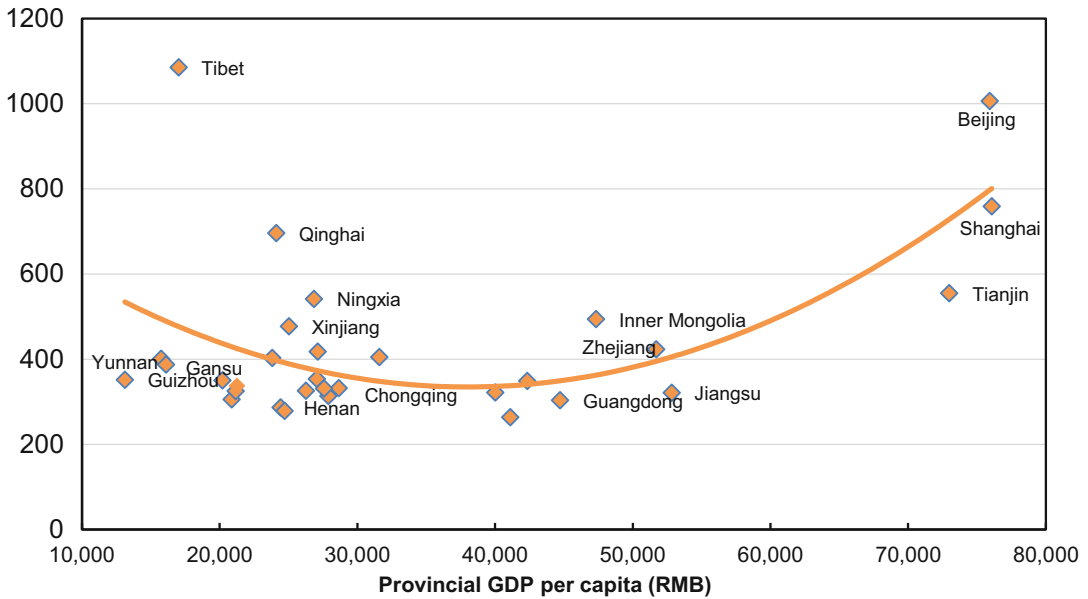
## Financing

### Sources of Funding and Accountability for Its Use

Subnational governments, even at county and township level, are responsible for about 90% of social sector financing and for the provision of essential services including health (National Bureau of Statistics 2011). Government expenditure on health depends heavily on local fiscal capacity (Yip et al. 2012; Wong 2010; Feltenstein and Iwata 2005); this varies widely across China, even after adjusting for formula-based "equalization transfers" from central government (Wong 2010; Bloom 2011). On average, tax revenue sharing and intergovernmental transfers finance up to 50% of subnational government expenditure (World Bank 2012).

This system bestows considerable power on provincial governments but also significant financial stress at the lowest levels of government. Each level of government has considerable discretion in transferring resources to successively lower levels. Provincial governments are the main recipients of the central government equalization

### Provincial expenditure on health per capita (RMB)



**Fig. 7** Provincial expenditure on health per capita in relation to provincial gross domestic product per capita, 2010 (Source: Ministry of Health, *China Health Statistics Yearbook*, 2011 (Ministry of Health 2012b))

grants and tax sharing and have significant autonomy in what they do with these funds. Prefecture governments in turn have similar autonomy. In this system, funding for public service delivery by poorer townships and counties tends to be insufficient (Wong 2010; Zhou 2010b).

Apart from earmarked transfers from the MoH and funds for selected nationwide priorities, local governments may withhold resources for lower levels or favor spending in more populous areas or on issues strategic to their career (Zhou 2010a; Liu 2007). This kind of bias at subnational levels can undermine progress on national development goals (Yang 2011; Uchimura and Jütting 2007).

To supplement resources received from the higher levels, subnational governments raise resources from various fees, the sale of land use rights, and taxes on real estate transactions (World Bank 2012). However, poor localities tend to have limited scope for such revenue generation. The imbalance between resources and expenditure responsibilities, particularly in poor jurisdictions, impacts on health service quality (Yang 2011) and on household health expenditure

(Blumenthal and Hsiao 2005; Meng et al. 2012; World Bank 2012).

Moreover, income disparities have widened across localities and population groups within local jurisdictions (Xing et al. 2008; Zheng et al. 2008; UNDP China and China Institute for Reform and Development 2008). The national urban-rural ratio of income per capita has risen from 2.4 in 1991 to 3.2 (up to 4 within certain provinces) in 2010 (Fig. 7) (National Bureau of Statistics 2011). At subnational level, only four provinces (Sichuan, Tibet, Xinjiang, and Yunnan) bucked this trend due to large subsidies to stimulate economic development and poverty reduction. Subsidies for these provinces impact the shape of the line of best fit in Fig. 2, which depicts provincial expenditure on health in relation to provincial GDP, per capita.

### Difficulties Using Available Health Financing for Policy Implementation

As mentioned, in China's decentralized environment, local government expenditures are not

aligned with policy priorities across sectors and programs. There are four distinct components of the national budget system, two of which impact on social sector spending: the general government budget (which relies on various taxation revenues and allocates funds to publicly funded services and activities) and the social security budget. The first of these allocates funds at the sectoral level; line ministries can then decide on and allocate earmarked transfers to the provinces (Wong 2010; Zhou 2010b). However, subnational government spending also relies on off-budget revenues (such as local taxes) for off-budget programs.

Monitoring is limited and there is little effort to align subnational budgets or plans with higher-level priorities. Moreover, apart from some individually monitored earmarked transfers, little information is available on whether governments actually spend money according to budgetary allocations or whether government expenditures and programs lead to the desired outputs and expected outcomes. Achievement of high-profile input and output HSR targets masks the absence of substantive analysis of outcome-level impact (Meng et al. 2012; Yip et al. 2012). Audits tend to focus on detecting malfeasance, not program performance.

Additionally, China's budget and expenditure cycles are not synchronous. The fiscal year starts with the calendar year, but the budget is not endorsed by the National People's Congress until the end of March. This delay reduces the budget's operational significance for subnational governments and central ministries (World Bank 2012). Fragmentation, information limitations, and delays in budget execution limit the ability of national authorities to transform policy priorities into resource allocation and results at the local levels (World Bank 2012).

### Health Expenditure and Sources of Revenue

Total health expenditure (THE) in China was US\$445.5bn in 2012, at US\$329 per capita, and 5.41% of GDP (China National Health Development Research Centre 2013). THE/GDP is modest

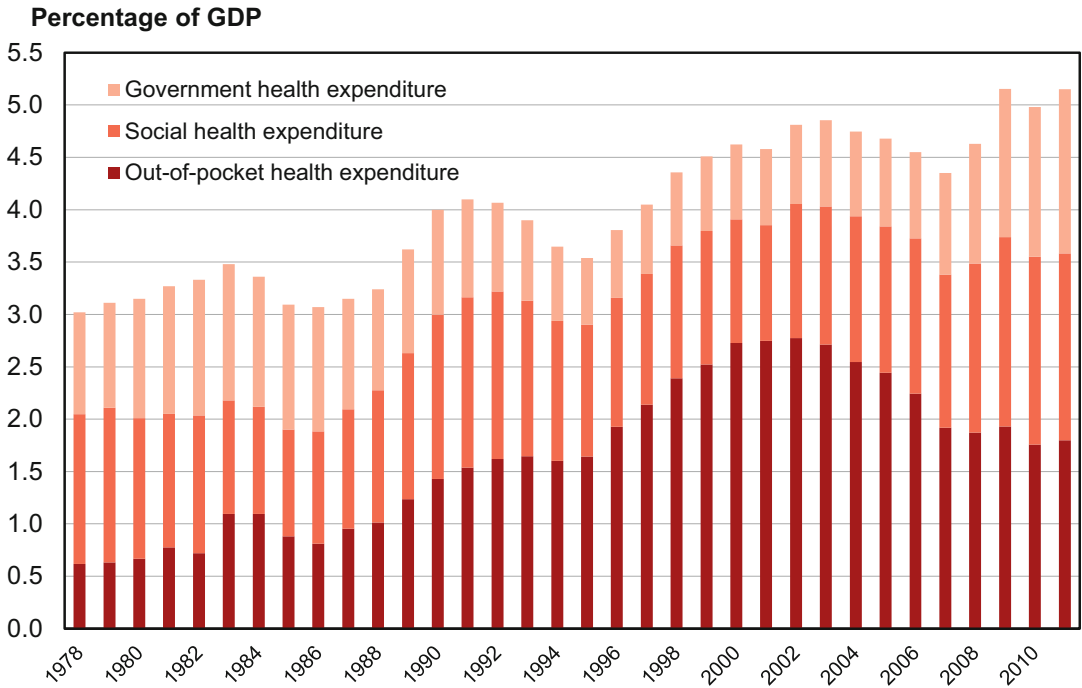
compared with industrialized countries, which averaged 9.7% in 2010 (OECD 2013), but is average among low- and middle-income countries (LMIC), whose THE/GDP ranges from 2.6% to 10% (e.g., Indonesia 2.6%, Thailand 3.9%, India 4.1%, Russia 5.1%, Vietnam 6.8%, South Africa 8.9%, and Brazil, 9.0%) (see data at <http://apps.who.int/nha/database>). Health expenditure as a proportion of GDP has increased from ~3% to ~5% since 1980, but numeric growth has been enormous due to China's rapid economic growth (Figs. 8, 9, and 10).

The sources of THE have changed dramatically over time, reflecting changes in the role of government. Marketization beginning in the 1980s led to historically high out-of-pocket expenditure in 2001 (60%), but this had decreased to ~34% in 2012 (China National Health Development Research Centre 2013), mostly through public subsidies for primary health programs, for health providers and for the social insurance schemes.

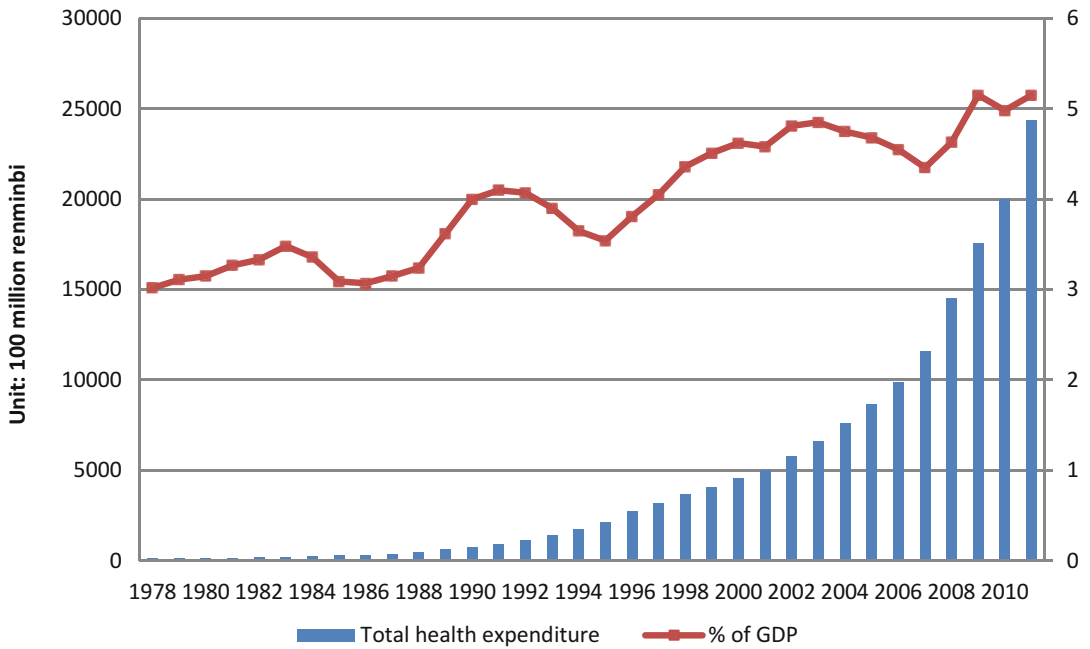
In 2011, tax-based government expenditures accounted for 30.7% of THE, social health expenditure 34.6%, and out of pocket 34.7% (Fig. 8). Overall, public expenditure on health as a share of THE is similar to that of many other LMIC and also to the United States (even higher if the government contribution to social health insurance is considered), but most high-income countries average around 71% (Tangcharoensathien et al. 2011). WHO calculates this figure differently and has China's figure at 56%; most nations in South and East Asia average around 41% (see <http://apps.who.int/nha/database> and Hipgrave and Hort 2014).

### Collection and Pooling of Funds

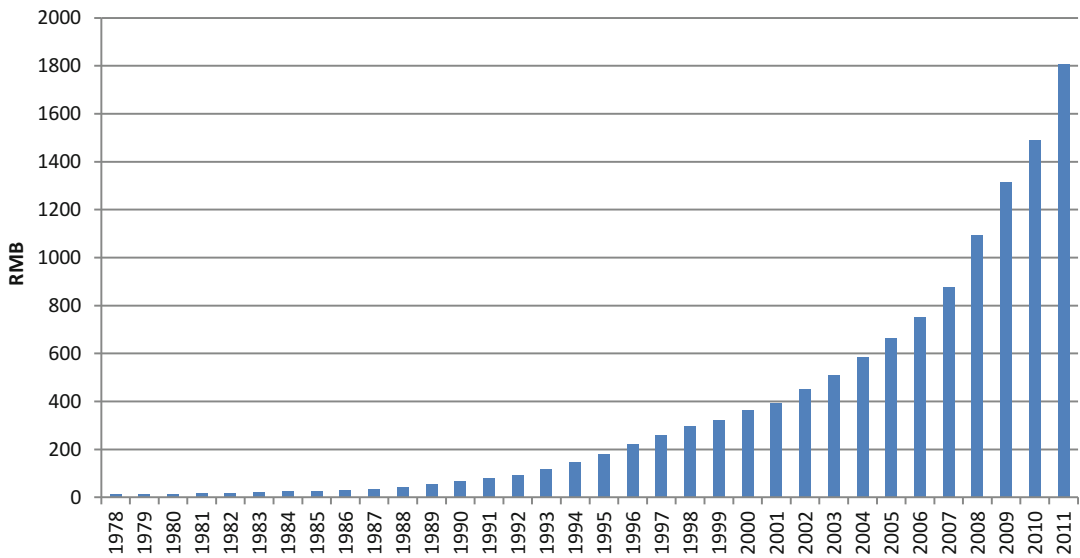
To provide essential health services, reduce inequity, and provide financial protection against catastrophic health expenditure, governments must mobilize sufficient resources via: (1) collecting revenues, (2) pooling of risk, and (3) purchasing goods and services (Gottret and Schieber 2006). Globally, three models of basic healthcare financing are practiced: nationalized health services, social insurance, and private insurance. China's



**Fig. 8** Government, social and out-of-pocket expenditure on health, 1978–2011 (Source: China Health Statistics Yearbook (Ministry of Health 2012b))



**Fig. 9** Total health expenditure (THE) in China, numerically and as a percentage of gross domestic product (China National Health Development Research Centre 2012) (2012: US\$1 = ~6 renminbi [RMB]) (Source: China National Health Accounts Report 2012)



**Fig. 10** China's per capita THE (Source: China National Health Account Report 2012 (China National Health Development Research Centre 2012))

healthcare financing has evolved to a structure dominated by three social insurance schemes with almost universal population coverage: the urban employees basic medical insurance (UEBMI) (financed by formal sector employers and employee contributions), the rural cooperative medical (insurance) scheme (RCMS), and urban residents' basic medical insurance (URBMI). The latter two receive heavy government subsidization in addition to individual contributions (in a roughly 4:1 ratio).

Government health expenditure stems from tax revenue, as described above. China does not have tax instruments specifically designated to health expenses; the funds are allocated from overall tax revenue. These funds are used to pay the salaries of health workers, purchase equipment, and build infrastructure at various levels and for various specific programs such as public health subsidies or other schemes earmarked by the MoH. Government also funds a social assistance program (the medical financial assistance scheme), which provides cash for designated poor households to purchase health services. There also remains "free medical treatment" for those on the government payroll and for retired military and Party cadres;

these arrangements are slated for phasing out. However, government does not as yet contribute substantively to the funding of hospital care, which remains predominantly managed in-house from various sources of revenue (in particular, out-of-pocket payments and insurance) (State Council 2014; Barber et al. 2014).

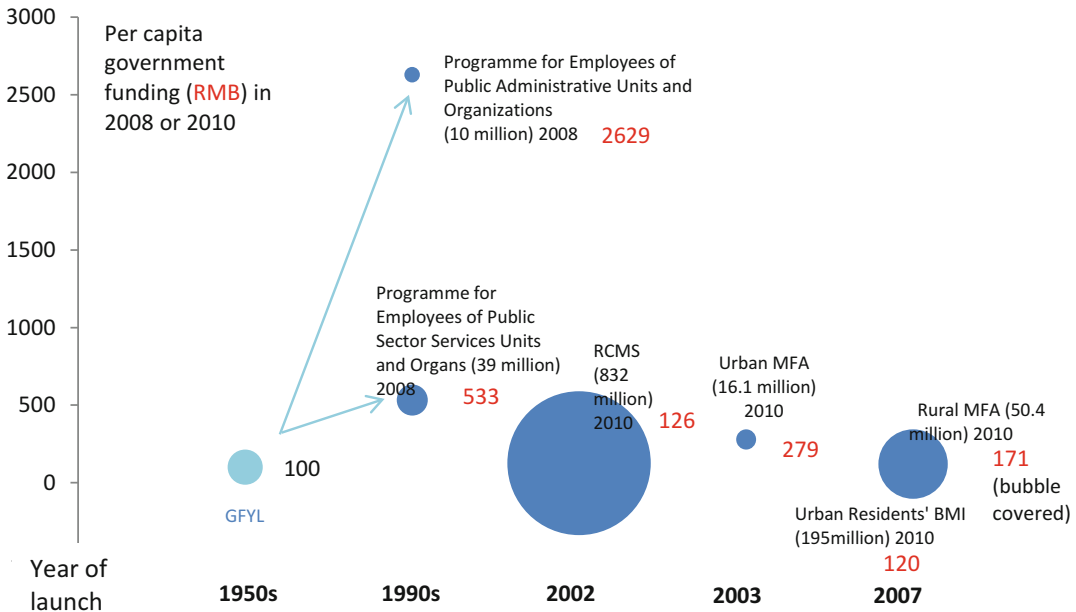
### Coverage, Benefit, and Cost Sharing

Table 1 summarizes the current basic health financing arrangements and benefit provided by the various health insurance schemes in China. It is evident that the major challenge remains fragmentation of the schemes and arrangements and the associated inequity and inefficiency. This is also highlighted in Fig. 11, which depicts the large variation in average numeric benefit and other information about the various schemes. In this context, and given China's highly mobile population and the limited access of migrant populations to urban health services (Di Martino 2011), the Government is prioritizing integration of the various insurance schemes (Ministry of Health 2012a), but this is a difficult and complex proposition.

**Table 1** Current healthcare financing arrangements, coverage, and benefit in China

	Three social health insurance schemes			Direct subsidizing of public providers	Medical financial assistance	Free healthcare	Other (private) insurance	Out of pocket (direct payment for care as needed)
	UEBMI	RCMS	URBMI					
Beneficiaries	Formal sector workers	Farmers	Urban residents not covered by the UEBMI	All citizens	Poor households with catastrophic health expenses or recognized recipients of China's social security payments	Public sector employees and special groups such as retired military and Party cadres	Voluntary purchasers of private insurance	Persons not covered by another scheme or having to choose a different health provider (includes many migrant workers)
Population coverage	14.8%	69.5%	9.5%	All citizens	809 million incidents funded in 2011	0.7%	0.3%	5.2% (excludes most migrant workers)
Benefit covered	Outpatient (OP) and inpatient (IP)	Mainly IP, but some counties experiment to cover some care for OPs (e.g., treatment of chronic diseases)	Mainly IP, but some cities experiment to cover some care for OPs (e.g., treatment of chronic diseases)	All services	IP incurring catastrophic cost to individuals	OP and IP	OP and IP	
Co-payment arrangement	Practices vary across regions; co-payment occurs for (1) expenses below scheme thresholds and also above ceilings, (2) expenses on high-end or special services excluded from schemes, and (3) the percentage not reimbursable for different services OP roughly 30–50%; IP: 10–20%	IP roughly 50–60%	IP roughly 60–70%	N.A.	Practices vary across regions	Very minimal	Varies across schemes	100%

Source: Authors' own compilation; figures from the 2012 China Health Statistics Digest (Ministry of Health 2012b)



**Fig. 11** Government financing per participant across health security schemes introduced during 1950–2007. Note: Bubble size is equivalent to the number of participants. Number of participants is shown in parentheses. Government spending per participant is shown in red.

Government funding figures are annual per person, except for the rural and urban medical financial assistance, reported per case (Source: National Health Account Report 2011 and China Health Statistical Digest 2011)

## Payment Methods for Health Services

Before the HSR, to ensure financial accessibility, the Chinese government priced primary healthcare services at below cost, but allowed providers to charge high prices for diagnostic tests using high-tech equipment, effectively cross-subsidizing primary services. Providers could also levy a 15% profit on drug sales. Under the prevailing fee-for-service payment modality, this created an incentive for providers to maximize profit by ordering tests and over-prescription of drugs. Cost-effective and efficient primary healthcare services were ignored by providers because they were not profitable; those who could not pay for services often chose to forego them (Tang et al. 2008).

The recent reforms to provider payment, and those mooted for the future, aim to: (1) encourage the provision of cost-effective and efficient primary healthcare services, (2) reduce provider reliance on drug income and curb over-prescription, and (3) curb cost inflation.

Innovative provider payment methods, such as capitation (for primary health mostly), gross budget, diagnosis-related groups (for hospitals), as well as performance-based payment for health workers, are being piloted at county and district level. Other related policy reforms include a zero markup policy (for essential drugs), implementation of essential drug list, and so on (Yang et al. 2013a).

## Physical and Human Resources

### Infrastructure and Its Funding

By international standards China's average health infrastructure level has been poor. For example, the number of hospital beds per 1000 population in 2011 was around 4, among the lowest in the world (Ministry of Health 2012b). Health infrastructure in China also suffered from a major urban-rural divide in the earlier stages of social and economic development. Not only did urban

health infrastructure enjoy greater public financial support, it attracted loans and other financial instruments because it was profitable and boosted the local economy. For many years, rural facilities received very limited government subsidy and relied on collective funding among farmers. Rural health infrastructure lagged seriously, in terms of both the basic condition of health facilities (buildings, beds, etc.) and the equipment, while big urban hospitals acquired technical equipment of high quality. In 2005, there were 3.6 hospital beds per 1000 urban residents, but only 0.78 in rural townships (Ministry of Health 2007). This inequity was recognized by national government, and in 2006 the majority of a national bond issue was used to finance a project earmarked for rural health, specifically to finance the rebuilding, renovation, and updating of medical equipment for rural providers, including primary health facilities such as CDC and MCH institutions. The NDRC and its local branches approved the funding proposals for physical health infrastructure. More recently, the 2009 HSR allocated large sums to further improve physical health sector infrastructure (focusing on rural remote rural areas, but also urban community health centers). Progress on this aspect of the Reform has been very positive (Yip et al. 2012).

### Health Workforce and Trends

For the majority of China's population, access to western and formally regulated traditional Chinese medicine (TCM) only commenced with the introduction of China's famed "bare-foot doctors" in the mid-1960s. These cadres numbered 1.8 million at their peak (around one per 600 people), but numbers fell rapidly with economic marketization and liberalization of population movement (Bien 2008). Moreover, village-level care lost its funding base with the dismantling of the rural cooperatives in the early 1980s, and training and supervision of the quality of care provided fell off. As recently as the late 1990s, many doctors lacked training to the level suggested by their rank and title (Youlong et al. 1997), and overprescribing of

drugs and inappropriate use of parenteral preparations continue to exemplify the low quality of care, especially in rural areas (Blumenthal and Hsiao 2005; Bloom and Xingyuan 1997; Zhan et al. 1998; Pavin et al. 2003; Dong et al. 2008; Chen et al. 2010).

With economic marketization, medicine at all levels became privatized, physician salaries were paltry, standard consultation fees were fixed below cost (Eggleston et al. 2008), and over 40% of doctors' and health facilities' income derived from the sale of drugs (Hu 2010). As a result, doctors worked where they could be assured of income, patients became disillusioned with the care at rural clinics, self-referral to urban clinics increased, and the distribution of doctors, nurses, and health facilities was heavily biased to urban areas (Yip et al. 2012; UNDP China and China Institute for Reform and Development 2008; Youlong et al. 1997; Anand et al. 2008) (Table 2). Residents of urban areas in China, particularly in the large eastern cities, enjoy physical access to health services to the same level as in most developed nations. However, like many other Asian nations, China has trained more doctors than nurses or midwives, and there are progressively fewer staff with formal health training in progressively poorer rural areas (Youlong et al. 1997; Anand et al. 2008) (Table 3). China includes TCM practitioners (13%) in headcounts of health staff (Anand et al. 2008).

China is still paying for the interruption of university education during the Cultural Revolution of 1966–1976, and the paucity of new village doctors trained since the breakup of the village cooperatives in the late 1970s. First, as of 2005, 67.2% of China's doctors and 97.5% of nurses had only completed junior college or secondary technical school level training, and 6% and 8% respectively had just high school or lower education (Anand et al. 2008). The duration and standard of professional education varies widely across the country (Youlong et al. 1997). Village doctors are an ageing cohort, with a likely high attrition rate in the coming decade (Xu et al. 2014).

However, with massive increases in the number of formal trainees since 1998, the distribution and quality of personnel are probably bigger



**Table 2** Health workers in China in 2011

Categories	Total		Urban		Rural	
	Number (1000s)	Density	number (1000s)	Density	number (1000s)	Density
All health workers	8616	4.58	3844	7.9	4762	3.19
Licensed doctors including assistant doctors	2466	1.82	1190	3	1275	1.33
Nurses	2244	1.66	1304	3.29	939	0.98
Other health professionals	1492	1.1				
Other health workers	2413	1.8				

Note: Urban areas refer to jurisdiction under China's four municipalities and prefecture-level cities. Rural areas refer to counties and county-level cities, as well township hospitals and village clinics. Density refers to the number of health workers per 1000 population

Source: China Health Statistics Yearbook 2012 (Ministry of Health 2012b)

**Table 3** Distribution of doctors and nurses by education level and health institution type, in 2011

	In hospitals (%)		In community health centers (%)		In township hospitals/clinics (%)	
	Doctors	Nurses	Doctors	Nurses	Doctors	Nurses
University and above	62.7	11.5	31.7	5.7	3.9	0.4
Secondary school and college	36.3	86.4	64.6	91	83	87.9
High school or less	1	2.1	3.7	3.3	13.1	11.7

Note: University and above refer to with at least a bachelor's degree. Secondary schools include technical or professional high schools

Source: China Health Statistics Yearbook 2012 (Ministry of Health 2012b)

problems than the overall number of China's health human resources. Indeed, some data suggest an excess of trainees and the likelihood that many health graduates do not take up professional service. Nonetheless, inequality and inequity in the distribution of doctors and especially nurses between and particularly within provinces remains extreme and has been linked to key health outcomes including infant mortality (Anand et al. 2008).

Authorities in China recognize the prevailing inequity in distribution of health human resources and have initiated training and other schemes to increase the number of qualified personnel and improve their distribution. The 12th Five-Year Plan for health sector development, released in 2012, sets targets for assistant physicians (1.88/1000 population) and nurses (2.07) and lays out plans for increased priority of staffing in rural areas and at community level, of personnel and financial support for poor rural and western health facilities by wealthier urban and eastern facilities,

of intensive efforts to fill known human resource gaps among various health and allied health providers, and of tiered registration for doctors that first requires a period of rural service. A focus on community general practice is reiterated in the plan, with a target of 150,000 staff newly trained or upgraded personnel to provide such services.

In addition, in a 2011 "Guidance" the State Council announced new roles for village doctors, recommending a wide range of tasks (Government of China 2011). By 2020, these cadres should be providing standardized primary care (following new clinical guidelines), implementing public health programs, undertaking disease surveillance, conducting community education, participating in health financing schemes, and maintaining individual e-health dossiers. In theory, it will be possible for the national HMIS to monitor their work. The official engagement of village doctors in a national system is positive development and should improve public confidence in their services. However, payment for

the planned elevation of village doctors' responsibilities will derive from a complex mix of funding streams (Government of China 2011; Ministry of Health 2011b) overseen and additionally funded by county-level authorities (Government of China 2011) whose accountability for this national initiative will be to local government (Wong 2010; Zhou 2010b), not health authorities.

### Remuneration of Health Workers

It is well established that marketization and the de facto privatization of clinical care by salaried doctors working in public facilities had, by 2000, resulted in China having one of the least equitable health systems in the world (The World Health Organization 2000), with over 60% of THE being out of pocket (Blumenthal and Hsiao 2005; Ho and Gostin 2009; Wang et al. 2007). One of the main objectives of China's HSR is to regulate the remuneration of doctors and to separate their income from choices on clinical care. However, while China has reduced the level of out-of-pocket expenditure on health to around 35% through increases in public funding and insurance initiatives (Yip et al. 2012), household health expenditure has not decreased either numerically or as a proportion of total household expenditure (Meng et al. 2012). Although there is indirect evidence of increased non-health expenditure by insured households in comparison to before the schemes were introduced (Bai and Wu 2014), this objective of the HSR is proving to be the most difficult to achieve. China's THE is increasing at around 17% per year, and a large proportion of the increase is due to payment of health facilities, doctors, and other providers by individuals or insurers. As patient expectations rise but out-of-pocket expenses remain numerically high, an increasing number of assaults of doctors by patients' families are being reported.

On the other hand, the scheduled fees payable to doctors for listed services are set below cost, forcing clinicians and facilities to charge for other services, investigations, procedures, and drugs (including those not on the essential drugs list with unregulated prices) (Blumenthal

and Hsiao 2005; Ho and Gostin 2009; Wang et al. 2007; Tian et al. 2008) or through accepting bribes and kickbacks (Yang and Fan 2012). While the government has committed to improving both the quality of care provided by health providers, and is exploring remunerating them through capitation, diagnostic-related groups and performance-based incentives (Ministry of Health 2012a), separating hospital management from doctors' income is proving to be the most difficult element of the current HSR (Yip et al. 2012).

---

## Health Services Delivery and Outcomes

### Primary Care and Public Health

As reviewed elsewhere (Hipgrave 2011a), public health services in China suffered badly under the marketization of the 1980s and 1990s. CDC in particular was weak, culminating in the SARS epidemic in 2003. Public funding for preventive health services fell dramatically and was insufficient to even cover salaries. Public health authorities were left to raise their own income through charging fees for services, including vaccination (for which fees were only completely dropped in 2007) and various inspections and screening. Community approaches to disease control were abandoned in favor of vertical programs reliant on national or external funding, and disease surveillance was poor.

SARS and health authorities' realization of the epidemic of NCDs due to ageing, urbanization, and decreasingly active lifestyles has led to major changes to public health programming in China. Disease surveillance is now conducted online, in real time, and funding for CDC and preventive health has increased dramatically. New vaccines were introduced in 2008, although globally recommended vaccines against *Haemophilus influenzae* type B, pneumococci, human papilloma viruses, and rotaviruses are only available privately (ironically, through government providers).

The largest boost to public health came with the 2009 HSR, when government introduced a

minimum 15 renminbi (RMB)/capita subsidy for public health/screening activities to be conducted across the nation. This had been pre-dated by various vertical preventive health programs, such as funding of hepatitis B vaccine since 2002 (Cui et al. 2007) and national funding of the EPI since 2007. The HSR public health funding is provided by a mix of national and local authorities according to their ability to pay (problematic for poor counties in rich provinces) and the RMB15 was increased to RMB25 in 2011; it is much higher in wealthy areas. The funds pay providers to conduct the following services, notionally free of charge: (1) maintenance of individual electronic health records, (2) health education, (3) vaccination, (4) infectious diseases' prevention and treatment, (5) screening and management of chronic diseases such as hypertension and diabetes, (6) mental healthcare, (7) child healthcare, (8) pregnancy and maternity care, and (9) healthcare for the aged.

For the elderly and those with chronic diseases, this kind of screening, along with the introduction of zero markup and full reimbursement for drug treatment of NCDs (Yang et al. 2013a), has made a huge difference to their care. However, rollout of this initiative is slow, and although most targets are being met (Yip et al. 2012), monitoring is hampered by the absence of local denominators. Moreover, some of the programs, such as management of mental illness, have not been founded upon a training program for staff ill-equipped to provide them. In addition, unpublished evidence gathered by UNICEF in 2010 suggests that some of the funds are being used as salary supplements to support the new responsibilities of village doctors (in public health and other programs) and that the volume of money allocated to some rural localities is actually too high, due to out-migration to cities. Meanwhile, the increasing proportion of China's population living in urban areas, including most rural-urban migrants, cannot access such services.

Another boost to public health came with the MoH's program, also introduced in 2009, to prioritize interventions for certain vulnerable populations. These include: (1) catch-up hepatitis B vaccination for those aged <15 years;

(2) cervical and breast cancer screening for women in rural areas; (3) an expansion of the hospital delivery subsidies first introduced in 2000, to cover women in all rural counties; (4) free cataract surgery for the poor; (5) free folic acid supplementation for rural women before and during pregnancy; (6) improved stoves and fuel to reduce fluorosis; and (7) introduction of eco-friendly toilets. Again, targets for introduction of these measures have been set and rollout is proceeding (Yip et al. 2012).

Finally, although firm evidence of impact is scant, local authorities in most Chinese cities have introduced public education and health literacy programs to enhance awareness on issues like diet, exercise, cigarette smoking, appropriate care of women before and during pregnancy, infants and young children, and the elderly. As usual, implementation of national guidelines on such activities depends on uptake and funding by other sectors and local authorities. The regular occurrence of outbreaks of food (Xinhua 2011) and environmental contamination (Human Rights Watch 2011) and other scandals with public health implications indicates the difficulty faced by national authorities in China's decentralized context.

## Clinical Services

Recent high-profile summaries of China's health system tend to focus on its administration and financing and neglect the considerable improvements in clinical care available to the local population. While standards at all levels of the service hierarchy vary very widely, health authorities have augmented the care available at virtually all public facilities across the nation. Moreover, access to services to services has improved for all the population, albeit at high cost to both government and individuals (Meng et al. 2012).

Clinical services in China are conducted through a hierarchically arranged network of facilities ranging from tertiary referral centers in the large cities (most having high-quality diagnostic and laboratory equipment) to second-tier hospitals at county and district level. Rural townships

**Table 4** Number of outpatient visits and inpatients in health institutions in China in 2011

Health institution type	Total visits (100 million person-times) ( <i>n</i> = 62.7)	Total inpatients (10,000 persons) ( <i>n</i> = 15,298)
Hospitals <i>n</i> (%)	22.6 (36)	10,755 (70.3)
General-acute hospitals	16.74	8431
Hospitals specialized in TCM	3.61	1349
Specialty hospitals	1.88	844
Sanitaria	0.05	98
Community health institutions (%)	38.05 (60.7)	3775 (24.7)
Health centers	8.8	3472
Urban health centers	0.11	23
Rural township hospitals	8.7	3449
Outpatient department	0.7	13
Clinics, health centers, and nurse stations	5.2	
MCH centers (stations) <i>n</i> (%)	1.76 (2.8)	682 (4.28)
Specialized disease prevention and treatment institutes	0.2	38

Source: China Health Statistical Yearbook 2012 (Ministry of Health 2012b)

and urban communities are served by clinics or hospitals with varying capacity for inpatient care and surgery. At village or neighborhood level, public or (mostly) private facilities provide basic outpatient care, usually with an attached dispensary and possibly with links to a laboratory or radiology service. Concern about the standard of care provided by local facilities has resulted in many patients self-referring to higher-level facilities and hospitals (Table 4). As a result, hospitals in China tend to provide care for all level of illness, resulting in inefficiency and overcrowding. Expenditure on hospital-based care as a proportion of THE in China far exceeds that in many OECD nations (Barber et al. 2014), resulting in the high priority given to improving

primary care, community general practice, and lower-level facilities in the HSR (Yip et al. 2012; Ministry of Health 2012a) and to moving outpatient care in particular from hospitals to primary care facilities (Barber et al. 2014).

As would be expected for a nation of this size and variation, clinical services in China vary widely, from the world-class care available to residents in Shanghai, Beijing, Guangzhou, and similar cities to the most basic care in rural clinics in far western China. Similarly, models for the care of chronic illness and the use of day-care and hospital in the home vary widely, but in general these options are not yet well developed in China. The average length of inpatient stay is high in China compared to OECD nations (Meng et al. 2012), particularly in public hospitals, which account for 89% of total beds and 92% of hospital admissions (Barber et al. 2014). Clinicians at community level have usually had training in TCM and many practice both western medicine and Chinese medicine.

However, the preparedness of clinicians in primary care for the wide range of conditions they treat varies widely. For example, China's current HSR acknowledges that the system's clinical focus has been ill-suited to the screening and outpatient care of chronic illness, an increasing priority as rates of noncommunicable diseases rise (The World Bank Human Development Unit 2011). Similarly, the high-volume model of clinical care in China is poorly suited to the management of mental illness (Qin et al. 2008), aged care and dementia, and prevention of tobacco-related illness and alcohol consumption, all of which are needed in China (The World Bank Human Development Unit 2011; Phillips et al. 2009; Yang et al. 2013b; Zhou et al. 2011; Chan et al. 2013).

With respect to quality of care, in the last decade China has moved to standardize many clinical pathways and practices, and the concept of evidence-based medicine is increasing. However, attention to such standards and their influence on clinical care is perceived to be low (Yang and Fan 2012). Moreover, funding for and the quality and independence of clinical research, access to information, and the ability of clinicians to practice independent of the profit motive are

major obstacles to the use of evidence-based guidelines in clinical care in China (Barber et al. 2014; Wang 2010).

## Pharmaceutical Care

China's pharmaceutical sector has been one of the most problematic for health authorities over recent decades and the focus of major reform efforts in the last few years. In 2008, 42.7% of China's THE was on drugs (Hu 2010), compared to 17% in developed nations (Seiter et al. 2010). Excessive drug prescription was common in rural China (Zhan et al. 1998; Pavin et al. 2003; Dong et al. 2008; Chen et al. 2010; Yu et al. 2010), and there is evidence that China's rural health insurance scheme was encouraging over-prescription (Chen et al. 2010; Sun et al. 2009). Drug sales continue to provide the largest income source for China's county health facilities; doctors have a pecuniary incentive to prescribe more and more expensive drugs (Chen et al. 2010; Yu et al. 2010). Hospitals and doctors profit significantly from the sale of drugs (Yu et al. 2010; The World Bank Group East Asia Pacific Region 2010), affecting financial access to healthcare (Tang et al. 2008; Meng et al. 2012). Weak regulation of drug manufacture and distribution raises safety concerns (Yu et al. 2010; Guan et al. 2011).

Previous efforts to improve the pharmaceutical sector had limited effect. The impact of laws, decrees, and 24 separate price reductions over 1996–2007 was constrained by hospital financing/income generation, market influences, and patient preferences (Chen et al. 2010; Yu et al. 2010). Price controls were undermined by manufacturers, wholesalers, and retailers and by hospitals and physicians controlling the prescription of price-controlled drugs (Hu 2010; Yu et al. 2010; Chen and Schweitzer 2008). New drug approvals were issued at astonishing rates (Ho and Gostin 2009) and the former head of the national drug administration authority was executed in 2007 for accepting bribes. Kickbacks and corruption continue to mar the sector (Yip et al. 2012; Yang and Fan 2012).

Acknowledging these problems, China's HSR included establishment of a National Essential Medicines Scheme (NEMS) to improve population access to and reduce the cost of essential medicines (State Council 2009), particularly at grassroots (township and village) level. The Scheme covers drug production, pricing, distribution, procurement, prescribing, and payment (Hu 2010) and a new National Essential Drugs List (NEDL) for primary healthcare institutions. The 2012 NEDL comprises 317 western drugs and 203 TCM commodities (increased from 205 western and 102 in 2009) for storage and use by grassroots facilities. Bidding prices for 296 NEDL drugs were capped (Schatz and Nowlin 2010), and a "zero markup" (no profit) policy was introduced, although markups remain allowed at county-level and higher facilities. By late January 2012, 99.8% of township hospitals and 58.1% of village clinics had implemented the policy (Ministry of Health 2012d). In addition, most (urban) districts and (rural) counties had made NEDL medicines reimbursable by the various health insurance schemes, with higher reimbursement rates than for nonessential medicines (Ministry of Health 2011c). Finally, to regulate the pharmaceutical market and distribution of essential drugs, the NEMS introduced province-wise, collective, internet-based public bidding and procurement for NEDL medicines.

These four elements – the NEDL, zero markup, reimbursement of certain drug costs by insurers, and public procurement – were designed by the government to wrest control of the public pharmaceutical sector from the private sector. However, the official HSR documents encourage local adaptation of the broad design (Ho 2010), including the NEDL (which has indeed been widely augmented (Guan et al. 2011; Shi et al. 2011)) and strategies to compensate providers for the zero markup policy. Few evaluations of the impact of the Scheme have emerged. Very early indications suggested little change in prescribing practices (Yip et al. 2012), but a small field evaluation found that while drug procurement has been systematized and the cost of care had declined coincident with reduced drug prices, manufacturers have not

uniformly supported the changes, and some drug prices have actually increased. Provider compensation for reduced income was mostly ineffective, forcing some to seek alternative sources of income within and outside the health sector. Rational drug prescribing had improved in this study. The loss of drug income had forced health facilities to rely more on public financing, and providers complained of higher workload and lower incomes (Yang et al. 2013a). Similar issues were found in another study in different locations (Xiao et al. 2013).

The NEMS particularly impacts small rural health facilities and will again rely on considerable local support for its implementation. Meanwhile, provinces are continuing to augment even a revised version of the NEDL (Tang et al. 2014a), and zero markup has not yet been applied in county or higher-level facilities. While insurance reimbursement and capitation may help to improve prescribing practices and reduce patient outlays, more control of procurement, manufacturer, and prescriber practices are required.

The recently announced reforms of county hospital funding and administration include a major focus on drug procurement, prescription, management, and pricing (State Council 2014).

## Private Healthcare

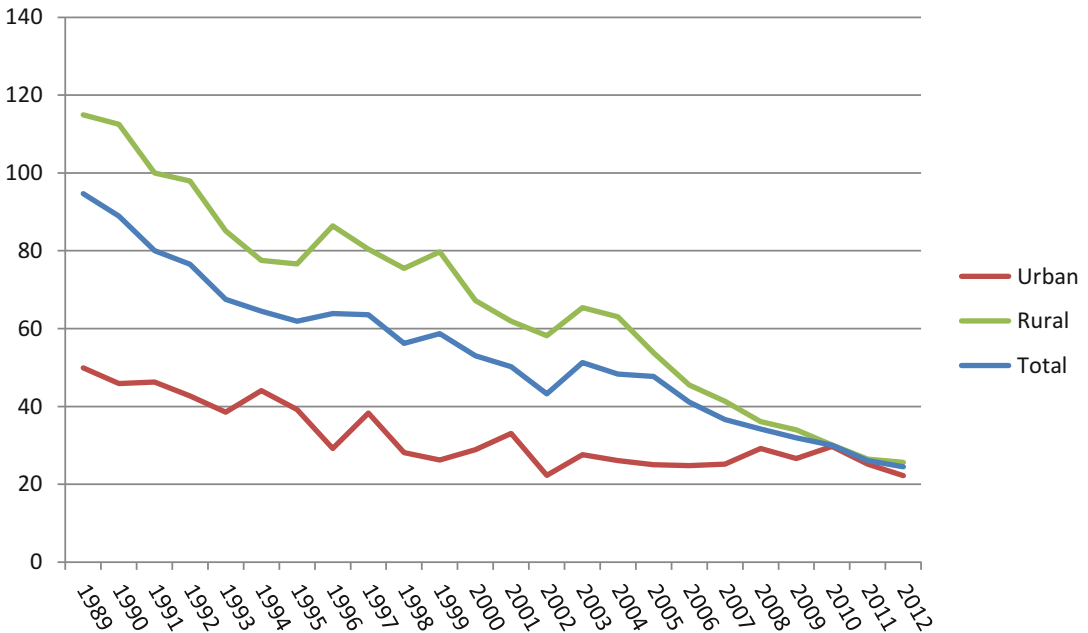
As a consequence of the marketization of China's health sector in the 1990s, provision of health services was opened significantly to private providers. The number of private providers increased rapidly and now comprises a significant proportion of the market. For example, in 2005, private hospitals accounted for only 17.2% of total hospitals, but the share had increased to 38.4% by 2011. In 2011 among all 954,389 health facilities (hospitals, clinics, and other institutions), 47% operated as "private" entities. Reports indicated that private health providers can offer services at a cheaper price and shorter physical distance and waiting time for patients (Deng et al. 2013) and are highly active in the provision of healthcare in China. However, most private facilities are small and poorly equipped, and collectively they only

employed 17.5% of the total labor force, owned 9.7% of total medical beds, and received 9.1% of total patient hospital visits (Ministry of Health 2012b). Compared with public facilities, a large percentage of elderly physicians and new laborers in health market are practicing in private clinics (Tang et al. 2014b). This staffing structure could have negative impact on quality of services.

In general, despite rapid development in recent years, private health services are at an early stage of development in China. One major reason is that the evolution and current standing of national policy generally still favors public providers in terms of resource allocation, stewardship (entry and registration control), opportunities for promotion, and social insurance entitlements. This accounts for common challenges in the private sector, i.e., lack of technical capacity, poor infrastructure, and thus compromised service quality. Health authorities are now promoting a robust private sector to encourage competition and efficiency within the health sector, aiming for 20% of beds and services to be privately provided by 2015. However, subsidization of grassroots level public institutions may prevent moves in this direction.

## Health Outcomes

While China's progress on major health indicators during the 30 years immediately following the foundation of the PRC is unparalleled (Jamison et al. 1984), marketization and the unaffordability of healthcare for a large proportion of the population stymied progress in the 1980s and 1990s. There are even suggestions that child mortality rates in China actually rose in the 1980s (Banister and Hill 2000), with the breakup of the commune-based health cooperatives. Moreover, improvement in certain indicators has been slow. For example, urban maternal mortality has been slow to fall, almost certainly because reductions in maternity risk for urban residents have been diluted by the much higher risk of death in pregnancy among urban migrants (Fig. 12) (Zhang et al. 2014). Geographic disparities also remain great, particularly between eastern and western



**Fig. 12** Maternal mortality per 100,000 live births by urban-rural location (Source: China Health Statistics Year Book (Ministry of Health 2012b) and NHFPC (China National Health and Family Planning Commission 2012))

provinces (Wang et al. 2012). In general, the priority given to China's recent HSR acknowledges that progress in its population's health status was less than could have occurred, given the nation's economic growth since the 1980s (Yip et al. 2012). Acknowledgement of this is the government target of a one-year increase in life expectancy by 2015 (Ministry of Health 2012a). The most comprehensive analysis of the causes of death and disability in China, published in mid-2013, highlighted the dramatic evolution of its demographic transition, with NCDs now making up all but two of the top 30 causes of lost life years, and most infectious diseases having fallen precipitously. The report also noted the contribution of air and household pollution to mortality and morbidity and the need for cross-sectoral action to tackle the major causes of ill-health in China (Yang et al. 2013b).

Nonetheless, in 2010, average life expectancy in China was 74.8 years, and in 2012 the maternal mortality ratio was 24.5/100,000 live births, infant mortality rate 10.3‰, and under-five mortality rate 13.2‰ (China National Health and Family Planning Commission 2012). These figures

compare favorably with other developing countries, and China's performance in reducing rural maternal and neonatal mortality has been outstanding (Feng et al. 2010b, 2011). China has already achieved all the health targets in MDGs 4, 5, and 6 and achieved the target on reducing child underweight in the early 2000s. Urban-rural disparity in under-five and particularly maternal mortality has declined since 1990, but remains high for child underweight and stunting and especially for child micronutrient deficiency (UNICEF China, unpublished data; (Hipgrave et al. 2014)).

Challenges to population health status have been alluded to already and include the rise of NCDs, especially smoking-related illness (The World Bank Human Development Unit 2011), illness due to environmental damage and air pollution (The World Bank Human Development Unit 2011; Millman et al. 2008), urbanization, and the provision of services for newly arrived migrants (Gong et al. 2012). The prevention of accidents and injury will also play an increasing role in maintaining China's trajectory on reducing preventable death and ill-health (Wang et al. 2008b). As the population ages,

private and institutional care of the elderly is another major issue for China's health and other social sectors.

---

## Assessment

China's progress in maternal and child health, urban health, and communicable disease control are very encouraging, but the nation's health system now faces a vastly different range of issues from those it faced before.

In addition to health insurance reforms that commenced in 2003, in many ways the comprehensive health system reforms announced in 2009 have been highly successful. Insurance coverage is almost universal, and the benefit package is gradually expanding, even for outpatient services, although a system for ensuring coverage for the huge population of rural-urban migrants remains under development. Introduction of public health screening and management, building of new health infrastructure and expansion of community-based services, measures to control profiteering from the sales of drugs, scale-up training of health personnel, and other measures were both needed and are being implemented. On the other hand, the reform of hospital management and financing remains at the pilot stage, with suggestions but no formal guidance on the model to be followed.

China's HSR is encouragingly specific but not prescriptive on strategy. Monitoring the reform remains predominantly output-based at macro-level; no detailed independent assessments have been undertaken, and population-level studies of health outcomes related to the reforms have not been undertaken. Moreover, mechanisms to incorporate patient feedback into health service provision have not been established and may be ignored if local economic, political, or vested interests override such input, as has been observed in relation to China's natural environment. Public financing of the health sector, although modest by global standards, has improved, particularly in relation to the proportion of THE that is out of pocket. But costs are rising faster than government inputs, and poorer constituencies

remain least able to fund public services, despite having the greatest needs. As a result, proportional household expenditure on healthcare has not declined.

Urban residents of China's industrialized eastern provinces enjoy a high quality of healthcare and access to trained personnel. This is not the case for poorer rural residents, particularly in the nation's vast western region. The official engagement of village doctors to provide publicly funded health services in rural areas should improve the standard of and public confidence in their care, but the burden on this ageing cadre of staff is rising and may be untenable; again, accountability for this national initiative will be to local government and health officials unused to the application of treatment algorithms, performance-based assessment, and clinical audit. Concern about the care provided by community providers continues to result in many patients self-referring to higher-level facilities and hospitals.

Population health in China is threatened by the rise of NCDs, especially illness due to diabetes, cardiovascular disease, overweight, tobacco smoking, environmental damage, and air pollution. The prevention of accidents and injury and management of mental illness will also play an increasing role in maintaining China's trajectory on reducing preventable death and ill-health. The required focus of the health sector on chronic illnesses, aged care, and outpatient services requires a dramatic increase in the engagement and stewardship of community providers.

This has been a major focus of China's health reforms, now well into their second phase, and it is likely that further major policy and financial inputs will be announced before this phase concludes in 2015. The private sector will play an increasing role in the provision of health services in China, but a higher level of stewardship and the use of financial mechanisms to reign in escalating costs will almost certainly be required, especially for hospital care. To ensure consistency and transferability, this may involve stronger oversight by and involvement of national health policy and financing authorities, notwithstanding the power vested in subnational authorities in China's system of government.



## References

- Anand S, Fan VY, Zhang J, Zhang L, Ke Y, Dong Z, et al. China's human resources for health: quantity, quality, and distribution. *Lancet*. 2008;372:1774–81.
- Anonymous. False data China's 'biggest source of corruption': statistics chief. *Want China Times*. 10 Apr 2012.
- Bai CE, Wu B. Health insurance and consumption: evidence from China's New Cooperative Medical Scheme. *J Comp Econ*. 2014;42:450–69.
- Banister J, Hill K. Mortality in China 1964–2000. *Popul Stud*. 2000;58(1):55–75.
- Barber SL, Borowitz M, Bekedam H, Ma J. The hospital of the future in China: China's reform of public hospitals and trends from industrialized countries. *Health Policy Plan*. 2014;29(3):367–78. <https://doi.org/10.1093/heapol/czt023>.
- Bien C. The barefoot doctors: China's rural health care revolution, 1968–1981. [Departmental Honors]: Wesleyan University; 2008.
- Bloom G. Building institutions for an effective health system: lessons from China's experience with rural health reform. *Soc Sci Med*. 2011;72:1302–9.
- Bloom G, Xingyuan G. Health sector reform: lessons from China. *Soc Sci Med*. 1997;45(3):351–60. Epub 1 Aug 1997.
- Blumenthal D, Hsiao W. Privatization and its discontents – the evolving Chinese health care system. *N Engl J Med*. 2005;353(11):1165–70. Epub 16 Sept 2005.
- Brixi H, Mu Y, Targa B, Hipgrave D. Engaging sub-national governments in addressing health equities: challenges and opportunities in China's health system reform. *Health Policy Plan*. 2012. <https://doi.org/10.1093/heapol/czs120>.
- Cai Y. An assessment of China's fertility level using the variable-r method. *Demography*. 2008;45(2):271–81. Epub 11 July 2008.
- Chan KY, Wang W, Wu JJ, Liu L, Theodoratou E, Car J, et al. Epidemiology of Alzheimer's disease and other forms of dementia in China, 1990–2010: a systematic review and analysis. *Lancet*. 2013;381(9882):2016–23. Epub 12 June 2013.
- Chen Y, Schweitzer SO. Issues in drug pricing, reimbursement, and access in China with reference to other Asia-Pacific Region. *Value Health*. 2008;11(Suppl 1):S124–S9.
- Chen W, Tang S, Sun J, Ross-Degnan D, Wagner AK. Availability and use of essential medicines in China: manufacturing, supply, and prescribing in Shandong and Gansu provinces. *BMC Health Serv Res*. 2010;10:211.
- China National Health and Family Planning Commission. Statistical bulletin on national health and family planning development. National Health and Family Planning Commission of the PRC, Beijing; 2012.
- China National Health Development Research Centre. China National Health Account 2011 (in Chinese). Beijing: China National Health Development Research Centre; 2012.
- China National Health Development Research Centre. China National Health Account 2012 (in Chinese). Beijing: China National Health Development Research Centre; 2013.
- China News Network. Ministry of Health signs "military-style order" on health reforms to improve poor accountability (in Chinese). China News Network [Internet]. 2010. Available at: <http://www.chinanews.com/jk/jk-ylgg/news/2010/05-24/2299821.shtml>. Last viewed 24 Oct 2014.
- Cui FQ, Wang XJ, Cao L. Progress in hepatitis B prevention through universal infant immunization – China, 1997–2006. *MMWR Morb Mortal Wkly Rep*. 2007;56(18):441–5.
- Deng G, Dou C, Gong Q. Ownership, fees and service quality by health providers. *Econ Rev (Jingji Pinglun)* (in Chinese). 2013;1:121–130.
- Di Martino K. China: ensuring equal access to education and healthcare for children of internal migrants. In: Bhabha J, editor. *Children without a state: a global human rights challenge*. Cambridge, MA: MIT Press; 2011.
- Dong L, Yan H, Wang D. Antibiotic prescribing patterns in village health clinics across 10 provinces of Western China. *J Antimicrob Chemother*. 2008;62(2):410–5.
- Eggleston K, Ling L, Qingyue M, Lindelow M, Wagstaff A. Health service delivery in China: a literature review. *Health Econ*. 2008;17(2):149–65.
- Feltenstein A, Iwata S. Decentralization and macroeconomic performance in China: regional autonomy has its costs. *J Dev Econ*. 2005;76(2):481–501.
- Feng XL, Shi G, Wang Y, Xu L, Luo H, Shen J, et al. An impact evaluation of the Safe Motherhood Program in China. *Health Econ*. 2010a;19(Suppl):69–94. Epub 14 Sept 2010.
- Feng XL, Zhu J, Zhang L, Song L, Hipgrave D, Guo S, et al. Socio-economic disparities in maternal mortality in China between 1996 and 2006. *BJOG*. 2010b;117(12):1527–36.
- Feng X, Guo S, Hipgrave D, Zhu J, Zhang L, Song L, et al. China's facility-based birth strategy and neonatal mortality: a population-based epidemiological study. *Lancet*. 2011;378:1493–500.
- Gong P, Liang S, Carlton EJ, Jiang Q, Wu J, Wang L, et al. Urbanisation and health in China. *Lancet*. 2012;379(9818):843–52. Epub 6 Mar 2012.
- Gottret PE, Schieber G. *Health financing revisited: a practitioner's guide*. Washington, DC: The World Bank; 2006.
- Government of China. State Council Guidance on further strengthening the ranks of rural doctors (in Chinese). 2011. Available at: [http://www.gov.cn/zwggk/2011-07/14/content\\_1906244.htm](http://www.gov.cn/zwggk/2011-07/14/content_1906244.htm). Last viewed 24 Oct 2014.
- Guan X, Liang H, Xue Y, Shi L. An analysis of China's national essential medicines policy. *J Public Health Policy*. 2011;32(3):305–19. Epub 27 May 2011.
- Hipgrave D. Communicable disease control in China: from Mao to now. *J Glob Health*. 2011a;1(2):223–37.

- Hipgrave D. Perspectives on the progress of China's 2009 – 2012 health system reform. *J Glob Health*. 2011b;1(2):142–7. Epub 1 Dec 2012.
- Hipgrave D, Hort K. Will current health reforms in south and east Asia improve equity? *Med J Aust*. 2014;200(9):514.
- Hipgrave D, Guo S, Mu Y, Guo Y, Yan F, Scherpbier RW, et al. Chinese-style decentralization and health system reform. *PLoS Med*. 2012;9(11):1–4.
- Hipgrave DB, Fu X, Zhou H, Jin Y, Wang X, Chang S, et al. Poor complementary feeding practices and high anaemia prevalence among infants and young children in rural central and western China. *Eur J Clin Nutr*. 2014;68:916.
- Ho CS. Health reform and de facto federalism in China. *China Int J*. 2010;8:33–62.
- Ho CS, Gostin LO. The social face of economic growth: China's health system in transition. *JAMA*. 2009;301(17):1809–11. Epub 7 May 2009.
- Hu S. Financing, pricing and utilisation of pharmaceuticals in China: the road to reform. Beijing: The World Bank East Asia and Pacific Region; 2010. Contract No.: 58410.
- Hu G, Baker T, Baker SP. Comparing road traffic mortality rates from police-reported data and death registration data in China. *Bull World Health Organ*. 2011;89(1):41–5. Epub 25 Feb 2011.
- Huang YZ. The sick man of Asia. *Foreign Aff*. 2011;90:119–36.
- Human Rights Watch. "My children have been poisoned": a public health crisis in four Chinese provinces. New York: Human Rights Watch; 2011.
- Jamison DT, Evans JR, King T, Porter I, Prescott N, Prost A. China: the health sector. Washington, DC: The World Bank; 1984.
- Kaiman J. Chinese statistics bureau accuses county of faking economic data. *The Guardian*. 7 Sept 2013.
- Li L, Chen Q-L. A rational evaluation of China's health sector reform over three years. *Health Econ Res*. 2012;5:7–12.
- Liu Y. China's public health-care system: facing the challenges. *Bull World Health Organ*. 2004;82(7):532–8. Epub 27 Oct 2004.
- Liu MD. Sub-provincial intergovernmental fiscal transfers. 2006 Annual China Fiscal Reform Forum. Beijing: UNDP; 2007.
- Liu Y, Rao K, Hsiao WC. Medical expenditure and rural impoverishment in China. *J Health Popul Nutr*. 2003;21(3):216–22. Epub 14 Jan 2004.
- Ma XVC. National Commission for Health and Family Planning. Quoted comments given at press conference during 13th National People Congress 2013 (in Chinese). 2013. Available at: <http://news.sina.com.cn/c/2013-03-15/035926536113.shtml>. Last viewed 24 Oct 2014.
- Meng Q, Xu L, Zhang Y, Qian J, Cai M, Xin Y, et al. Trends in access to health services and financial protection in China between 2003 and 2011: a cross-sectional study. *Lancet*. 2012;379(9818):805–14.
- Millman A, Tang D, Perera FP. Air pollution threatens the health of children in China. *Pediatrics*. 2008;122(3):620–8. Epub 3 Sept 2008.
- Ministry of Health. The National Health Statistics reporting system (in Chinese). Beijing: Chinese Academy of Medical Science; 2007.
- Ministry of Health. Report on women and children's health development in China. Beijing: China Ministry of Health; 2011a.
- Ministry of Health. China's Minister of Health: rural doctors will continue to serve the masses indefinitely (in Chinese). 2011b. Available at: [http://www.gov.cn/gzdt/2011-02/18/content\\_1805889.htm](http://www.gov.cn/gzdt/2011-02/18/content_1805889.htm). Last viewed 24 Oct 2014.
- Ministry of Health. China 2010 health statistical yearbook. Beijing: China Ministry of Health; 2011c.
- Ministry of Health. China's State Council announcement on deepening medical and health system planning and implementation of the program during the 12th Five Year Plan. 2012a. Available at: [http://www.wpro.who.int/health\\_services/china\\_nationalhealthplan.pdf](http://www.wpro.who.int/health_services/china_nationalhealthplan.pdf). Last viewed 24 Oct 2014.
- Ministry of Health. China health statistics yearbook. Beijing: Chinese Academy of Medical Science; 2012b.
- Ministry of Health. Three years of significant progress in health reform (in Chinese). 2012c. Formerly available at: <http://www.moh.gov.cn/publicfiles/business/htmlfiles/mohbgt/s3582/201201/53883.htm>. Last viewed 20 Aug 2012 – MoH website now deleted.
- Ministry of Health. Health statistical monthly reports. Beijing: Ministry of Health; 2012d.
- Ministry of Health Centre for Health Statistics and Information. An analysis report of the fourth national health services survey in China in 2008. Beijing: China Union Medical University Press; 2009.
- Mulholland K, Temple B. Causes of death in children younger than 5 years in China in 2008. *Lancet*. 2010;376(9735):89.
- National Bureau of Statistics. China statistical yearbook (in Chinese). Beijing: National Bureau of Statistics; 2011. Available at: <http://www.stats.gov.cn/tjsj/ndsj/2011/indexch.htm>. Last viewed 24 Oct 2014.
- National Bureau of Statistics. Tabulation of the 2010 population census of People's Republic of China. Beijing: China Statistics Press; 2012.
- National Bureau of Statistics. China statistical yearbooks. Beijing: Published annually; 2016.
- OECD. Health at a Glance: OECD Indicators, OECD Publishing. 2013. [https://doi.org/10.1787/health\\_glance-2013-en](https://doi.org/10.1787/health_glance-2013-en)
- Osno E. China's censored world. *New York Times*. 2 May 2014.
- Pavin M, Nurgozhin T, Hafner G, Yusufy F, Laing R. Prescribing practices of rural primary health care physicians in Uzbekistan. *Trop Med Int Health*. 2003;8(2):182–90.
- Phillips MR, Zhang J, Shi Q, Song Z, Ding Z, Pang S, et al. Prevalence, treatment, and associated disability of mental disorders in four provinces in China during

- 2001–05: an epidemiological survey. *Lancet*. 2009;373(9680):2041–53. Epub 16 June 2009.
- Qin X, Wang W, Jin Q, Ai L, Li Y, Dong G, et al. Prevalence and rates of recognition of depressive disorders in internal medicine outpatient departments of 23 general hospitals in Shenyang, China. *J Affect Disord*. 2008;110(1–2):46–54. Epub 12 Feb 2008.
- Rudan I, Chan KY, Zhang JS, Theodoratou E, Feng XL, Salomon JA, et al. Causes of deaths in children younger than 5 years in China in 2008. *Lancet*. 2010;375(9720):1083–9. Epub 30 Mar 2010.
- Schatz G, Nowlin P. Drugs for the masses. *China Bus Rev*. 2010;2010:22–5.
- Seiter A, Wang H, Zhang S. A generic drug policy as a cornerstone to essential medicines in China. 2010. Contract No.: 58413.
- Shi LW, Ma YQ, Xu LP, Zhao DH, Zhang Y. Review of adjustment of essential medicine list at provincial level in China. *Value Health*. 2011;14(3):A14.
- State Council. Opinions of the Communist Party of China Central Committee and the State Council on deepening the health care system reform. 2009. Available at: [http://www.china.org.cn/government/scio-press-conferences/2009-04/09/content\\_17575378.htm](http://www.china.org.cn/government/scio-press-conferences/2009-04/09/content_17575378.htm). Last viewed 24 Oct 2014.
- 'State Council. Opinions on promoting the comprehensive reform of county public hospitals (in Chinese). 2014. Available at: [http://baike.baidu.com/link?url=jPS0SkO7GzuPcuhNHwpFhXEF0aPjpJ8PKVX7\\_GzRMhMUYNg\\_u-THEs\\_dlop84b77tO3Y2PXMS0XyKJ5gLPtyg](http://baike.baidu.com/link?url=jPS0SkO7GzuPcuhNHwpFhXEF0aPjpJ8PKVX7_GzRMhMUYNg_u-THEs_dlop84b77tO3Y2PXMS0XyKJ5gLPtyg). Last viewed 5 July 2014.
- Sun X, Jackson S, Carmichael GA, Sleigh AC. Prescribing behaviour of village doctors under China's New Cooperative Medical Scheme. *Soc Sci Med*. 2009;68(10):1775–9. Epub 4 Apr 2009.
- Tang S, Meng Q, Chen L, Bekedam H, Evans T, Whitehead M. Tackling the challenges to health equity in China. *Lancet*. 2008;372(9648):1493–501. Epub 22 Oct 2008.
- Tang S, Brixi H, Bekedam H. Advancing universal coverage of healthcare in China: translating political will into policy and practice. *Int J Health Plann Manage*. 2014a;29(2):160–74. <https://doi.org/10.1002/hpm.2207>
- Tang C, Zhang Y, Chen L, Lin Y. The growth of private hospitals and their health workforce in China: a comparison with public hospitals. *Health Policy Plan*. 2014b; 29(1):30–41. <https://doi.org/10.1093/heapol/czs130>
- Tangcharoensathien V, Patcharanarumol W, Ir P, Aljunid SM, Mukti AG, Akkhavong K, et al. Health-financing reforms in southeast Asia: challenges in achieving universal coverage. *Lancet*. 2011;377(9768):863–73. Epub 29 Jan 2011.
- The World Bank. World development report 2004: making services work for the poor people. Washington, DC: The World Bank; 2003.
- The World Bank Group East Asia Pacific Region. Fixing the public hospital system in China. Washington, DC; 2010. Contract No.: 58411.
- The World Bank Human Development Unit. Toward a healthy and harmonious life in China: stemming the rising tide of non-communicable diseases. Washington, DC: The World Bank; 2011.
- The World Health Organization. The world health report 2000: health systems: improving performance. Geneva: World Health Organization; 2000.
- Tian Y, Hua LJ, Chao WM. Chinese doctors' salaries. *Lancet*. 2008;371:1577.
- Uchimura H, Jütting, J. Fiscal decentralization, Chinese style: good for health outcome? OECD Development Centre Working Paper #264. Paris: OECD; 2007.
- UNDP China, China Institute for Reform and Development. China national human development report 2007/2008: access for all: basic public services to benefit 1.3 billion people. Beijing: UNDP; 2008.
- Walter CE, Howie F. Red capitalism: the fragile financial foundation of China's extraordinary rise. Singapore: Wiley; 2011.
- Wang J. Evidence-based medicine in China. *Lancet*. 2010;375(9714):532–3. Epub 18 Feb 2010.
- Wang H, Xu T, Xu J. Factors contributing to high costs and inequality in China's health care system. *JAMA*. 2007;298(16):1928–30. Epub 24 Oct 2007.
- Wang L, Wang Y, Jin S, Wu Z, Chin DP, Koplan JP, et al. Emergence and control of infectious diseases in China. *Lancet*. 2008a;372(9649):1598–605. Epub 22 Oct 2008.
- Wang SY, Li YH, Chi GB, Xiao SY, Ozanne-Smith J, Stevenson M, et al. Injury-related fatalities in China: an under-recognised public-health problem. *Lancet*. 2008b;372(9651):1765–73. Epub 22 Oct 2008.
- Wang YP, Miao L, Dai L, Zhou GX, He CH, Li XH, et al. Mortality rate for children under 5 years of age in China from 1996 to 2006. *Public Health*. 2011;125(5):301–7. Epub 29 Apr 2011.
- Wang Y, Zhu J, He C, Li X, Miao L, Liang J. Geographical disparities of infant mortality in rural China. *Arch Dis Child Fetal Neonatal Ed*. 2012;97(4):F285–90. Epub 17 Jan 2012.
- Wong C. Public Sector Reforms toward Building the Harmonious Society in China. Paper prepared for the China Economic Research and Advisory Programme. University of Oxford; 2010.
- World Bank. China 2030: building a modern, harmonious, and creative high-income society. Washington, DC: The World Bank; 2012.
- Wu N, Yang HW. Retrospective evaluation of the achievements of the implementation of essential medicine system in medical reform of past three years. *China Pharm*. 2013;10(5–6):78–82.
- Xiao Y, Zhao K, Bishai DM, Peters DH. Essential drugs policy in three rural counties in China: what does a complexity lens add? *Soc Sci Med*. 2013;93:220–8. <https://doi.org/10.1016/j.socscimed.2012.09.034>.
- Xing L, Fen S, Luo X, Zhang X. Intra rural income disparity in West China. *China Econ Q*. 2008;1(1):329–50.
- Xinhua. China penalizes 113 over chemical tainted pork. *China Daily*. 26 Nov 2011 10:36.
- Xu H, Zhang W, Gu L, Qu Z, Sa Z, Zhang X, et al. Aging village doctors in five counties in rural China: situation

- and implications. *Hum Resour Health*. 2014;12:36. Epub 30 June 2014.
- Yang DL. The central-local relations dimension. In: Freeman CW, Lu XQ, editors. *Implementing health care reform policies in China*. Washington, DC: Center for Strategic and International Studies; 2011. p. 21–9.
- Yang ZP, Fan DM. How to solve the crisis behind Bribe-gate for Chinese doctors. *Lancet*. 2012;379(9812):e13–5.
- Yang L, Cui Y, Guo S, Brant P, Li B, Hipgrave D. Evaluation, in three provinces, of the introduction and impact of China's National Essential Medicines Scheme. *Bull World Health Organ*. 2013a;91:184–94.
- Yang G, Wang Y, Zeng Y, Gao GF, Liang X, Zhou M, et al. Rapid health transition in China, 1990–2010: findings from the Global Burden of Disease Study 2010. *Lancet*. 2013b;381(9882):1987–2015. Epub 12 June 2013.
- Yip WC-M, Hsiao WC, Chen W, Hu S, Ma J, Maynard A. Early appraisal of China's huge and complex health-care reforms. *Lancet*. 2012;379(9818):833–42.
- Youlong G, Wilkes A, Bloom G. Health human resource development in rural China. *Health Policy Plan*. 1997;12(4):320–8. Epub 3 Nov 1997.
- Yu X, Li C, Shi Y, Yu M. Pharmaceutical supply chain in China: current issues and implications for health system reform. *Health Policy*. 2010;97(1):8–15. Epub 24 Mar 2010.
- Zhan SK, Tang SL, Guo YD, Bloom G. Drug prescribing in rural health facilities in China: implications for service quality and cost. *Trop Doct*. 1998;28(1):42–8. Epub 3 Mar 1998.
- Zhang W, Navarro V. Why hasn't China's high-profile health reform (2003–2012) delivered? An analysis of its neoliberal roots. *Crit Soc Policy*. 2014;34:175–98.
- Zhang J, Zhang X, Qiu L, Zhang R, Hipgrave D, Wang Y, et al. Maternal deaths among rural-urban migrants in China: a case-control study. *BMC Public Health*. 2014;14:512.
- Zheng M, Fu Q, Wang X. Comparative study on structural changes in income disparities in urban households in Chongqing Municipality, Shanghai Municipality and Sichuan Province. *J Reform Strategy*. 2008;5:98–101.
- Zhou LA. Reforming China's local government governance. In: *Incentives and governance: China's local governments*. Singapore: Cengage Learning Asia Pte. Ltd.; 2010a.
- Zhou LA. Incentives and governance: China's local governments. Singapore: Cengage Learning Asia Pte. Ltd.; 2010b.
- Zhou L, Conner KR, Caine ED, Xiao S, Xu L, Gong Y, et al. Epidemiology of alcohol use in rural men in two provinces of China. *J Stud Alcohol Drugs*. 2011;72(2):333–40. Epub 11 Mar 2011.



Christian A. Gericke, Kaylee Britain, Mahmoud Elmahdawy,  
and Gihan Elsis

## Contents

<b>Introduction</b> .....	811
<b>Organization and Governance</b> .....	813
Overview .....	813
Historical Background Until 2011 .....	814
Public System .....	815
Private System .....	816
Information Systems .....	817
<b>Financing</b> .....	817
Overview .....	817
Expenditure .....	817
External Sources of Financing .....	818
Insurance Coverage .....	818
Health Payments .....	820
Paying Health Workers .....	820

---

C. A. Gericke (✉)  
Anton Breinl Centre for Health Systems Strengthening,  
James Cook University, Cairns, Australia

University of Queensland School of Public Health,  
Brisbane, Australia  
e-mail: [c.gericke@uq.edu.au](mailto:c.gericke@uq.edu.au)

K. Britain  
University of Queensland School of Public Health,  
Brisbane, Australia  
e-mail: [kaylebritain@gmail.com](mailto:kaylebritain@gmail.com)

M. Elmahdawy  
Ministry of Health, Cairo, Egypt  
e-mail: [mahmoud77@yahoo.com](mailto:mahmoud77@yahoo.com)

G. Elsis  
Ministry of Health, Cairo, Egypt  
Faculty of Pharmacy, Heliopolis University, Cairo, Egypt  
e-mail: [gihanhamdyelsisi@hotmail.com](mailto:gihanhamdyelsisi@hotmail.com)

<b>Physical and Human Resources</b> .....	820
Physical Resources .....	820
Human Resources .....	821
<b>Provision of Services</b> .....	821
Overview .....	821
Inpatient Care .....	822
Outpatient Care .....	822
Mental Health Care .....	822
Pharmaceuticals .....	822
<b>The Arab Spring Revolution</b> .....	823
<b>Reforms</b> .....	823
Overview .....	823
Past Reforms .....	824
Proposed Plans .....	824
<b>Assessment</b> .....	825
<b>References</b> .....	826

### Abstract

With over 95 million inhabitants, Egypt is the second most populous country in the Middle East and North Africa. Poverty has nearly doubled over the last 15 years. Egypt has a very young population, and youth unemployment has become a major societal issue.

Egypt's health-care system is pluralistic combining both public and private providers and financiers. The largest public health-care payers are the Health Insurance Organization (HIO) and the Curative Care Organization (CCO). HIO covers 60% of the population, and provides basic coverage to employees, students, and widows through their own hospitals and clinics. CCO contracts with individuals and companies to provide inpatient and outpatient care that was developed through the privatization of Egypt's health-care providers over the last two decades. Although the public system provides basic universal coverage, it is plagued by chronic underfunding, low service quality, and high out-of-pocket payments.

The private sector comprises private hospitals, doctors, and pharmacies, perceived as of higher quality than public services. Most private services are paid for out-of-pocket; private health insurance is insignificant.

With only 4.75% of GDP spent on health, total health expenditure (THE) in Egypt is low compared to other lower-middle-income countries. Out-of-pocket payments comprise over 60% of THE. Spending on pharmaceuticals is relatively high with over 25% of THE, mostly in the form of out-of-pocket costs. Another problem is the lack of communication between public and private providers.

Widespread public dissatisfaction with basic living conditions spurred the Arab Spring revolution in 2011. Since then, the country has seen sustained political instability and slow economic growth which have thwarted most long-term plans for health reform. Several reform measures have been publicly discussed, but only few were implemented such as the introduction of a pharmacoeconomics unit in the Ministry of Health to curb the disproportionately high spending on pharmaceuticals.

A long-term national strategy is needed to address issues of growing inequalities in financial access to care, the perceived low quality of public services, as well as the growing privatization of health care which furthers the existing inequalities in access to care.

## Introduction

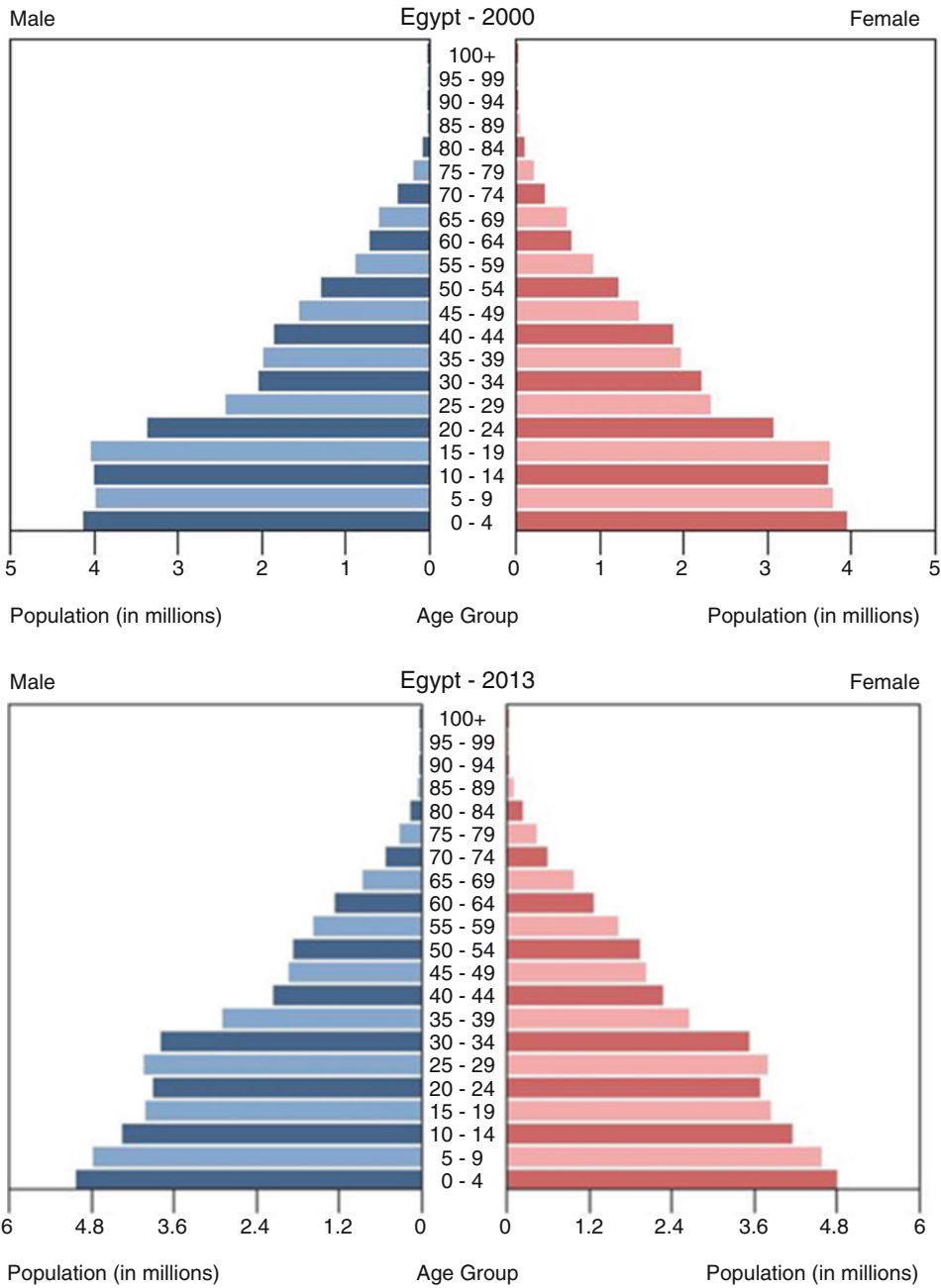
Egypt is the second most populous country in the World Health Organization's (WHO) Eastern Mediterranean region (WHO 2010). Egypt's continued population growth along with increases in urbanization puts much strain on the country's capacity to provide enough food through agriculture (CIA 2013) and on many publicly provided services including health care. With its location in the northeast corner of Africa, Egypt has been the cultural bridge between the African continent and the Middle East for millennia. The country consists mostly of desert and relies on the limited stretch of fertile land along the river Nile and its branches as its only perennial water source (CIA 2013). The rapid growth in population and urbanization has been a continued source of threatening health concerns due to the unsanitary, polluted, and overstrained environment (CIA 2013; Anwar 2003). In recent years, with increasing fertility rates along with decreasing infant mortality rates, the largest age group is 0–14 years (Fig. 1; CIA 2013). This implies a highly dependent population impeding economic growth. With a part of this young population moving into the workforce in recent years, unemployment rates for them have become a huge issue.

Egypt is a lower-middle-income country with the majority of its income coming from tourism, remittances from working abroad, the Suez Canal, and oil sales (WHO 2010). Poverty within the country has continued to decrease through recent decades aided by substantial international donor support (Egypt's progress 2010). Despite this, poverty rates did increase in 2008 to 25.2% as a result of the global economic crisis. However, extreme poverty continued to decline to 4.8% (Egypt's progress 2010). Most of the poverty is found in rural areas (most notably in Upper Egypt) which consists of just over 50% of Egypt's population (Egypt's progress 2010). These inequalities have carried over into both health and literacy indicators for the respective areas (Table 1; Egypt's progress 2010). Along with poverty rates, the economic crisis has also made an impact

on reducing economic growth rates from 7% to 4.7% which resulted in one fifth of the Egyptian population now falling into the “near-poor” category (Egypt's progress 2010).

Politically, Egypt is a constitution-based republic. This consists of a mixed legal system based on Napoleonic civil law and Islamic religious law and judicial review by a Supreme Court and Council of State (CIA 2013). The constitution created three separate branches of government: (a) the executive, headed by the president; (b) the legislative, consisting of a People's Assembly and the Advisory Council; and (c) the judicial branch, with a Supreme Constitutional Court (consists of the court president and ten members), Court of Cassation, and subordinate courts: Courts of Appeal, Courts of First Instance, Courts of Limited Jurisdiction, and a Family Court (CIA 2013). After the independence from British colonialism and the ousting of the last Egyptian King Farouk in 1952 in the great revolution led by General Gamal Abdel Nasser, Egypt came under socialist rule in the 1950s and 1960s (Jabbour 2012). After the death of General Nasser in 1970, his vice president and close ally during the revolution Anwar Al-Sadat replaced him and moved Egypt toward a market-based economy. When president Al-Sadat was assassinated in 1981 by Islamist fundamentalists, his vice president General Hosni Mubarak took over the presidency. He continued Sadat's market-based economic and international open political approach. Mubarak served as president for five consecutive terms up until the Arab Spring revolution in 2011 (Jabbour 2012).

The opening of the country to the world economy under presidents Al-Sadat and Mubarak allowed continued economic growth through the years as well as implementation of various economic reforms in order to balance economic inequalities as well as reduce foreign debt (Ministry of Health, Egypt 2010). Unfortunately, economic reforms were accompanied by problematic social effects giving rise to unemployment and poverty. It seems that the Egyptian Social Fund for Development which was instituted in



Source: United States Census Bureau, International Programs 2013

**Fig. 1** Population pyramids for Egypt (2000 and 2013). (Source: United States Census Bureau, International Programs 2013)

1991 to counter some of these undesired side effects had some positive impact through its microcredit and community financing initiatives (Abou-Ali et al. 2010). Despite some social

improvements and a growing economy, widespread public dissatisfaction with basic living conditions and high levels of poverty remained and spurred the Arab Spring revolution in 2011.



**Table 1** Socioeconomic and demographic indicators for Egypt

Indicator	Year	Country
<b>Socioeconomic</b>		
Total population	2013	85,294,388
Population living in urban areas (%)	2010	43%
Gross national income per capita	2010	6120
Gross domestic product	2010	\$255 billion
GDP growth rate (%)	2012	2.00%
Poverty rate (%)	2010	25.20%
Unemployment rate (%)	2012	12.50%
Rate of urbanization	2010	2.10%
Literacy rate males (%)	2010	80%
Literacy rate females (%)	2010	64%
<b>Demographic</b>		
Total fertility rate (per woman)	2013	2.9
Population 0–14 (%)	2013	32.30%
Population 65 years and over (%)	2013	4.80%
Death rate (per 1000 population)	2013	4.79
Birth rate (per 1000 population)	2013	23.79

Sources: CIA Factbook 2013, WHO 2010, World Bank 2013

The uprising has caused economic growth to slow down (CIA 2013) in the past few years due to the political uncertainty along with a significant reduction in tourism (Haley and Beg 2012). At the same time, the revolution has resulted in increased social spending to address public dissatisfaction and has also led toward a reduction in foreign exchange reserves contributing to a rising deficit (CIA 2013).

Overall health in Egypt prior to 2011 had been steadily improving over time with marked increases in life expectancy and decreases in infant mortality rates since 1990 (Table 2). Communicable disease control, in particular, for endemic tropical diseases such as schistosomiasis has also made great improvements during this time; however, diarrheal diseases, acute respiratory infections, and hepatitis are still reported from health facilities (CIA 2013). Compared to other MENA countries, the population percentage of communicable diseases is

in fact low, while noncommunicable diseases in Egypt are higher than other countries in the geographic region (WHO 2013). Of the non-communicable diseases, like in many other countries, obesity is a growing factor with over 33% of the population being obese as of 2008 (WHO 2013). HIV/AIDS has also been an increasing health issue with over 11,000 known persons with the infection as of 2009. Today the top three diseases causing mortality are essential primary hypertension, intracerebral hemorrhage, and fibrosis/cirrhosis (WHO 2013). In contrast, Egypt has put little emphasis on controlling environmental risks to health and well-being (Anwar 2003; Gericke 2006).

## Organization and Governance

### Overview

Egypt's health-care system is pluralistic and complex combining both public and private providers and financiers. The government has committed to provide health care to the poor; however, with a system pluralistic in nature, health-care providers compete, and clients are free to choose services based on their needs along with the ability to pay (WHO-EMRO 2006). Subsequently, the health-care system relies upon four financing agents:

- Government sector
- Public sector
- Private organizations
- Household payments (out-of-pocket)

The government sector represents the various ministries and departments of the government financed primarily through the Ministry of Finance (MOF). Other government agents are the Ministry of Health and Population (MOHP), the Ministry of Higher Education, and the Ministries of Interior and Defense. The MOHP is responsible for policy formulation and the regulation of the health sector including public, nongovernmental, and private organizations

**Table 2** Health trends in Egypt

Indicators	1990	1995	2000	2004	2005	2010	2013
Life expectancy at birth (total)	65.3 (92)	66.9 (98)	67.1 (01)	70.1 (02)	–	73	73.19
Infant mortality rate	63	66	24.5	22.4	20.5	–	23.3
Under five mortality rate (per 1000 live births)	–	3.9 (97)	33.8	28.6	26.3	21	–
Maternal mortality ratio (per 100,000 live births)	174 (92)	96 (98)	84 (01)	68 (02)	63	66	–

Sources: CIA 2013; WHO 2006, 2010

covering over 29 ministries and organizations (WHO-EMRO 2006). The MOHP is also responsible for providing preventative and curative care throughout all of Egypt making it the largest provider of health-care services in the country (WHO-EMRO 2006).

The public sector comprises financially independent governmental organizations. The largest of these are the Health Insurance Organization (HIO) and the Curative Care Organization (CCO) (Haley and Beg 2012; WHO-EMRO 2006). The HIO is Egypt's public health insurance with the goal of providing sustainable and universal coverage to employees, students, and widows through their own hospitals and clinics. The CCO contracts with individuals and companies to provide inpatient and outpatient curative care that was developed through the privatization of Egypt's health-care system (Ministry of Health, Egypt 2010). The private sector includes private health insurance companies that can be either non-profit or for profit (Ministry of Health, Egypt 2010). The private sector has been the fastest-growing source of health provision as the country has continuously moved toward privatization in the last two decades (WHO 2010). The private sector comprises private pharmacies, doctors, and private hospitals and overall provides care that continues to remain much higher rated than its public counterparts.

For all of these levels of care, out-of-pocket payments have consistently remained the largest source of health financing in Egypt (Ministry of Health, Egypt 2010; WHO-EMRO 2006) with an adoption of the idea of "fee-for-service." This idea forces households to pay at the point of care in both private and public health facilities.

## Historical Background Until 2011

The progression of Egypt's health-care system today begins with the implementation of socialist rule under the Nasserite regime (1950s–1960s). This social movement nationalized many services including hospitals (Jabbour 2012). It was also during this time that the HIO and the CCO were established as the idea of health insurance based on actuarial premiums was unacceptable (Jabbour 2012). The health-care system grew significantly under Nasser furthering not only primary care but also secondary and tertiary care through a system of fee-for-service (Jabbour 2012). After these improvements were made under the Nasserite regime, new economic policies under Sadat to Mubarak were introduced to bring up a newly falling economic performance. These actions lead toward an increased privatization of the health-care system in Egypt. This began with the introduction of the "Infitah" policy under Sadat which was created to reduce the government's role in the economy to allow for more private involvement and investments (Salem 2002). Furthermore, this started the development of the health-care system in accordance with international agencies and standards (Jabbour 2012). With the help of mostly USAID policies, investments were made in expanding the private health-care sector (Jabbour 2012).

In the 1990s, the Egyptian government made a declaration to focus on improving health for the nation. The aim of this statement was to initiate the provision of a universal health-care system along with the adoption of the family health model for the provision of primary care (WHO 2010). This led to the government created Health

Sector Reform Program (HSRP) established through the Family Health Fund along with the HIO (Salem 2002). To aid this development, Egypt received substantial foreign aid and assistance by the World Bank, USAID, and the European Commission (Salem 2002). Egypt also became a party to International Health Regulations (IHR) to improve practice, surveillance, and preparedness for health issues (WHO 2010). Egypt's HSRP was officially introduced in 1997 to address how health in Egypt is organized, financed, and delivered (Haley and Beg 2012). This program has worked to improve upon the disjointed and complex health system through the private and public sector that existed then and now in Egypt. A few years after its establishment, the Healthy Egyptians 2010 Initiative was launched in 2000 to foster disease prevention and control (Anwar 2003).

The accumulation of reforms has benefited the health system in Egypt by implementing a social health insurance model, successfully increasing surveillance, and reducing communicable disease incidence and prevalence (WHO 2010). However, given Egypt's lower-middle-income status, its overall population health is relatively poor in comparison with other lower-middle-income countries. Furthermore, despite some improvements, the burden of noncommunicable diseases has increased, putting further strain on Egypt's health system (Roberts et al. 2013). Universal health care still has to be achieved due in large part to the privatization and its subsequent reduction in public spending which forced an increase in prepaid private and in out-of-pocket health expenditure (WHO-EMRO 2006). Compared to other lower-middle-income countries, Egypt spends comparatively little on health care: only 4.75% of GDP (2007–2008) (Ministry of Health, Egypt 2010).

## Public System

The primary organization behind the public system in Egypt is the MOHP. The MOHP offers health service free of charge to every Egyptian citizen covering all inpatient and outpatient care

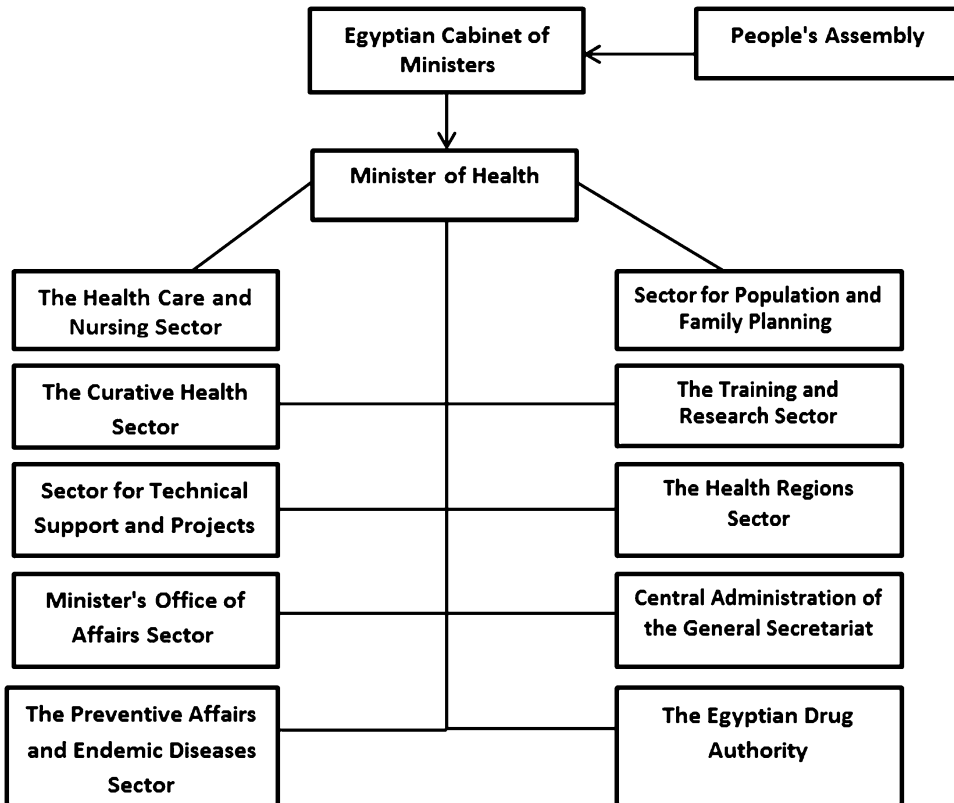
(Elgazzar 2009). This organization is headed by the minister and further employs over 5,000 personnel in managing and delivering public health services (WHO-EMRO 2006). However, due to poor salary bases for doctors along with income-based inequality in service utilization, the quality of public health care in Egypt is known to be poor which shifts both suppliers and demand to private health care (Elgazzar 2009). Despite this, the MOHP is the major provider of primary, preventative, and curative care with over 4,500 health facilities throughout the country (WHO-EMRO 2006). The MOHP delivers its functions through four separate levels which correlate to the following levels of health care (WHO-EMRO 2006):

- Central
- Health directorates (government level)
- Health districts
- Health-care providers

Centrally, the MOHP is divided into ten sectors (MOHP 2013) depicted in Fig. 2.

These sectors, in accumulation, control the policy and regulation of health and health services in all of Egypt. The governorate level of the MOHP operates in purchasing and financing health care for the Egyptian population by balancing income and expenditure in compliance with health sector regulations (WHO-EMRO 2006). The district health structure is simply a replication of the government level in functionality except on smaller scales (WHO-EMRO 2006). Finally, the provider level of the MOHP is divided based on services as well as location (WHO-EMRO 2006). Despite a consistent discrepancy between rural and urban health care in Egypt, the MOHP does try to provide a large variation of all necessary services to all populated areas of Egypt.

A main component of the public sector of the Egyptian health system is the HIO. While most Egyptians rely on private care provision in addition to the services provided by the MOHP, the HIO is the largest health insurer in Egypt with continuous increases to its utilization through the years (Haley and Beg 2012). From 1990 to 2008, the percentage of population insured by the HIO



Source: MOHP, 2013

**Fig. 2** Organization of the MOHP. (Source: MOHP 2013)

increased from 10% to 55% (Table 4) showing not only its growing use but also improvements to the public sector by increased access (Ministry of Health, Egypt 2010). However, the provision of health from all public-sector services has suffered from the government and MOHP's inability to keep up with increasing costs (WHO 2010). This has turned not only patients but also doctors to the private health system which can provide both better salaries and physical resources.

### Private System

Increased privatization along with poor maintenance of public care has driven substantial development of the private health-care system in Egypt. Moreover, the private system has achieved a

perception of high quality within the country. However, the system prior to 2011 has not set up sufficient regulations on governing its service and finance, forcing much of the service to be provided through purely out-of-pocket payments (WHO-EMRO 2006). This increases the inequality of health-care access within the country as the private services are only for those who can afford them. Furthermore, because there are less regulations, more doctors are relying on private care work as supplemental payment which has been a key factor in the private system's perceived better quality of care (WHO 2010). The lack of governmental regulation along with competing health insurers and providers has resulted in a severe absence in communication between the private and public sectors. This has been a key source of Egypt's health system's continuous dysfunction.

## Information Systems

Egypt's health information systems have continued to be developed through increased surveillance implementation within the country. In 2000, the Epidemiology and Disease Surveillance Unit (ESU) was created through the MOHP in order to assess health staff and monitor health patterns, risk factors, and diseases (WHO-EMRO 2006). In total, Egypt has increased surveillance to cover 26 communicable diseases which has helped to reduce the incidence of tuberculosis as well as the elimination of polio within the country (WHO 2010). Egypt, experiencing a dual burden of disease, has only recently implemented the STEPwise surveillance framework for non-communicable disease. The STEPwise survey was successfully conducted in 2011 using a standard survey instrument and a methodology adapted to Egypt's resource setting in accordance with WHO (2013). There is also currently a supply chain system megaproject underway working to centralize and computerize drug ordering, procurement, delivery, and other associated logistics. In conjunction with this, an electronic medical records (EMR) system is also a work in progress. This would be an integral component of the national health insurance project in order to avoid service duplication and abuse of the system by any group of patients.

---

## Financing

### Overview

In 2007/2008, Egypt invested 42.5 billion Egyptian pounds (LE) on health. For a middle-income country in the region, this amount of spending is relatively low (see Table 3 for comparisons) (Ministry of Health, Egypt 2010). Breaking this down, financing derives from direct tax revenues, HIO premium payments and direct out-of-pocket spending from private households, private health insurance premiums, and health spending from employers to employees, and finally assistance also comes from a cigarette tax as well as minor donor assistance (Ministry of Health, Egypt

2010). Overall government investments in the Egyptian health system have been declining over the years (World Bank 2013). This has subsequently forced financing from private households (out-of-pocket) to continue rising distinguishing it further as the single largest source of health financing in the country. Moreover household expenditure has risen past 60% of all health investments (Ministry of Health, Egypt 2010).

The second largest source of health financing is the Ministry of Finance (MOF). The MOF accounts for the public financing of the health system most notably in regard to programs and services from the MOHP as well as some support to the social health insurance (HIO). The public sector of health financing totals approximately one third of all health investments (Ministry of Health, Egypt 2010). Private and other external sources account for the remaining two thirds (Fig. 2; Ministry of Health, Egypt 2010).

In summary, Egypt spends a mere 4.7% of GDP on health (Ministry of Health, Egypt 2010). Only 1.6% of GDP accounts for public spending, and 2.1% of GDP accounts for private spending on health care (WHO-EMRO 2006). Given the country's economic status, this value is low, and the percentage of out-of-pocket spending by individuals is high, compared to regional comparators (Table 3).

## Expenditure

In Egypt, public funding for the health system flows to financial agents and then onto providers under mutually exclusive tracts known as silos. This impedes care coordination and effective allocation of resources between the public and private sectors (Ministry of Health, Egypt 2010). From this, expenditure moves into various parts of the health system ranging from both private and public service providers to pharmaceuticals. The largest part of health financing is expenditure for pharmaceuticals (Fig. 3). Pharmaceuticals account for 25.9% of total health expenditure which is a relatively high percentage in comparison of comparable health systems (Ministry of Health, Egypt 2010). Funding for pharmaceuticals mostly

**Table 3** Comparison of health spending in Egypt to other WHO's Middle Eastern countries (2007/2008)

	Percent of GDP spent on health (%)	Government spending as the percentage (%)	Health spending as the percentage of total government budget (%)	Out-of-pocket expenditure as the percentage (%)	Per capita health spending (Constant 2005 US\$)
Algeria	4.49	83.85	10.65	15.30	205
Dji bouti	8.54	76.07	14.15	23.60	81
<b>Egypt</b>	<b>4.75</b>	<b>33.00</b>	<b>5.00</b>	<b>60.00</b>	<b>111</b>
Iran	6.30	45.72	11.40	51.68	294
Jordan	9.10	62.20	11.35	33.40	273
Lebanon	8.76	48.99	12.39	39.95	551
Li bya	2.80	75.88	5.38	24.12	383
Morocco	5.33	34.87	6.17	56.13	133
Syria	3.23	45.13	6.01	54.87	76
Tunisia	5.95	49.17	8.90	42.52	213

Sources: WHO NHA data, Egypt NHA results, Jordan NHA report, cited in Egypt MOH 2010

comes from out-of-pocket spending as a result of a lack of communication and education about proper use and distribution. Private clinics receive the next highest amount of expenditures similarly as a result of out-of-pocket spending by households (Figs. 4 and 5).

### External Sources of Financing

While the majority of public health financing comes from the Egyptian government's Ministry of Finance, Egypt has progressed its health policies and system through the aid of outside country's resources. USAID has been a leading resource in aiding the achievement of various health-related Millennium Development Goals (Egypt's progress 2010). USAID was responsible for policies in improving health care through further privatization of the health system. However, USAID pulled out funding in 2009 as a result of Egypt's significant progress in improving health status. Other countries have also helped in funding health policies and goals following international standards and ideals, in particular the European Union and some of its member states. The International Monetary Fund (IMF) has planned to loan \$4.8 billion to the country in recent years; however, ongoing political tensions have prolonged Egypt's bid to secure this (World Bank 2013).

### Insurance Coverage

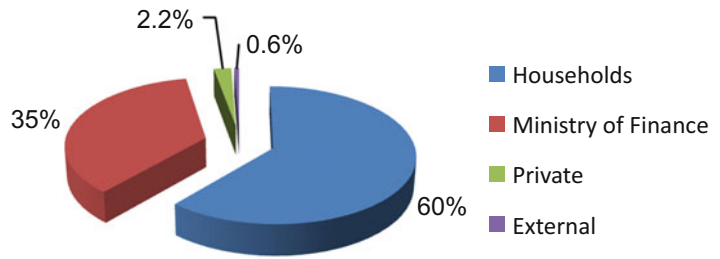
The largest source of health insurance for Egyptians remains the Health Insurance Organization (HIO), essentially a social health insurance system that is supposed to complement the tax-financed services provided by the MOHP. The aim of this organization is to provide a universal health-care coverage for all Egyptians. While this organization has not achieved this goal, its coverage has continued to rise significantly since 1990 (Table 4). As such, the HIO represents the second largest health financing organization in Egypt (WHO-EMRO 2006). The rise in coverage has come with the inclusion of new population groups such as newborns and school-age children (WHO-EMRO 2006).

Given these inclusions, there are four classes of HIO beneficiaries (WHO-EMRO 2006):

1. All employees working in the government sector
2. Private- and public-sector employees and pensioners and widows
3. Beneficiaries of the Student Health Insurance Program (SHIP)
4. Newborn children up to age 5 years

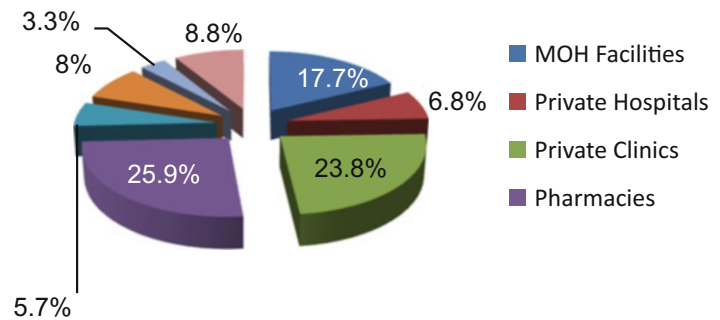
Currently, Egypt has no health insurance laws. The HIO instead operates under different social insurance laws, ministerial decrees, and regulations.

**Fig. 3** Egyptian health investments (2007–2008). (Source: Ministry of Health 2010)



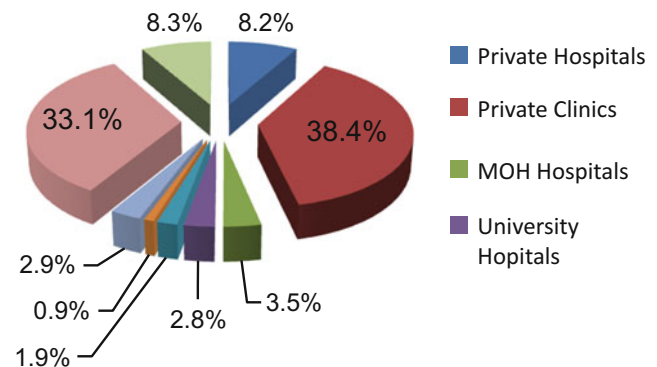
Source: Ministry of Health 2010

**Fig. 4** Health expenditure by type of provider and ownership (2007–2008). (Source: Ministry of Health 2010)



Source: Ministry of Health 2010

**Fig. 5** Out-of-pocket expenditure by provider (2007–2008). (Source: Ministry of Health 2010)



Source: Ministry of Health 2010

**Table 4** Insurance coverage (1990–2011)

	1990	1995	2000	2004	2008	2011
Social insurance	10%	37%	45%	52%	55%	59%
Uninsured/Uncovered	90%	63%	55%	49%	N/A	N/A

Sources: Egypt NHA data, cited by WHO 2006, 2010

The goal of the HIO is to be a provider of health services to its beneficiaries under a low and fixed premium structure with an extensive benefit package (WHO-EMRO 2006). These benefits include transplants, plastic surgery, and treatments abroad

with no cap on the quantity of services (WHO-EMRO 2006). However, there remains inadequate management of HIO service providers resulting in poor care and low responsiveness in the public system (Mosallam et al. 2013). Therefore, with the

higher demand for private care, the benefits of social insurance force most patients insured with the HIO to continue to pay out-of-pocket for most of their health care.

Private health insurance is not significant in Egypt in terms of financing or population coverage, but it is on offer for the select few who can afford it, for example, Egycare. The regulations have made it difficult for organizations to profit from health insurance schemes, which remain the biggest barrier to their spread. However, from time to time, new private health insurance programs appear which benefit upper-middle and upper-class populations (WHO-EMRO 2006).

A major development post the 2011 Arab Spring revolution era was the enactment of a new health insurance law, a project in the pipeline that was hotly debated at the ministerial cabinet's level, in addition to the parliament floor. Unfortunately political instability with frequent changes in the government executive has delayed the project launch and implementation. A new draft of the law was presented to the Higher Health Council (HHC) in April 2012 which requested some modifications. The goal of this development is to improve universal coverage within the country in both in a cost-effective way. The new social health insurance (SHI) would be intending to cover 90% of the population and reduce out-of-pocket payments to 35% at the end of implementation phase.

## Health Payments

Out-of-pocket payments have always been the largest source of service payment. Regardless of insurance status, there are formal user fees for both outpatient and inpatient public services, the MOHP facilities having the smallest fees due (WHO 2010). Overall there have been three separate pathways in provider payment (WHO-EMRO 2006):

1. MOF funding
  - (a) Funds to government care providers
2. Social insurance
  - (a) Funds services as a combined provider and commissioner (half of revenues to finance services by itself and the other half to

purchase services and goods from other providers)

3. Direct household funding

- (a) Over 90% of this goes directly to private health-care providers.

Another source of health revenues has been the Family Health Fund (FHF). The FHF pays performance-based incentives to health workers in the public sector (WHO-EMRO 2006).

## Paying Health Workers

Prior to 2011, over 50% of health professionals were employed by MOHP facilities (WHO-EMRO 2006). With the limited funds that the MOHP receives, salaries for individuals are limited forcing most professionals to practice privately for further sources of income. In turn, 89% of medical doctors had been found to hold more than one job prior to 2011 (WHO-EMRO 2006). This allows for a balance in salary payments as a result of out-of-pocket payments to the private facilities.

Post January 2011, the health ministry has worked closely with health-care practitioners (HCPs) and their respective syndicates, in addition to the MOF to establish a new payroll system for government-employed health-care workers. It is designed to reduce the gap between different payroll categories. In addition to further appreciate those willing to serve in distant geographical locations, new incentive schemes are being developed. Once implemented, this would encourage the recruitment of more competitive health-care practitioners into the government health-care system. In addition, it would promote more health-care workers to serve in remote locations. The new payroll system needs parliament approval prior to implementation.

---

## Physical and Human Resources

### Physical Resources

The number of health facilities has been growing rapidly over the last two decades (Table 5). These consist of both public and private facilities with



**Table 5** Summary of health facilities (2005)

	Number	Beds
MOHP	1,166	78,502
Rural	669	11,093
Rural (integrated) hospitals	439	8,509
Urban	497	67,406
General and district hospitals	233	34,656
Obstetric and pediatric hospitals	10	752
Mental hospitals	17	6,415
Teaching hospitals and institutes (THO)	18	5,639
Curative care organization (CCO)	11	2,129
Health insurance organization (H10)	40	9,828
Other ministries' hospitals	119	29,851
Medical schools	71	25,891
Police and prison	26	1,382
Private hospitals	1,329	15,302

Source: Egypt MOHP data, cited by WHO 2006

varying services and amenities. Most facilities offer at a minimum basic structural needs (i.e., electricity and water) along with at least one doctor (MOHP 2004). Maternal, child, and reproductive health services have continued to increase with urban areas showing the highest percentage (MOHP 2004). Overall the private sector has access to more/better medical equipment compared to public facilities which have been continuously underfunded (WHO 2010; Ministry of Health, Egypt 2010; WHO-EMRO 2006). This has not changed since the Arab Spring revolution.

## Human Resources

Historically in efforts to make the health-care system more independent, health professions were encouraged and looked upon highly in society. However, with that status, many health-care workers have been unwilling to practice in rural areas of Egypt. This has continued the cycle of a poor distribution of health services in these areas (Jabbour 2012). Efforts had been made by previous governments to encourage practitioners to work in rural areas, but discontent among health professionals with

**Table 6** Summary of health workers (2005)

	Number
Physicians	12,917
Dentists	3,885
Pharmacists	1,277
Nurses	44,300
Lab technicians	3,575

Source: CAI HC data, cited by WHO 2006

these policies has hindered their implementation (Jabbour 2012). Progress has been made in eliminating a job guarantee following medical school. However, there has always been a lack of communication between the universities producing doctors and the government overseeing policies. Prior to 2011, there were 6.53 physicians and 13.75 nurses per 10,000 people registered in Egypt (WHO-EMRO 2006). There is little to no data as to whether the size of the workforce has been adequate to this point (WHO-EMRO 2006). Little has changed in human resources apart from new payroll incentives to encourage health professionals to practice in rural areas (Table 6).

## Provision of Services

### Overview

The utilization of health services in Egypt is highly reliant upon the division of public and private sectors of health care. For the most part, the majority of health facilities are run by the MOHP. However, the dichotomous system does not allow for health provision completely by the public sector because the latter is chronically underfunded. In contrast, inpatient care is mainly provided by the MOHP/public sector, while ambulatory and pharmaceutical care is mostly private (WHO-EMRO 2006). While universal health-care coverage has not yet been achieved within the country, 100% of the Egyptian population has access to basic health services (WHO 2010).

For the most part, the MOHP oversees hospitals and outpatient facilities. Other public facilities

consist of HIO service providers, university and teaching hospitals, and institutes along with other various government-run facilities. The private sector is also extensively providing hospitals, pharmacies, outpatient facilities, as well as traditional healers to name a few. Taking all of these providers into account, 95% of the population is within 5 km of a medical facility (Elgazzar 2009). Of this, public facilities coordinated by the MOHP had grown to over 5000 health facilities with over 80,000 beds in 2011 (Haley and Beg 2012). Despite the growing number, these facilities are underutilized by up to 50%, and over 60% of primary care visits happen in the private sector (WHO 2010; Gericke 2006). This figure has shown little improvement over time as in 1994–1995, public hospitals showed an occupancy of only 45% (with other sources showing even lower utilization rates) (WHO-EMRO 2006).

While so much of the population is located close to health facilities, the actual provision of services has been somewhat of an issue between rural and urban areas along with discrepancies among different socioeconomic populations. This has been the result of a lack of optimal development in rural areas despite the Health Reform Program as well as the lack of affordability of different forms of care. As a result, 70% of outpatient care is obtained privately by wealthier populations along with longer inpatient stays (Elgazzar 2009).

### **Inpatient Care**

Inpatient care is mainly provided through government-funded health facilities. Eighty-five percent of all inpatient care in 2005 was through government MOHP or public facilities (i.e., HIO) (WHO-EMRO 2006). Admission rates on average were 0.029 per capita per year which is within the upper-middle range for comparable developing countries (WHO-EMRO 2006). The average length of stay for inpatient care was 3 days; however, wealthier patients have reported up to 1.5 times that (Elgazzar 2009).

### **Outpatient Care**

Unlike inpatient care, the majority of outpatient care is provided by private health facilities. On average only 1.4 out of 3.98 total outpatient visits per capita per year occurred within MOHP facilities (WHO-EMRO 2006).

### **Mental Health Care**

There are few large psychiatric hospitals in Egypt. For the most part, these facilities have remained fairly centralized providing inpatient care considered inadequate by many observers (WHO 2010). The majority of the issues stem from inpatient care surrounding the provision of acute mental health care as 60% of the beds are occupied by long-stay patients (WHO 2010). Overall, there has been an increased recognition of the importance of mental health care. However, in Egypt spending and mental health regulations have not kept up pace with these increased expectations (Jenkins et al. 2010). This has led toward a severe lack of staff, resources, and information regarding mental illness (Roberts et al. 2013). Therefore, increased funding and recognition by the MOHP, as well as other service providers, are necessary in order to redress this situation.

### **Pharmaceuticals**

Pharmaceutical expenditure accounts for the second largest part of total health spending in Egypt. This is a result of the majority of pharmaceuticals being produced in Egypt. The benefits of this have allowed for increased immunization rates in children and the general population. However, many of the pharmaceuticals have not fully met world standards, and with the lack of communication within the health system, there have been issues with management based on needs (WHO-EMRO 2006). Leading the control of pharmaceuticals in Egypt is the Central Administration for Pharmaceutical Affairs (CAPA) within the MOHP. This group has been able to positively influence pharmaceutical developments through a decree

establishing a clinical pharmacy unit and drug information center in every public and private hospital in order to empower and educate patients on medication issues. Furthermore, the use of pharmaceuticals will continue to develop as the MOHP health technology assessment and pharmacoconomics unit will work to better utilize pharmaceutical resources and expenditure.

---

## The Arab Spring Revolution

Taking inspiration from Tunisia, in January of 2011, Egyptian protesters working alongside the only organized opposition force, the Muslim Brotherhood, stormed the streets of major cities in Egypt in order to protest against the current Egyptian regime. This revolution succeeded to overthrow President Hosni Mubarak. Under the Mubarak presidency between 1981 and 2011, there were many grievances over questions of freedom of expression, other human rights issues, as well as social and economic issues. The revolution followed a number of years with high unemployment rates, low wages, as well as food price inflation. The overall goal of the revolution was to end the president's regime along with the country's policy on emergency law, lack of freedom of speech, and overall corruption from the government. The protests varied from peaceful to violent and lasted a total of 28 days until the president was finally overthrown.

In efforts to disassemble the protests, the Egyptian government attempted to eliminate social media the night before the protests started. While this was somewhat successful, the protests still filled the streets the next day resulting in President Mubarak dismissing his government, appointing a new cabinet and vice president Omar Suleiman (the first in 30 years) all in hopes of satisfying the uprising masses. However, protests did not resist until President Mubarak handed power over to the Armed Forces placing Egypt in a truly transitional state (Abou-El-Fadl 2012). Under the Armed Forces oversight, a new prime minister, Essam Sharaf, was announced, the Egyptian Parliament was dissolved, and the Egyptian constitution was put on hold. Following the revolution, Egypt

worked to recover with the help of other countries such as financial assistance from the United States which further increased Egypt's already large debt (Hamilton 2013).

It was not until June 2012 that Egypt finally elected a new president. Promising to end years of presidential abuse of power, Mohammed Morsi was sworn in (Hamilton 2013). Within his first year of office, Egypt began importing natural gas. This investment was to the benefit of the nation's richest businessmen and increased public spending on fuels to 25% of all public expenditure – more than what the country spends on health and education combined (Hamilton 2013). Also within this term, with influence from the Muslim Brotherhood, Morsi broke a number of electoral promises. In summary, his actions brought no improvements to social issues nor fulfilled the goals of a new constitution to be improved following the revolutionary demands. Because of this, the Egyptian people once again took to the streets in order to overthrow their new president along with the newly developed constitution. The Armed Forces sided with the people. On 3 July 2013, President Morsi was overthrown by the military's coup d'état, and he alongside with other leaders of the Muslim Brotherhood was arrested and put to trial. After a series of violent demonstrations and bombings on police and military institutions as well as on Coptic Christians and churches, the Muslim Brotherhood was declared a terrorist organization in December 2013. The return to a military government has led to new uncertainty and a continuation of an economically and socially unstable condition. In 2014, General Abdel Fattah El-sisi was elected as the sixth president of Egypt.

---

## Reforms

### Overview

The 2011 revolution made way for huge changes within the country. For the most part, the population recognizes the challenges caused by a rapidly growing population alongside an out-of-date public-sector health-care provision (Devi

2013). And while attention has been drawn toward this along with the long-standing demands for more funding for health care and prevention, the unstable government and economy have not been able to make it a priority (Devi 2013). The past Health Minister Hamed has even stated recently the “20% of hospitals in the rural south had no doctors and only 40% of necessary medicines were available in government hospitals and clinics.” There is much optimism that the revolution will ultimately increase public spending in the health-care sector in the future. However, so far it has only resulted in addressing some immediate health concerns (i.e., deaths, injuries, public displacement, and general deterioration of public health-care facilities) (Coutts et al. 2013). The revolution has hurt the economy by not only reducing Egypt’s number one source of income (tourism) but also reducing global investments and directing the limited funds in other places of need (Haley and Beg 2012). Therefore, regarding the health-care sector, Egypt is mostly left with the reforms set in place prior to the revolution along with the hope of improved health care in the public sector through future increased spending and development.

### Past Reforms

Within the past couple years, the government of Egypt has made an improvement in salary bases for doctors and pharmacists to meet people’s dissatisfaction. Besides this, there have been few minor reforms since the implementation of the Health Sector Reform Program strengthening primary care. This allowed for greater service delivery innovation with the implementation of the Family Health Model to provide better access to integrated services at a higher quality (WHO-EMRO 2006). This also accounts for the development of the Family Health Fund to help with payment and financing for the program. Another step was made toward the goal of universal health care through increasing public health insurance benefits and general coverage. In 2014, Health Minister Elrabat stated that “there is a plan to cover all people by the new health insurance system within 7 years.”

### Proposed Plans

In the near future, the goal will be to finalize the implementation of the new social health insurance in order to drastically improve health insurance coverage for the general population (both rural and urban), to reduce out-of-pocket spending (from 72% to 35%), and to provide Egyptians with the freedom of choice in terms of provider and treatment location. This plan requires the government budget to grow from 4.7 to 8% at the end of the implementation with health insurance spending to represent more than 50% of total health expenditure. The implementation of a health technology assessment (HTA) system and the recent establishment of a pharmacoeconomics unit at the Egyptian Drug Authority are seen as a promising step to reduce the exceedingly high amount of pharmaceutical expenditures compared to other subsectors in the health system. Implementation of these plans will start in three to five governorates for 3 years as a pilot followed by a gradual rollout to the whole nation if successful.

Besides these plans, a focus of future reforms in Egypt will be to continue to develop primary care and prevention outreach (Roberts et al. 2013). One of the key goals in this area includes ensuring that primary care is provided in rural areas, free medical treatment in hospitals, subsidized treatment of children under 6 who are not covered by insurance, and regulation of private health-care providers (Roberts et al. 2013). The Egyptian government (through the HSRP) will continue to work on developing their universal health care through further development of the Family Health Model and ensuring more equity and access to care through increased public spending on health (WHO 2010; WHO-EMRO 2006). Health sector reform is paving the way through the new health insurance system that aims to provide easy access to affordable basic health services to all Egyptians, rich and poor, urban and rural, young and old. Plans also include increasing the percentage of GDP allocated to health care, in addition to better utilizing those resources.

## Assessment

The state of Egyptian health care is fragile and fragmented given the country's unstable conditions over the last few years. However, the motivation of protest gives hope that the new government will improve the health-care system through further reforms. The largest issue currently hindering the system in 2014 is very limited funding. The government has continuously underfinanced the public health sector and thus created a huge gap in inequality based on what individuals or their families can afford. Therefore, with the organization of the new constitution and government following the current unrest, one of the first issues to be addressed will be increasing public funding for health care (Gericke 2005). This will allow for the much-needed finances to go toward upgrading and maintaining public health resources as well as increased payment to health-care workers in the public sector. This should also come with subsequent policies to help distribute the resources as well as recognize doctor salaries in order to create an equal distribution of health workers throughout the country (Haley and Beg 2012). With little to no change in regard to spending on health within the last decades, this assessment is not new.

This stagnant health system has maintained minimal communication between private and public health care. This is problematic for overall health. With a lack of government finance and focus on the public sector comes a subsequent lack in the regulation of the private sector which needs immediate change (Gericke 2005). While the private sector continues to develop and grow in Egypt, there is an uneven distribution of care between the two sectors. There need to be an increased focus and spending in the public sector in order to further come close to achieving the goal of a universal coverage because as of now only the private sector seems to continue developing and providing good care to those who can afford it. Care provision also needs to spread to the rural areas of Upper Egypt to reduce inequalities in access to care. Likewise, the private health-care sector needs to be better monitored for quality and, in particular for the pharmaceutical market, cost-effectiveness and price controls.

Health indicators in general have improved throughout the years despite the low levels of health spending. However, concerns with regard to noncommunicable and chronic diseases are rising. Therefore, a new strategy is needed to fight these as both smoking and obesity are increasing along with their subsequent poor health outcomes. There is also a need to increase disease surveillance and to work on improving chronic disease control and prevention (Devi 2013; Coutts et al. 2013). Hepatitis C (with its high incidence and prevalence) in Egypt poses what some government officials have publicly addressed as a national security concern. There is a growing need to set priorities along with a written and documented plan on how to proceed to face and solve some of these problems. Finally, greater efforts are required to address both the constraints and gaps in provision of comprehensive reproductive health care in Egypt, including stronger coordination mechanisms among various stakeholders and the need for more effective partnerships with civil society and the private sector.

In conclusion, the Egyptian health system does provide an extensive infrastructure with regard to the number of physicians, clinics, pharmaceuticals, and physical access to hospitals (Gericke 2006). However, with remaining inequalities, there ultimately needs to be a national strategic plan for the next 5–10 years to address the persisting issues. The very frequent change in health ministers (seven between 2010 and 2013) makes it very difficult to move forward with a constant set of objectives. Health care needs to grow in Egypt by diminishing inequality through the spread of affordable care to even the most remote areas and all population segments. Therefore, there is a need to build up the MOHP to create a more structured organization from which health reforms and improvements will come from. Improvements in the health sector cannot be seen in isolation and require parallel, substantial improvements to education, health promotion, safe water and housing, and traffic regulations to name only a few. Despite the recent stabilization of government, the continuing economic problems and low public spending on health have thwarted most attempts at reforming health that have been discussed in recent years.

## References

- Abou-Ali H, et al. Evaluating the impact of Egyptian social Fund for Development programmes. *J Dev Eff*. 2010;2(4):521–55.
- Abou-El-Fadl R. Beyond conventional transitional justice: Egypt's 2011 revolution and the absence of political will. *Int J Transit Justice*. 2012;6(2):318–30.
- Anwar WA. Environmental health in Egypt. *Int J Hyg Environ Health*. 2003;206(4–5):339–50.
- CIA. The World Factbook: Egypt. 2013 [cited 2013 June 27th]. Available from: <https://www.cia.gov/library/publications/the-world-factbook/geos/eg.html>.
- Coutts A, et al. The Arab SPRING and health: two years on. *Int J Health Serv*. 2013;43(1):49–60.
- Devi S. Women's health challenges in post-revolutionary Egypt. *Lancet*. 2013;381(9879):1705–6.
- Egypt's progress towards achieving the Millenium Development Goals. Ministry of Economic Development, Cairo; 2010. p. 1–154.
- Elgazzar H. Income and the use of health care: an empirical study of Egypt and Lebanon. *Health Econ Policy Law*. 2009;4(4):445–78.
- Gericke CA. Comparison of health care financing in Egypt and Cuba: lessons for health reform in Egypt. *East Mediterr Health J*. 2005;11(5–6):1073.
- Gericke CA. Financing health care in Egypt: current issues and options for reform. *J Public Health*. 2006;14(1):29–36.
- Haley DR, Beg SA. The road to recovery: Egypt's healthcare reform. *Int J Health Plann Manag*. 2012;27(1):e83–91.
- Hamilton OR. Egypt's latest revolutionary act was profoundly democratic: London: The Guardian; 2013.
- Jabbour S. Egypt in crisis: politics, health care reform, and social mobilization for health rights. In: *Public health in the Arab world*. Cambridge/New York: Cambridge University Press; 2012. p. 477–88.
- Jenkins R, et al. Mental health policy and development in Egypt – integrating mental health into health sector reforms 2001–9. *Int J Ment Heal Syst*. 2010;4:17.
- Ministry of Health, Egypt. National Health Accounts 2007/2008: Egypt report. In: *Health systems 20/20 project*. Bethesda: Abt Associate Inc; 2010. p. 1–45.
- MOHP. Egypt service provision assessment survey 2004. 2004. p. 1–410.
- MOHP. 2013 [cited 2013 July]; Available from: <http://www.mohp.gov.eg/about/OrgChart/default.aspx>.
- Mosallam RA, Aly MM, Moharram AM. Responsiveness of the health insurance and private systems in Alexandria, Egypt. *J Egypt Public Health Assoc*. 2013;88(1):46–51.
- Roberts B, et al. The Arab Spring: confronting the challenge of non-communicable disease. *J Public Health Policy*. 2013;34(2):345–52.
- Salem MA. Policy Research in Egypt's Health Sector Reform. The Alliance for Health Policy and Systems Research; 2002. Working paper no. 13.
- WHO. Country cooperation strategy for WHO and Egypt 2010–2014: Geneva: World Health Organization; 2010. p. 1–52.
- WHO. Egypt: health profile: Geneva: World Health Organization; 2013. p. 1–2.
- WHO-EMRO. Health System Profile Egypt: Regional Health Systems Observatory, EMRO, World Health Organization, Cairo; 2006. p. 1–111.
- World Bank. World Bank Data: Egypt. 2013 [cited 2013 June 12th]; Available from: <http://www.worldbank.org/en/country/egypt>.



Karine Chevreur and Karen Berg Brigham

## Contents

<b>Introduction</b> .....	828
<b>Organization and Governance</b> .....	828
<b>Financing</b> .....	829
<b>Physical and Human Resources</b> .....	831
<b>Delivery of Health Services</b> .....	831
Primary Care .....	831
Hospital Care .....	832
Integrated Care .....	832
Long-Term Care .....	833
Disabled Adults and Children .....	833
Mental Health Care .....	833
Pharmaceutical Care .....	834
Public Health .....	834
<b>Reforms</b> .....	834
<b>Assessment</b> .....	835

## Abstract

The French Republic is comprised of metropolitan France located in western Europe and a collection of overseas islands and territories on other continents. It is a unitary state with administrative subdivisions: 100 departments (local authorities) embedded in 27 regions. On

January 1, 2013, the French population totaled 63.7 million inhabitants in metropolitan France and 2.1 million inhabitants in the overseas territories. France is the second most populous country in the European Union (EU), and over three-quarters of its population lives in urban areas. It has the fifth largest economy in the world. The French political system is a parliamentary democracy with a president and a bicameral parliament consisting of a National Assembly and a Senate. France is a welfare state that developed its social security system after the Second World War with the aim of covering the financial risks associated with

K. Chevreur (✉)  
Health Economics and Health Services Research Unit,  
URC ECO Ile de France, Paris, France  
e-mail: [karine.chevreur@urc-eco.fr](mailto:karine.chevreur@urc-eco.fr)

K. B. Brigham  
University of Washington, Seattle, WA, USA

getting sick, being injured in the workplace, getting old, and growing families.

---

## Introduction

The overall picture of the state of health in France contains apparent contradictions. On the one hand, indicators such as life expectancy, life expectancy without disability, and healthy life expectancy show that the health of the population is good. The French average life expectancy is now over 80 years and is the second highest in the world for women. Moreover, the French population is aging, and from 2020 onwards, those aged over 60 will outnumber those aged under 20 (accounting for 27% and 23% of the population, respectively). The aging of the population is not due to a decreasing fertility rate as in other European countries. Indeed, France has the third highest fertility rate in the EU. In addition, older people remain in better health than in many other European countries.

The main causes of death in France are cancer, cardiovascular diseases, accidents, and diseases of the respiratory system. However, France also compares well with regard to cardiovascular diseases, while its relative position with respect to mortality caused by alcoholism, cirrhosis, and cancer of the cervix is improving. Nonetheless, France suffers from a high rate of premature male deaths from accidents and unhealthy habits such as smoking and alcoholism that are the most common causes of avoidable mortality in France. Additionally, France has long reported health inequalities across socioeconomic groups that are wider than in most other European countries. These inequalities result not only from risk factors, but also from disparities in access to health services that require the highest out-of-pocket expenditure by patients.

---

## Organization and Governance

The French health care system is of a mixed type, structurally based on a Bismarckian approach with Beveridge goals reflected in the single public payer model, the increasing importance of tax-based revenue for financing and strong state intervention.

There is Statutory health insurance (SHI), which covers virtually 100% of the resident population under various noncompeting schemes. The delivery of care is shared among private, fee-for-service physicians, private profit-making hospitals, private non-profit-making hospitals, and public hospitals. In addition to the health care sector and the social sector, there is a health and social care sector, known as the third sector, which provides care and services to elderly and disabled people.

Jurisdiction in terms of health policy and regulation of the health care system is divided among the state (parliament, government, and the Administration of Health and Social Affairs), SHI, and local authorities, particularly at the regional level. Reforms over the last two decades have attempted to devolve a greater remit in governance and health policy decision-making to the regional level, particularly with respect to planning. This trend culminated in the 2009 Hospital, Patients, Health and Territories Act (*loi hôpital patients, santé et territoires*; HPST), which merged institutions representing the main stakeholders (the state, SHI schemes, health professionals, and public health actors) at the regional level into “one-stop shops,” the 26 regional health agencies (*agences régionales de santé*; ARS). Cutting across the traditional boundaries of the health care sector, the public health preventive sector, and the health and social care sector for disabled and elderly persons, the ARSs are responsible for ensuring that health care provision meets the needs of the population by improving coordination between the ambulatory and hospital sectors and health and social care sector services while respecting national health expenditure objectives.

Planning and regulation involve negotiations among provider representatives (hospitals and health professionals): the state, represented by both the Ministry in charge of Health and the Ministry in charge of the Economy and Finances, and SHI. The outcome of these negotiations is translated into administrative decrees and laws passed by the parliament. These include public health acts, social security funding acts, and reform acts. In the context of increasing health



care expenditure and the increasing SHI deficit, the role of the state in planning and regulation has increased over the past two decades. The responsibility for capacity planning is shared by the central and regional levels. At the regional level, the ARSs coordinate ambulatory and hospital care and health and social care for the elderly and disabled through a regional strategic health plan (*plan stratégique régional de santé*; PRS) based on population needs. Each sector's planning process must comply with the PRS which, starting in 2010, represents the first attempt at regional planning of the ambulatory care sector.

Providers are paid by SHI (or directly by patients who are later reimbursed). The statutory tariffs are set through negotiations between providers and SHI and are approved by the Ministry in charge of Health. Quality of care is regulated at the national level. Hospitals must undergo a certification process every four years, but there is no formal re-certification or re-licensing process for health professionals. However, doctors, pharmacists, dentists, and midwives are required to follow lifelong learning activities through professional continuous development.

The role of patients in regulation and planning has slowly increased in recent years, although their participation remains marginal. The 2009 HPST law created the Regional conference on health and autonomy (*Conférence régionale de la santé et de l'autonomie*; CRSA) through which patients and their representatives may participate in defining public health priorities at the regional level, including development of the PRS. Patient input is stronger at the services level.

Health information systems and technologies have been developed to help in planning and regulation. The SHI inter-schemes system (*système national d'information interrégimes de l'assurance maladie*; SNIIR-AM) was established in 2003. It encompasses information on patient health care consumption for which a claim has been sent to SHI, regardless the type of care (hospital inpatient stays, self-employed doctor visits, drugs...) as long as it is covered by SHI. This system has been facilitated by the development of electronic billing, which has been implemented in the ambulatory sector since the mid-1990s via an

individual health insurance electronic card (*carte Vitale*) on the patient side and an electronic identification card for health workers (*carte de professionnel de santé*; CPS) on the provider side.

Additionally, in order to improve quality of care and decrease redundancy in consumption, the development of an electronic patient record (*dossier médicale personnel*; DMP) to group medical information and care consumption in ambulatory and hospital settings for patients on a voluntary basis was initiated in 2004. Implementation has not been smooth due to both technical and patient privacy concerns. However, by June 2013, nearly 350,000 patients had DMPs, which are now used by 4800 health professional in the ambulatory sector and 350 institutions in the hospital sector.

---

## Financing

Financial responsibility for health care in France is mainly borne by SHI. However, SHI only funds around three-quarters of health spending, leaving considerable scope for complementary sources of funding, such as private voluntary health insurance (VHI). Moreover, funding for long-term care for the elderly and disabled is financed differently. It is partly financed by a dedicated fund created in 2004, the National Solidarity Fund for Autonomy (*Caisse nationale de solidarité pour l'autonomie*; CNSA). Its resources come from SHI and the "solidarity and autonomy contribution" that is generated from the revenue of an unpaid working/solidarity day (*journée de solidarité*) contributed by the French working population. Local authorities and households also participate in financing these categories of care.

SHI resources mainly come from an earmarked tax called the "general social contribution" (*contribution sociale généralisée*) based on total income and not only on earned income as was previously the case. Additional revenue accounts for around 13% and comes from specific taxes such as "sin" taxes or taxes on the pharmaceutical companies' turnover. Funds are pooled at the national level, and there is no formal allocation mechanism in France.

SHI coverage is established according to resident status, and entitlement is based on employment, unemployment, student, or retiree status. Since the introduction of universal medical coverage (*couverture maladie universelle*; CMU) in 2000, the state has covered the health care costs of residents not otherwise eligible for SHI. Illegal residents who have applied for residency are covered by a special program (*aide médicale de l'état*; AME).

SHI covers a broad range of services and goods that are provided in hospital or defined in positive lists for outpatient care. In Europe, the level of coverage is considered quite generous, offering rapid access to the latest innovations. The rate of coverage varies across goods and services; for example, the co-insurance rate is 30% for physician and dentist care, 40% for ancillary services and laboratory tests, and 20% for hospitalization. For most drugs, co-insurance amounts to either 35% or 70% but ranges from 0% for nonsubstitutable or expensive drugs to 85% for "convenience medications." However, there are several conditions for which patients are exempted from co-insurance, such as chronic conditions covered under the ALD scheme (*affections de longue durée*) or pregnancy after the fifth month. Co-insurance amounts are generally covered by VHI, which provides reimbursement for co-payments and better coverage for medical goods and services that are poorly covered. However, deductibles introduced after 2004 with the aim of improving coordination of care and reducing patient consumption cannot be covered by VHI or else the insuring entity will be subject to financial penalties.

Over recent decades, VHI has gained an important role in ensuring equity of access and financing of health care. It covers 88% of the population on a private basis. Since 2000, in order to ensure that the measures increasing patients' co-insurance would not result in increased social inequities in access, public complementary insurance (*couverture maladie universelle complémentaire*, CMU-C) has been offered on a voluntary basis to lower socioeconomic groups and covers 6% of the population.

SHI pays for hospital acute care by means of a DRG-type payment method (*tarification à l'activité*; T2A). In addition to the 20% co-insurance amount, a hospital catering flat fee amounting to €18 per day is the responsibility of patients or their VHI. Self-employed professionals are paid on a fee-for-service basis and patients are reimbursed based on official tariffs. However, certain self-employed doctors are allowed to practice extra-billing, which impairs the equity of access objective of the system. Financial incentives to improve the quality and efficiency of doctors' practices and to decrease the level of extra-billing exist. Individual contracts with general practitioners including with pay for performance target were initially implemented in 2009 and extended to specialists in 2012. From 2012, measures designed to rein in excessive extra billing include a new voluntary "Access to health care." In exchange for maintaining their extra-billing fee practices at 2012 levels, doctors benefit from social and fiscal advantages.

In 2012, total expenditure on health in France was estimated at €243 billion or 12% of gross domestic product (GDP). Expenditure on personal health care accounted for three-quarters of total health expenditure (€183.6 billion), representing an average €2806 per person. Of this, 75.5% was publicly funded, with complementary voluntary health insurance (VHI) financing 13.7% and households covering 9.6% in out-of-pocket costs. As in other European countries, health care expenditure has steadily increased. As a result, since the late 1990s, SHI annual expenditure has been capped by a national ceiling on SHI expenditure (*objectif national des dépenses assurance maladie*; ONDAM) approved by the parliament. It is split into subtargets that cover hospital expenditure, social, and health care services for elderly and disabled, privately delivered care. While there is no formal allocation mechanism, this has provided SHI with a tool to allocate health care expenditure between broad sectors. If the health care system is found to exceed its projected budget by more than 1%, a special parliamentary Alert Committee can ask the head of the Directorate of Social Security (the watchdog for

all social security branches) to present a financial rescue plan.

---

## Physical and Human Resources

In France, there is a high level of facilities, equipment, and other physical resources. However, there are strong disparities in geographic distribution, and France is well below the EU average for MRI units (7 per million population, compared to the EU23 average of 10.3) and CT scanners (11.8 per million population, compared to 20.4).

There are four main categories of hospitals: regional hospitals, general hospitals, local hospitals, and psychiatric hospitals. Capital investment is either covered by reimbursements for services delivery or funded through specific programs. Two nationwide investment plans were launched in the last decade in order to improve quality and safety standards. The ARSs are responsible for the control of capital investment and purchases of major medical equipment.

Following the general trend in European countries, the number of full time acute beds per 1000 inhabitants has steadily declined over the last 20 years. In 2010, it was 6.4, which is above the EU27 average of 5.3. Reduction in acute care capacity was accompanied by the transformation of acute beds into rehabilitation and long-term care units and the development of day surgery and hospitalization at home.

Nurses and nursing aides form the largest group of professionals, accounting for approximately half of the health care workforce. Registered health professionals also include medical professionals (physicians, dentists, and midwives), pharmacists, professionals involved in rehabilitation (physiotherapists, speech therapists, vision therapists, psychomotor therapists, occupational therapists, and chiropodists) and technical paramedical professions (hearing aid specialists, opticians, and radiographers). The other professions usually identified as contributing to health care include clerical and technical staff working in hospitals, laboratory technicians, pediatric auxiliaries, dieticians, psychologists, and ambulance drivers.

About 7% of the French population works in the health care sector. The number of practicing doctors per 1000 population is slightly lower than the EU27 average (3.3 vs. 3.4), although in France the number includes not only those providing direct patient care, but also managers, educators, researchers, etc. The number of practicing nurses exceeds the EU27 average (8.5 vs. 7.9), and the ratio of nurses to physicians is 2.6, just above the EU average. Workforce forecasting and careful planning of educational capacity is mostly made at the national level through the use of *numerus clausus* for medical professionals. It seeks to prevent shortages or oversupply of health professionals. However, it does not control for the geographical distribution of medical professionals, as self-employed professionals are free to choose where they practice. In order to solve the resulting great disparities in the distribution of medical professionals, there has been increasing transfer of tasks from medical to other professionals such as nurses and development of incentives for attracting health professionals to under-served areas.

---

## Delivery of Health Services

The delivery of care is shared among private physicians, private profit-making hospitals, private non-profit-making hospitals, and public hospitals. In addition to the health care sector and the social sector, there is a so-called “third sector” which provides both care and social services to elderly and disabled people.

### Primary Care

Primary care is mostly delivered in the ambulatory care sector by self-employed professionals who are paid on a fee-for-service basis by patients who receive partial reimbursement from the SHI funds (i.e., co-insurance payments apply). Since the late 1990s, GPs have gained a major role in the coordination of care with the implementation of a semi-gatekeeping system that provides incentives to people to visit their GP prior to consulting a specialist.

## Hospital Care

Hospital care is delivered by public, private non-profit-making, and private profit-making hospitals. Acute medical, surgical, and obstetric care is provided by public as well as private hospitals, with different areas of specialization.

Acute medical care is mainly provided by public hospitals, which account for three-quarters of acute medical care capacity (80% of medical beds and 70% of day-care beds) and perform 75% of full-time episodes and 55% of day-care episodes. Private profit-making hospitals account for 10% of full-time beds and 20% of day-care beds, and they provide 15% of full-time episodes and 40% of day-care episodes; they specialize in a small number of technical procedures for which there are profit opportunities, such as invasive diagnostic procedures (e.g., endoscopies or angiograms). The balance of the acute medical activity is performed by the private non-profit-making sector, which are the main providers in the area of cancer treatment.

Surgical care is mainly delivered by private profit-making hospitals, which perform more than half of all surgical procedures, including 75% of the surgical episodes performed in day-care settings. Surgical care accordingly represents more than half of the acute care activity of the private profit-making sector. These hospitals tend to specialize in procedures that can be performed routinely within a short stay with a predictable length; for example, they perform three-quarters of surgery for cataracts and varicose veins and two-thirds of surgery for carpal tunnel syndrome. Public hospitals perform a third of surgical procedures, with a much wider scope than profit-making hospitals, including the most complex procedures. Surgical procedures performed in the private non-profit-making sector are mostly related to cancer treatment, as for medical stays.

Two-thirds of obstetric procedures are performed within public hospitals, while the private sector accounts for the remaining third, mainly within profit-making hospitals, which account for one-quarter of all obstetrical stays.

Because of concerns about excess acute care capacity, alternatives to full-time inpatient care have been promoted since the late 1980s. Specifically, authorizations to develop “hospital at home” (*hospitalisation à domicile*; HAD) units, as well as ambulatory care places, have been granted in return for reducing the number of acute beds.

HAD units, which have existed in France for about 50 years, send medical or paramedical staff to the patient’s home on a daily basis in order to provide continuous and coordinated care in cases where a hospital stay would have been otherwise necessary. Administratively, the units are either hospital departments or private mainly non-profit-making associations. Each unit is led by a physician, who takes responsibility for the overall coordination of medical care, while nurses coordinate individual treatments; actual care is provided by salaried staff from the hospital or self-employed professionals. In 2011, there were about 305 HAD units that cared for more than 100,000 patients, mainly in the areas of palliative care, cancer treatment, and perinatal care.

Ambulatory surgery accounted for only 40% of surgical hospital stays in France in 2011, compared to nearly 80% in the UK. The Minister of Health has set a target for ambulatory surgeries to exceed 50% of all surgeries by 2016.

## Integrated Care

The 2002 the Patients’ Rights and Quality of Care Act brought together diverse provider network initiatives under the concept of “health networks” with the aim of strengthening coordination and continuity through the interdisciplinary provision of care, particularly for selected population groups and targeted diseases. The disease management provided by these networks also includes experimentation with new models of care delivery (e.g., nurses performing tasks previously reserved for doctors). Participation is voluntary both for patients and providers. Patients may benefit from services not usually covered by SHI (e.g., podiatric care and dietary advice for diabetics), and physicians may be reimbursed for preventive

services and patient education not otherwise covered. Physicians receive additional compensation for coordinating the care of patients with certain chronic diseases (€40 per patient per year).

### Long-Term Care

Long-term care for elderly and disabled is provided through both residential care and home care and falls under the third sector, which combines elements of medical and social care. The French population aged over 75 years is expected to nearly double by 2050, when it will constitute 15.6% of the population compared to 8% today. Thus, there is an increasing need for long-term care services for frail elderly persons at home or in nursing facilities or other residential care settings. In 2010, French long-term care spending was estimated at €34 billion, or 1.73% of GDP, of which 70% was publicly funded.

Home care is mainly provided by self-employed physicians and nurses and, to a lesser extent, by community nursing services (*services de soins infirmiers à domicile*; SSIAD), which deliver nursing care at home mainly using employed auxiliary nurses and to a lesser extent nurses, who are mostly self-employed.

Residential care for elderly people is provided by many types of institution offering different levels of service. These include **collective housing facilities** (*foyers logements*), offering a range of nonmedical facilities (such as catering and laundry) and almost no medical care; **retirement homes** (*établissements d'hébergement pour personnes âgées*; EHPA), which accommodate the elderly but also offer medical care; and **long-term care units** (*unités de soins de longue durée*; USLD), which accommodate people whose care requires constant medical monitoring. These units are provided in autonomous nursing homes or in hospital wards for very sick and dependent people.

In the early 2000s, intermediary services were created to receive frail elderly persons not living in residential services for short periods. They care for patients on a daily basis (*accueil de jour*) or on a temporary basis (*accueil temporaire*) with the

goal of offering respite care for families and day care for patients with Alzheimer's disease and other dementias.

### Disabled Adults and Children

About 3.2 million people are registered as disabled in France, of whom 1.8 million are affected by a severe disability that limits their functional autonomy. Disability is measured in terms of an incapacity rate, which takes into account the degree of difficulty with daily living. Specific committees for children and for adults at the department level evaluate the rate of incapacity and determine the right to certain benefits. They also have the authority to refer the disabled person to a specialized institution.

Around 200,000 disabled adults are accommodated in 4800 dedicated facilities. Different institutions provide a range of services for disabled adults with different levels of functional autonomy. Nearly 130,000 disabled children are cared for in 2500 facilities. A large number of institutions offer treatment, special education, and vocational training to children affected by motor, cerebral, or intellectual disabilities.

Disabled individuals may be eligible for monetary allowances. The disability compensation allowance (*prestation de compensation du handicap*; PCH) may be used to finance the wages of aides to disabled people or their families or any necessary technical devices. The allowance is funded by the general councils, the CNSA, and the CSG funds and is not means tested.

### Mental Health Care

Mental health care is delivered by both the health sector and the social and health care sector. As in many other European countries, mental health care policy in France during the second half of the twentieth century was influenced by a general movement towards community-based organization of mental health care services – the so-called “deinstitutionalization” process. Services provided by the health sector take the

form of both public and private outpatient and inpatient care.

Adult public mental health care is provided within around 800 geographical areas that cover theoretically equivalent populations of approximately 60,000 inhabitants aged 16 or more, called mental health care areas (*secteurs de soins de santé mentale*; MHC). Care within each area is coordinated by a hospital (a public hospital in more than 90% of the cases) and includes a wide range of preventive, diagnostic, and therapeutic services, which are provided in both inpatient and outpatient settings. In particular, ambulatory care centers (*centres médico-psychologiques*; CMP) are present in almost every MHC area; they provide primary ambulatory mental health care, including home visits, and direct the patients towards appropriate services. The size and resources of MHC areas are quite heterogeneous.

Public mental health care for children follows a similar territorial organization, with 321 areas covering an average of 46,000 people aged under 20 years (corresponding to an average of 210,000 inhabitants). These MHC areas for children show even wider geographical inequalities.

## Pharmaceutical Care

France is the fourth largest market for pharmaceutical drugs in the world and the second in Europe after Germany. Drugs are dispensed by self-employed pharmacists, while the price of drugs is set administratively for all drugs covered by SHI. Pharmacies have a monopoly on the dispensing of medicines. As a general rule, retail pharmacies must be owned by a qualified pharmacist or by a group of pharmacists associated in a company; these pharmacists or companies cannot be proprietors of more than one pharmacy. This number of pharmacies is regulated by a numerus clausus that takes into account both the size of the population to be served and the distance involved in getting to the nearest pharmacy. There were about 22,000 retail pharmacies in 2012. Since June 2008, pharmacies have been allowed to sell a limited range nonprescription

drugs “over the counter” on shelves directly accessible to patients.

A number of measures have been taken to try to improve and limit the prescribing behavior of physicians and as well as the consumption patterns of patients. The promotion of generic drugs, largely nonexistent until recently owing to the relatively low price of drugs in France, first occurred in the 1990s. The rate of generic substitution increased to 83% in 2012 from 76% in 2011. The volume of drug consumption has slowed since 2010 due to fewer prescriptions, the effect of publicity campaigns, including those to reduce antibiotic use, and removal of certain drugs from the positive list.

## Public Health

Public health policy and practice in France have historically been difficult to describe because they involve numerous actors and sources of funding, and large discrepancies exist between legislative texts and actual practice, which relies on the initiative of local actors. The 2004 Public Health Act provided a new framework for public health policy, firmly establishing the responsibility of the state in public health matters and emphasizing the role of the regional level for organizational issues. The Act also created a quantitative assessment framework for health policies encompassing public health objectives for 5-year periods that must be monitored on an annual basis and set 5-year targets for most of the related indicators. In order to meet some of these goals, several national plans have been established, such as those related to cancer; violence, addictions, and risky behaviors; environment and health; quality of life of patients with chronic diseases; and the provision of health care for patients with rare diseases.

## Reforms

The main objectives of the reforms to the health care system of the last decade were to contain SHI expenditures without damaging equity in financial access, to increase geographic equity in access to

care, and to meet the increasing demand for long-term care. Decentralization and a change in the balance of power between the state and SHI were the main instruments used to achieve these objectives.

To contain SHI expenditure, two categories of measures were used. The first, called the “strict accounting cost-containment policy,” primarily focused on decreasing the size of the benefit basket and levels of coverage, resulting in a shift towards VHI coverage. After 2004, several new mechanisms were introduced. A coordinated care pathway was implemented with higher co-insurance for patients consuming care out of this pathway, and new categories of co-payment for patients were created with the introduction of deductibles on some categories of care such as drug packages, doctor and nurse consultations, or patient transportation. Finally, there was stricter control of statutory tariffs, and starting in 2013 economic considerations have been introduced in health technology assessment of innovations.

The second category of measures was called the “medically based cost-containment policy”; it was developed in the 1990s after a long period of strict accounting policies that led to ongoing conflicts between doctors and SHI. Medically based cost-containment focuses on the reduction of financial and equity loss due to medical practice variations and aims to improve medical practice. The main tools used are the implementation of lifelong learning, the development of practice guidelines by national agencies, and the introduction of good practice commitments within professionals’ collective agreements with SHI. At first, coercive measures such as fines for not following continuous education were used to enforce this new policy, but this was slowly abandoned for a move towards the development of incentives, most recently the introduction of payment for performance for individual doctors based on meeting good practice targets. Overall, it appears that the coercive medically based cost-containment policy did not lead to major improvements in collective practice and much is expected from the pay-for-performance approach.

In order to facilitate geographical equity in access to care, the HPST reinforced local planning

and simplified regional governance of the health care system by creating the ARSs. In addition to creating the PRS, which should lead to a common approach in planning for the hospital, ambulatory, and health and social care sectors, it made formal legal provisions for the transfer of tasks between professionals. It also linked the regional medical numerus clausus to needs. In order to optimize the distribution of doctors without impairing freedom of settlement, incentives to increase the attractiveness of underrepresented specialties and medically under-served areas are being developed. For instance, wages for hospital doctors will possibly increase in contexts where there is a high need for their specialties, and contracts with medical students and self-employed health professionals with financial incentives to practice in under-served areas will be implemented on a voluntary basis.

The increasing demand for long-term care is a major concern, as the need for public funding in the coming decades is estimated to be three times higher than the expected growth of the population, thereby threatening equity in financing. Since 2005, various financing reform proposals have been debated, ranging from a newly covered risk under the social security system to targeted subsidies for private long-term care insurance. However, to date no reform measure has been enacted.

---

## Assessment

The French health care system has long enjoyed the reputation of being one of the best in the world. It has become synonymous with universal health coverage and a generous supply of health services. This reputation comes in large part from success in meeting its goals of full coverage, access without waiting lists, patient choice, and satisfaction. The combination of a basic universal public health insurance system and voluntary complementary private insurance, which provides reimbursement for co-payments required by the public system as well as coverage for medical goods and services that are poorly covered by the public system, results in low out-of-pocket costs and high medical care utilization. France’s average life

expectancy of over 80 years is in part testament to the strong combination of good health care and good public health policies in France.

Despite these positives, there also are some shortcomings, especially when considering efficiency and socioeconomic inequality in health. Major problems include lack of coordination between hospital and ambulatory services, between private and public provision of care,

and between health care and public health. Health expenditures per capita are higher than the OECD average, ranking usually third or fourth after the United States, Germany, and Switzerland, depending on the data used and year. The high level of health expenditure has become increasingly important at a time when the public system is facing chronic deficits, which are likely to increase with the current economic downturn.





Ryozo Matsuda

## Contents

<b>Introduction</b> .....	838
<b>Organization and Governance</b> .....	839
Stewardship/Governance in Health System .....	839
Dimensions of Coverage (Breadth, Scope, Depth) .....	839
Typologies of Health System .....	840
Regulating and Planning; Actors and Responsibilities .....	841
<b>Financing</b> .....	841
Sources and Collection of Revenue .....	841
Pooling of Funds and Resource Allocation .....	842
Purchasing Process and Paying for Health Services .....	842
Health Spending .....	842
<b>Physical and Human Resources</b> .....	842
Physical Resources .....	842
Intermediate Care Facilities .....	843
The Health Workforce .....	843
<b>Provision of Services: Providers, Services, Access, and Quality</b> .....	844
Public Health .....	844
Primary Care/Ambulatory Care .....	844
Specialized Ambulatory Care/Hospital Care .....	844
Pharmaceuticals .....	845
Long-Term Care .....	845
Mental Health Care .....	845
Dental Care .....	846
Complementary and Alternative Medicines .....	846
<b>Assessment</b> .....	846
<b>References</b> .....	846

---

R. Matsuda (✉)  
Ritsumeikan University, Kyoto, Japan  
e-mail: [ryoza.matsuda@gmail.com](mailto:ryoza.matsuda@gmail.com)

---

**Abstract**

The health-care system in Japan has been based on the Statutory Health Insurance System, consisting of more than 3,000 community-based and employment-based insurance plans with significant subsidies from the general budget. The system, supplemented by the Public Assistance Program, covers the entire residents for most medical and dental services. The national government decides its benefit basket and prices of covered services and pharmaceuticals after negotiations with providers and insurance organizations. Two-tiered local governments are involved in regulating the system, developing supplementary measures, and providing public health services. Patients are free to choose providers when they use health services. Physicians are not differentiated into general physicians and specialists: ambulatory care is provided both at clinics and at hospital outpatient departments. With different mixes of providers in different regions, the government has been developing regional regulations. The system works fairly well: access to healthcare is good though financial and geographical barriers have been occasionally reported, particularly in the era of increasing poverty. Mechanisms to monitor and regulate quality of care are becoming more important with increasing pressures on resources.

---

**Introduction**

The health-care system in Japan has been based on a combination of community-based and employment-based statutory health insurances with significant subsidies from the general budget. The system, which is called the statutory health insurance system (SHIS) here, has been governed at the national level, although local governments have been heavily involved in the system by operating community-based health insurances and implementing regulations on health-care providers.

The current system is an accumulation of layers that were molded in different periods and

contexts. In the nineteenth century, when Japan adapted to the changing world order and economy, the health care there drastically changed with the introduction of the Western medicine. The expansion of health-care services and the increased population of waged workers until the 1910s lead to the establishment of statutory health insurance for workers, an idea learned from Germany. Facing poverty and sickness particularly in rural areas having traditions of mutual assistance, the government made a legal framework to establish statutory community-based insurances in the 1930s. The framework enabled a municipality to establish first voluntary and then compulsory health insurance in the 1940s for the uninsured residents in its jurisdiction (Ikegami et al. 2011; Tataru and Okamoto 2009). The framework worked well until the national economy deteriorated during the WWII.

During the occupation by the Allied Nations since 1945–1952, new ideas and measures, including hospital planning, had been developed under the influences of the United States. The government expanded eligibility of the employment-based health insurance and imposed the obligation of establishing compulsory health insurance on all municipalities by 1961. The two types of insurance have been the main compositions for funding health care.

The health system in Japan has been principally universal since the implementation of the community-based compulsory statutory insurances across the country in 1961. Health-care services, pharmaceuticals and medical devices covered, and the coverage rates are almost the same across all statutory health insurances although some insurances provide additional coverage for, e.g., preventive medical examinations. The rules of payment for providers are also common across insurances.

Delivery of health care has been market-oriented under regulations of the government. Meanwhile, local governments have established their hospitals in their own initiatives, supported by subsidies from the national government. Private providers are principally supposed to behave as not-for-profits although some for-profit companies have their hospitals that began their operation

before the establishment of that principle. To administer the system, the national and local governments have developed complex regulations and incentives more than half century.

According to the structure given by the volume, this case study focuses on basic issues of the health system and mostly excludes descriptions on details on differences between statutory insurances and innovations in policy making. Also, its descriptions are limited to the period up to 2013. In translation of Japanese language, words are selected so that they are clear for international readers in references to previous articles (Ikegami et al. 2011): some English names of insurance plans, acts, and organization are different from the official translation.

---

## Organization and Governance

### Stewardship/Governance in Health System

Health-care policies are developed predominantly by the national government with involvement of concerned actors, including statutory health insurers, medical professions, and experts. Within the government, the ruling party, the Cabinet, and the Ministry of Health, Labour and Welfare (MHLW) as well as the Ministry of Finance and other ministries join health policy making. Statutory and non-statutory councils and committees of the government, involving concerned actors, usually discuss policy options to build consensus for enacting legislation and ministerial ordinances (Rodwin 2011).

The national government developed and enforced laws and regulations on health and long-term care. Coverage rates and policies are usually decided by bills that shall be passed by the National Diet. The Social Security Council within the MHLW develops national strategies on quality and safety, cost control, and payment reforms in health care. The Minister of Health, Labour and Welfare decides services covered and their prices, pharmaceuticals covered and the rule for deciding each price of each pharmaceutical, and other payments rules in the statutory health insurance

systems. The decision is made usually according to decisions of the Central Social Insurance Medical Council, which is a major arena for policy debates with representatives from insurers, providers, ministry officials, researchers and other experts. Meanwhile, technology assessment of pharmaceuticals and medical devices is conducted by the Pharmaceutical and Medical Devices Agency, a regulatory agency of the government.

Two-tiered local governments implement policies established by the national government as well as develop their own policies. Forty-seven prefectures, at the upper level, develop strategies for health-care development and health promotion, implement regulations on health facilities, monitor activities of providers, and collect data on health and health care. There are more than 1,700 municipalities at the lower level, each of which operates its community-based health insurance for residents that are not covered by other statutory insurers and the long-term care insurance. Prefectures and municipalities also implement regulations on clinics and home care providers and hospitals, respectively. Meanwhile, since local governments have omnipotent power to develop new policies unless they are against current law, they occasionally develop innovative policies for collaboration and supplemental measures to decrease cost-sharing of children in their jurisdictions.

### Dimensions of Coverage (Breadth, Scope, Depth)

The SHIS covers all residents in Japan except those with social assistance (or livelihood protection) and some exceptional cases. In practice, the insurance is operated by the following three types of compulsory insurance: employment-based health insurance (EHI), community-based health insurance (CHI), and health insurance for elderly (HIE).

The EHI covers employees and their dependents under age 75. It is operated by more than 1,400 society's established at large companies for their employees, by more than 75 mutual aid associations for public servants and other defined

groups, and by the National Health Insurance Association (NHIA) for those working at medium to small companies (Ikegami et al. 2011). Some groups of professionals (e.g., doctors in private practice) are covered by the purposely established associations by themselves.

Municipalities are in charge of administration both of the CHI and the HIE in different ways. They operate the CHI by themselves, while they delegate their responsibilities to 47 statutory insurers, also purposely established, at the prefecture level. A HIE insurer is governed by representatives of all municipalities in a prefecture.

The SHIS provides the same national benefit package, which covers hospital care, ambulatory care, mental health care, approved prescription drugs, home care, physiotherapy, and most dental care. Health checks, health education, and counseling are delivered by statutory insurers to those ages 40 and older. Social assistance provides similar coverage. Cancer screenings are delivered by municipalities outside the SHIS.

Co-payment rate is 30% in general but 20% for children under 3 years old and 10% for people ages 70 and over with lower incomes. To mitigate high financial burdens, catastrophic insurance covers most of co-payments over a monthly threshold which varies according to enrollee's age and income. Also, cost-sharing is reduced for those with low-income, disabilities, mental illness, and specified chronic conditions. A part of expenditure on health services and goods can be deducted from taxable income.

Providers are prohibited from charging extra fees in general, although they can make extra fees for some services specified by the MHLW, including amenity beds, "experimental treatments," the outpatient services of large multi-specialty hospitals, after-hours services, and hospitalizations of 180 days or more.

Catastrophic coverage stipulates a monthly out-of-pocket threshold which varies according to enrollee age and income (e.g., 80,100 yen for people under ages 75 with an average income); above this threshold, a 1% co-payment rate is applied. Alternatively, the threshold works as a ceiling for low-income people, who do not pay more than 35,400 yen a month in 2013.

Since 2000, the long-term care insurance (LTCI) covers all residents ages 40 and over. It is compulsory and covers both institutional care and home care. The co-payment rate is 10% in 2013.

## Typologies of Health System

The health system in Japan is principally a type of social insurance-based systems, but since the government has been involved in making decisions on some details of the system and more than a third of its funds comes from tax, the state involvement is far stronger than most social insurance-based systems in Western countries (Blank and Bureau 2010).

On the one hand, it has been partly based on statutory health insurances: the EHIs are funded by contributions both from employees and employers and the CHI by contributions of beneficiaries and subsidies from tax. On the other hand, the government has been holding strong power, particularly of deciding the payment system and levels. Although the system is operated by more than 3,000 statutory insurers, financing administration is highly concentrated with little discretion to each insurer except limited issues. Provision of health-care services is based on market mechanisms without gatekeeping mechanisms, where public and not-for-profit providers compete with each other as well as collaborate.

Lee et al. (2008) describe the system as a hybrid of a hybrid model between social health insurance and the national health insurance, where the financing administration of health systems is concentrated into a national entity, and private sectors are dominant in health-care provision.

Private voluntary health insurance, historically developed as a supplement to life insurance, appears to play a marginal role (Paris et al. 2010). Traditional plans usually pay a lump sum when insured persons are hospitalized over a defined period and/or diagnosed with cancer or any of other specified chronic diseases. In the last decades, however, varieties of complementary private insurance policies, sold separately from life insurance, have been increasing.

## Regulating and Planning; Actors and Responsibilities

The national, prefectural, and municipal governments regulate health care and conduct planning activities in various fields within the SHIS, structured by law. Statutory insurers are responsible for operating themselves within the framework and regulations stipulated by acts and ordinances.

The national government decides which services and pharmaceuticals are covered by the SHIS and the rules for paying them. It revises the rules every 2 years by building consensus between providers, insurance organizations, and experts in health policy. Once the rules are proclaimed, they are valid countrywide in the SHIS. The government also set requirements and quality standards for health-care facilities, most of which local governments enforce in their jurisdictions.

The national government shall and can develop its plans on health promotion and health care. Their objectives include promotion of healthy behavior and environment, higher utilization of personal preventive services, increase of efficiency in health-care delivery, and higher utilization of generic drugs (OECD 2009). It also makes guidelines for implementing regulations. It directly supervises the operation of the largest insurer, the NHIA. Seven regional bureaus of health and welfare supervise the operation of the insurance societies, local branches of the NHIA, and the CHI insurers.

Prefectural governments supervise and support the CHI insurers in its jurisdictions both in financial and technical terms. They shall develop their plans on health promotion and health care. They are usually supposed to consider policy and technical guidelines developed by the national government. Prefecture shall develop and publish health-care plans in its jurisdiction, which shall include assessment of needs, directions for strategic development, and descriptions of providers. The power and capacity of prefectures for implementing the plans have been limited to place the cap on hospital beds (Hashimoto et al. 2011). Prefectures

implement regulations on quality of hospital services and can develop their own policy measurement, including subsidies and regulations, with their budget. Prefectures shall have public health centers to which many of regulatory responsibilities concerned with health care and public health in their jurisdictions are usually delegated from the governors' office (Tatara and Okamoto 2009).

Municipalities are responsible for operating the CHI and the LTCI, delivering home and welfare services, and promoting health in the population. More autonomous large cities than usual municipalities shall establish public health centers.

The public can participate in every level of political decision-makings. In the last two decades, critical committees concerned with health care are more likely to have members who put patients' interest first.

---

## Financing

### Sources and Collection of Revenue

In 2010, 82.1% of total health expenditure was financed through the SHIS, meanwhile 14.4% by out-of-pocket (OECD 2013). The national and local government paid around a quarter and a ninth of national health spending, respectively. Contributions are collected by each insurer. Each CHI insurer decides its complex method of calculating premiums for households. Usually it is based on the number of CHI member in the household and the member's household income. Rates,, therefore, vary between municipalities. Each HIE insurer at a prefecture levies premium on per-capita and income basis.

The EHI insurers levy premiums on wages. Employers pay half of these premiums for their employees. Premium rates of the EHI societies vary between 3% and 10% of their income whereas rates of the NHIA, which differ between branches, are around 10%.

There are various types of direct and indirect tax both at the national and local levels, politically controlled. By law, the national

and local governments have obligations of paying funds, calculated with actual spending, to the SHIS.

### **Pooling of Funds and Resource Allocation**

Each insurer in the SHIS is expected to be financially healthy. Subsidies from the national and local governments are granted mainly to the CHI insurers and the HIE insurers and, to a lesser extent, the JHIA. There are cross-subsidies from the CHI and the HI insurers to the HIE insurers, calculated by factoring in the number of enrollees ages 65–74.

### **Purchasing Process and Paying for Health Services**

Providers are paid by a national payment rule, which combines various kinds of activity-based funding methods: fee-for-service payments, per-diem payments, and per-monthly payments for chronic outpatient care.

Providers send claims for the CHIs to the Central Federation of National Health Insurance and claims for the EHIs to the Health Insurance Claims Review & Reimbursement Services, a statutory body to manage claims in the SHIS.

### **Health Spending**

The total health expenditure (THE) on health as percentage of GDP is similar to the average of OECD countries. It continuously increased in the last decades. In 2010, 63.3%, 9.1%, 21.4%, 3.0%, and 1.6% of the THE were spent for services of curative and rehabilitative care, services of long-term nursing care, medical goods, prevention and public health services, and administration, respectively (OECD 2013). Hospitals, nursing and residential care facilities, and ambulatory care providers spent 47.1%, 3.8%, and 27.1% of the THE. More than 20% of the THE was spent for pharmaceuticals.

Increasing health-care demand, partly due to demographic changes and the introduction of new technologies, is considered as cost drivers in Japanese health-care system (Ikegami and Anderson 2012).

---

## **Physical and Human Resources**

### **Physical Resources**

Hospitals, clinics, intermediary facilities, long-term care facilities, and other facilities have developed. The number of hospitals and beds in them per population is high, compared to other OECD countries (Tatara and Okamoto 2009). Health facilities are owned and managed both publicly and privately. Private providers include health facilities owned by physicians as well as medical corporations, which are not-for-profit private legal entities, usually controlled by physicians, for health-care provision.

To decrease geographical variations, the national government increased the number of medical courses with a policy aiming that every prefecture has at least a university with medical faculties and educational hospitals in the 1960s and 1970s. Also, since 1956, the government has developed and implemented its Rural Healthcare Plan with subsidies to local governments since 1956.

Health facilities need to announce such specialties and/or subspecialties as “internal medicine,” “surgery,” “orthopedics,” and “circulatory medicine.” Which specialties and sub-specialties can be announced, the nomenclature is regulated by the government. It has not so far included “general practice,” “family practice,” nor “primary care.” The argument to make “general practice” or “primary care” recognizable has been discussed recently.

Health facilities can install licensed medical devices with its resources and, in some cases, with subsidies from the governments. Since there have been no regulations on their diffusions, magnetic resonance imaging (MRI) and computed tomography (CT) scanners spread widely (Anderson et al. 2005).

A hospital is defined by law as a health-care facility that provides medical care with at least 20 beds. Hospitals provide both inpatient care and outpatient care, particularly specialist outpatient care. Hospitals are either publicly or privately owned and/or operated. A fifth of hospitals is publicly owned and shares 30% of hospital beds. Small hospitals are common: half of hospitals have less than 150 beds. Psychiatric and long-term care account for around a fifth of hospital beds, respectively. The number of hospital beds is about four times larger than the OECD average (OECD 2009).

The government makes standards on health workforce, buildings, room spaces, instruments, and other necessities of hospitals. Hospitals need permission from prefectural governments when they increase number of beds. A prefecture has its plan on the number of hospital beds in its jurisdiction, according to which it can deny applications from hospitals.

In 2012, there are three types of hospitals: (usual) hospitals, psychiatric hospitals, and hospitals with infectious diseases. The standards vary between the types.

Prefectural governments designate 378 hospitals, making up approximately 5% of general hospitals, as “community hub hospitals,” which shall operate in close connections with community physicians working at clinics.

A clinic is defined by law as a health-care facility that provides medical or dental care without or with less than 20 beds for inpatient care. Most clinics are privately owned and operated. In 2010, only 4.9% of general clinics (clinics excluding dental clinics) are operated by public bodies (Health Statistics Office, Ministry of Health, Labour and Welfare 2011), although their presence is critical in rural areas (Matsumoto et al. 2010). Clinics have varieties of medical functions: most clinics provide primary care in reality, but some provide specialists care. For example, 0.4%, 4.0%, 0.2%, 3.7%, 3.3%, and 11.7% of clinics announce hematology, rheumatology, respiratory surgery, urology, proctological surgery, and dermatology, respectively, as one of their specialties (Health Statistics Office, Ministry of Health, Labour and Welfare 2012b).

## Intermediate Care Facilities

In the mid-1980s, a new type of health care facility was created to provide “intermediate care” between the hospital and the community (Ishizaki et al. 1998; Ikegami et al. 2003). Most services at the new facilities were covered by the SHIS since 2000. Then the coverage was transferred to the LTCL. Meanwhile, a new interdisciplinary post-acute rehabilitation unit has been incorporated in the SHSI (Miyai et al. 2011). Measures to develop community-based integrated care have been progressively implemented in the 2010s (Tsutsui 2014).

## The Health Workforce

### Physicians

Anyone without a license given by the government is prohibited by law to use the title, “physician” (*Ishi*). Physician license is given to those who pass the national medical board examination after graduating medical courses at universities and colleges. The capacity of those courses is strictly regulated by the government. Physicians that just pass the board examination must take mandatory 2-year trainings aimed at developing general clinical knowledge and skills (Teo 2007). After that, they freely practice in principle but usually continue to take trainings in various specialties (Teo 2007).

Most physicians working at hospitals are employed by the hospitals and receive salaries. The contract can be either individually or collectively through labor unions. Those salaries are usually not related to payments to hospitals from the SHIS.

Physicians working at clinics are usually owners of them and are responsible for their overall management in addition to clinical issues. So after paying costs for operating clinics, including human resources, buildings, and instruments, they can principally decide how to use it.

### Nurses and Other Co-medical Staff

There have been two qualifications in nursing: registered nurses and assistant nurses, who need licenses, awarded by the national or prefectural governments, to practice. Most nurses are

employed and get salaries from their employers. Approximately 60% of nurses work at hospitals, most of others at clinics. Some nurses operate home nursing service providers, in which case they earn money as owners of providers. Japanese Nursing Association has developed certification programs (Japanese Nursing Association 2011). Public health nurses, who are supposed to work in the field of public health, and midwives also need licenses to practice. One must take courses for the two professions with qualification as nurses. Qualifications for long-term care, including home helper and care worker at caring institutions, exist besides nursing qualifications.

Other qualified professionals in health care include physical therapists, occupational therapists, radiology technologists, and clinical medical technologists. For alternative medicines, licenses are needed to practice therapeutic massage, acupuncture, moxa cautery, and judo chiropractic treatment (Tatara and Okamoto 2009).

---

## **Provision of Services: Providers, Services, Access, and Quality**

### **Public Health**

Public health administration has been a part of general administrative structure of the governments and been separated from the SHIS. According to legislations by the national government, prefecture governments have a responsibility of public health and environmental health in their jurisdictions (Tatara and Okamoto 2011). Large cities, designated by ministerial ordinances, also have the same responsibility. Those prefectures and cities also have an obligation of establishing and operating public health centers and delegate most of their responsibilities and powers on public health to directors of those centers.

Municipalities delivered almost personal preventive services, including vaccination, health checks, and cancer screenings, until 2008. Since the 2008 Reform, statutory health insurers deliver health checks and behavioral modification programs, while municipalities continue to deliver other personal preventive services

(Matsuda 2008). The aim of current health checks, delivered by the insurers, is not checking general health but detecting possible metabolic syndromes so that insurers intervene to decrease health-care expenditures. The government established targets for uptake rates of health checks and introduces a financial incentive: insurers that fail to achieve the target have to pay more cross-subsidies to the HIE.

Regarding health promotion, the national government has the national plan and strategies for health promotion, “Health Japan 21,” and municipalities organize health activities for their residents using their local health centers.

### **Primary Care/Ambulatory Care**

Ambulatory care is provided by clinics and hospital out-patient departments. The number of ambulatory patients at medical clinics are 2.5 times than that of hospitals (Health Statistics Office, Ministry of Health, Labour and Welfare 2012a).

Since physicians in Japan are trained as specialists and primary care or family care medicine has not been established as a specialty in clinical medicine, it is difficult to distinguish primary care physicians, although it is easy to recognize such specialists as ophthalmologists, otolaryngologists, and dermatologists. It has been argued that “general practice” shall be a specialty and included to the nomenclature of specialties (Matsuda 2008).

There is no gate-keeping. Patients are free to choose either clinics or outpatient departments of hospitals when they need medical consultations. Meanwhile, highly specialized hospitals can make extra charges when patients visit them without referral from other providers. Physicians at clinics or outpatient departments deal with first-contact patients, although their performance might not be satisfactory by the standards of trained family physicians.

### **Specialized Ambulatory Care/Hospital Care**

Specialized ambulatory care is provided both at clinics and at outpatient departments of hospitals.



Patients can directly use the care without referral in principle, although they shall pay extra charges for the direct utilization. There has been a financial incentive to avoid direct utilization of patients of specialist care: hospitals with highly specialized care functions can charge extra fees to patients. Hospitals vary in scale from small hospitals with 20 beds to large with more than 1,000 beds.

Remuneration for specialist physicians depends on their status, i.e., whether they are employed physicians or owners or executives of health-care organizations, as described above.

The payment method to hospital inpatient care is based on their activities but has been gradually changing from payment on fee-for-service basis to payment on per-diem basis with case-mix modifications using the Diagnostic Procedure Combination (DPC), a case-mix classification system similar to the Diagnostic-Related Groups (Matsuda et al. 2008; Okamura et al. 2005). However, the payment system with the DPC is unusual because it includes both a DPC component and a fee-for-service component. The former is a per-diem payment that declines as the length of the hospital stay increases and covers services other than such specified services as surgical procedures and rehabilitation basic charge, which are covered by the fee-for-service component (OECD 2009). A specific coefficient to multiply DPC rate for a hospital is determined in consideration of different scales and functions of hospitals. Hospitals using the DPCs must submit detailed data on their services. In 2012, more than half of beds were paid with the DPC. The government uses the data to analyze hospital behaviors and impacts of financial incentives on them.

Integration or coordination of care has been emphasized in health and long-term care policy. Particular policies toward integration of care include development of disease-oriented clinical care pathways (Okamoto et al. 2011).

## Pharmaceuticals

Prescribed pharmaceuticals for outpatients and inpatients are covered by the SHIS. In principle, patients bring prescriptions of physician to

pharmacies in the community, which dispense prescribed pharmaceuticals to patients. Some pharmacies operate only for prescribed pharmaceuticals in the SHIS; the others sell OTC drugs and other goods in addition to provision of prescribed pharmaceuticals. There was a tradition that physicians dispense pharmaceuticals at their offices by themselves in Japan and the tradition still has remained: 41% of outpatient prescriptions were still dispensed by physicians in 2008 (OECD 2009).

Patients pay the same proportions of cost-sharing for prescribed drugs as described above. Pharmacists can replace prescribed brand-name drugs with generic drugs unless physicians explicitly prohibit it on their prescriptions. Generic drugs count for 47.9% in its quantity and 11.4% in monetary terms among prescribed drugs dispensed at pharmacies.

## Long-Term Care

With the mandatory Long-Term Care Insurance, established in 2000, person with disabilities can use monthly budgets, allocated according to their assessed needs, to purchase long-term care services. Long-term care services are classified largely into institutionalized care and community care. The government prohibits private companies to operate institutionalized care in the LTCI, although they can outside the LTCI. Most providers of institutionalized care, therefore, are not-for-profit organizations. Private for-profit companies can enter the community care market and account for around half of all community care providers (Oliveras-Tirado and Tamiya 2013).

## Mental Health Care

Psychiatric hospitals, psychiatric departments of general hospitals, and psychiatric clinics provide mental health care covered by the SHIS. In addition to those providers, prefectures have mental health centers, which are mostly funded with tax, to support providers with expertise and develop collaboration between concerned organizations.

Community mental health care has been developed.

## Dental Care

Dental care for children as well as adults is covered by the SHIS. Some common services, including orthodontics and expensive artificial teeth, are excluded from the SHIS coverage.

To become a dentist, one shall graduate from a dental school and pass the national board examination. Most dentists own and operate their clinics, who are paid on the fee-for-service basis, and employ dental hygienists and technicians who work with dentist.

## Complementary and Alternative Medicines

The government issues licenses of massage therapists, acupuncturists, moxa cauterists, and judo chiropractitioners for providing care. The licensed practitioners can provide defined services in the SHIS provided that physicians order them.

## Assessment

One difficulty for anyone trying to assess the Japanese health-care system is that the fragmented system and lack of system-level robust data make it difficult to assess it quantitatively. The long life expectancy in Japan suggests that the system works at least fairly well even if strong health consciousness and prevalent healthy behaviors are taken into consideration (Ikeda et al. 2011). Looking parts of the system, however, inefficiency in delivering health care and imbalances between regions have been pointed. Health expenditure has been fairly controlled, but its projected increase in the near future jeopardizes the sustainability of the system (OECD 2010).

Access to health care has been good since patients can choose any providers principally. However, in some rural areas, patients have difficulties to find physicians, particularly such specialists as

obstetricians and pediatricians. Although there have been much differences in health-care resources between prefectures, reasons of the differences and whether they are inequitable or not have not firmly assessed. Furthermore, in the era of increasing poverty, fair and good access to quality health services have encountered new challenges. Those challenges include delinquency in paying premiums to the CHI and cost-related access problems with the current co-insurance rates, particularly in ambulatory care (Matsuda 2016; Murata 2010; OECD 2009).

Quality of care is another area lacking systematic evidences. However, new institutions for hospital certification and policy incentives have been developed since 2000. More and more hospitals publish their clinical indicators, which are supported by the government. With increasing financial pressures on health-care resources, mechanisms to monitor and regulate quality of care are becoming more important (Hashimoto et al. 2011).

## References

- Anderson GF, Hussey PS, Frogner BK, Waters HR. Health spending in the United States and the rest of the industrialized world. *Health Aff.* 2005;24(4):903–14. <https://doi.org/10.1377/hlthaff.24.4.903>.
- Blank RH, Bureau V. *Comparative health policy*. Basingstoke: Palgrave Macmillan; 2010.
- Hashimoto H, Ikegami N, Shibuya K, et al. Cost containment and quality of care in Japan: is there a trade-off? *Lancet.* 2011;378(9797):1174–82. [https://doi.org/10.1016/s0140-6736\(11\)60987-2](https://doi.org/10.1016/s0140-6736(11)60987-2).
- Health Statistics Office, Ministry of Health, Labour and Welfare. Summary of 2010 static/dynamic surveys of medical institutions and hospital report. Tokyo: Ministry of Health, Labour and Welfare; 2011.
- Health Statistics Office, Ministry of Health, Labour and Welfare. Summary of 2011 patient survey. Tokyo: Ministry of Health, Labour and Welfare; 2012a.
- Health Statistics Office, Ministry of Health, Labour and Welfare. Summary of 2011 static/dynamic surveys of medical institutions and hospital report (in Japanese). Tokyo: Ministry of Health, Labour and Welfare; 2012b.
- Ikeda N, Saito E, Kondo N, et al. What has made the population of Japan healthy? *Lancet.* 2011;378(9796):S32#1094–105.
- Ikegami N, Anderson GF. In Japan, all-payer rate setting under tight government control has proved to be an effective approach to containing costs. *Health Aff.* 2012;31(5):1049–56. <https://doi.org/10.1377/hlthaff.2011.1037>.

- Ikegami N, Yamauchi K, Yamada Y. The long term care insurance law in Japan: impact on institutional care facilities. *International Journal of Geriatric Psychiatry*. 2003;18(3):217–221.
- Ikegami N, Yoo B-K, Hashimoto H, et al. Japanese universal health coverage: evolution, achievements, and challenges. *Lancet*. 2011;378(9796):1106–15.
- Ishizaki T, Kobayashi Y, Tamiya N. The role of geriatric intermediate care facilities in long-term care for the elderly in Japan. *Health Policy*. 1998;43(2):141–151.
- Japanese Nursing Association. *Nursing in Japan*. Tokyo: Japanese Nursing Association; 2011. Available at: <http://www.nurse.or.jp/jna/english/pdf/nursing-in-japan2011.pdf>
- Lee S-Y, Chun C-B, Lee Y-G, Seo NK. The National Health Insurance system as one type of new typology: the case of South Korea and Taiwan. *Health Policy*. 2008;85(1):105–13.
- Matsuda R. Arguments for instituting “general physicians”. *Health Policy Monitor*, April 2008. 2008. Available at: [http://www.hpm.org/en/Surveys/Ritsumeikan\\_University\\_-\\_Japan/11/Arguments\\_for\\_Instituting\\_General\\_Physicians\\_.html](http://www.hpm.org/en/Surveys/Ritsumeikan_University_-_Japan/11/Arguments_for_Instituting_General_Physicians_.html)
- Matsuda R. Public/Private Health Care Delivery in Japan: and Some Gaps in Universal Coverage. *Global Social Welfare*. 2016;3:201. <https://doi.org/10.1007/s40609-016-0073-1>
- Matsuda S, Ishikawa KB, Kuwabara K, Fujimori K, Fushimi K, Hashimoto H. Development and use of the Japanese case-mix system. *Eurohealth*. 2008;32#14(3):25–30.
- Matsumoto M, Inoue K, Kajii E, Takeuchi K. Retention of physicians in rural Japan: concerted efforts of the government, prefectures, municipalities and medical schools. *Rural Remote Health*. 2010;10(2):1432.
- Ministry of Health, Labour and Welfare. *Pharmaceutical expenditures at dispensing pharmacies, FY2013 (in Japanese)*. Tokyo: Ministry of Health, Labour and Welfare; 2014.
- Miyai I, Sonoda S, Nagai S, Takayama Y, Inoue Y, Kakehi A, Kurihara M, Ishikawa M. Results of New Policies for Inpatient Rehabilitation Coverage in Japan. *Neurorehabilitation and Neural Repair*. 2011;25(6):540–547.
- Murata C, Yamada T, Chen CC, Ojima T, Hirai H, Kondo K. Barriers to Health Care among the Elderly in Japan. *International Journal of Environmental Research and Public Health*. 2010;7(4):1330–13413
- OECD. Health-care reform in Japan: controlling costs, improving quality and ensuring equity. In: OECD, editor. *OECD economic surveys: Japan 2009*. OECD Publishing; 2009. [https://doi.org/10.1787/eco\\_survey\\$32#s-jpn-2009-6-en](https://doi.org/10.1787/eco_survey$32#s-jpn-2009-6-en).
- OECD. Value for money in health spending. 2010. <https://doi.org/10.1787/9789264088818-en>.
- OECD. Health data 2013 [database on the Internet]. 2013.
- Okamoto E, Miyamoto M, Hara K, et al. Integrated care through disease-oriented clinical care pathways: experience from Japan’s regional health planning initiatives. *Int J Integr Care*. 2011. Available at: <http://www.ijic.org>. URN:NBN:NL:UI:10-1-101572.
- Okamura S, Kobayashi R, Sakamaki T. Case-mix payment in Japanese medical care. *Health Policy*. 2005;74(3):282–6.
- Olivares-Tirado P, Tamiya N. Trends and Factors in Japan’s Long-Term Care Insurance System: Japan’s 10-year Experience. Dordrecht: Springer; 2013.
- Paris V, Devaux M, Wei L. Health systems institutional characteristics. OECD health working papers, No. 50. Paris: OECD; 2010.
- Rodwin MA. Conflicts of interest and the future of medicine: the United States, France, and Japan. New York: Oxford University Press; 2011.
- Tatara K, Okamoto E. Japan. Health system review. *Health Syst Transit*. 2009;11(5):1.
- Tatara K, Okamoto A. *Public health of Japan 2011*. Tokyo: Japan Public Health Association; 2011.
- Teo A. The current state of medical education in Japan: a system under reform. *Med Educ*. 2007;41(3):302–8. <https://doi.org/10.1111/j.1365-2929.2007.02691.x>.
- Tsutsui T. Implementation process and challenges for the community-based integrated care system in Japan. *International Journal of Integrated Care*. 2014;14(1).



Julio Frenk and Octavio Gómez-Dantés

## Contents

<b>Health Conditions</b> .....	850
<b>History of the Mexican Health Care System</b> .....	851
<b>Organization and Governance</b> .....	851
Organization .....	851
Planning and Regulation .....	852
Health Information Systems and Technology .....	852
Role of Patients .....	853
<b>Financing</b> .....	853
Coverage and Benefits .....	853
Sources of Revenue, Collection, and Pooling .....	854
Health Expenditure .....	854
<b>Physical and Human Resources</b> .....	855
Pharmaceuticals .....	855
<b>Delivery of Personal and Public Health Services</b> .....	855
Quality of Care .....	856
<b>Recent Reforms</b> .....	857
<b>Assessment</b> .....	857
<b>References</b> .....	858

---

President of the University of Miami and former Minister of Health of Mexico (2000–2006)  
Senior researcher, Center for Health Systems Research, National Institute of Public Health, Mexico

J. Frenk (✉)  
University of Miami, Coral Gables, FL, USA  
e-mail: [president@miami.edu](mailto:president@miami.edu)

O. Gómez-Dantés  
National Institute of Public Health, Cuernavaca, MOR, Mexico  
e-mail: [ocogomez@yahoo.com](mailto:ocogomez@yahoo.com)

---

## Abstract

This chapter discusses the Mexican health system. We first describe the general characteristics of Mexico and the health conditions of the Mexican population, with emphasis in non-communicable diseases, which are now the main cause of death and disability. The following section is devoted to the description of the basic structure of the system: its history; its main institutions; the population coverage; the health benefits of those affiliated to the

different health institutions; its financial sources; the availability of physical, material, and human resources for health; the delivery of personal and public health services; the stewardship functions displayed by the Ministry of Health; and other actors. This part also discusses the role of citizens in the monitorization and evaluation of the health system, as well as the levels of satisfaction with the rendered health services. In part three, the most recent innovations and its impact on the performance of the health system are discussed. Salient among them are the System of Social Protection in Health and the Popular Health Insurance. The chapter concludes with a discussion of the most recent health initiatives and reforms, and a brief analysis of the short- and middle-term challenges faced by the Mexican health system.

Mexico is the largest Spanish speaking country in the world. It covers 1.9 million km<sup>2</sup> of land in North America ([Central Intelligence Agency](#)). It borders to the north with the USA, and with Guatemala and Belize to the south.

Mexico is an upper middle income country with a GDP of US\$<sub>ppp</sub> 1.788 trillion (2012) and a per capita GDP of US\$<sub>ppp</sub> 15,100.1. Its human development index is 0.775 (2012), above the world average of 0.694 and ranking 61 out of 187 countries ([UNDP](#)). Inequality, as measured by the Gini index, is 47.2, higher than all other high human development countries except for Brazil ([The World Bank](#)). Its principal source of income is services (61.8%), with industry running second (34.2%) and agriculture representing a small and waning portion (4.1%) ([Central Intelligence Agency](#)). Its annual economic growth rate during the period 1990–2010 was 2.8% ([The World Bank](#)).

Mexico has a population of 116.2 million (2013 est.) that is witnessing: ([Central Intelligence Agency](#); Partida 1999)

- A decline in general mortality explained mostly by a reduction in infant mortality from 79 per 1000 live births in 1970 to 16.2 in 2013 (2013 est.)

- An increase in life expectancy at birth from 49.6 years in 1950 to 79.8 years in women and 74.0 in men in 2013 (2013 est.)
- A reduction in fertility from 6.8 children per women of reproductive age in 1970 to 2.2 in 2013 (2013 est.)

The rapid decline in fertility is driving an aging process which implies an increasing proportion of older adults in the population structure. Children under 5 will represent less than 10% of the total population in 2050 while older adults will concentrate over 20% of the total population ([Ham-Chande 2012](#)).

Mexico is also going through an accelerated process of urbanization. Eight out of every 10 Mexicans now live in urban areas ([Central Intelligence Agency](#)). This is associated to a parallel process of rural population dispersion which increases the problems of access to health care of a population with major health needs ([Reynabernal and Hernández-Esquivel 2006](#)).

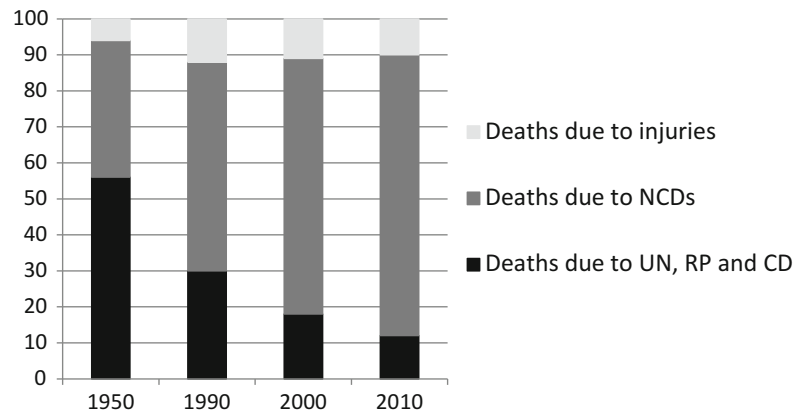
---

## Health Conditions

The increase in life expectancy and a growing exposure to unhealthy life styles in urban dwellings are modifying the main causes of disease, disability, and death. Mexico is going through a health transition characterized by an increasing predominance of noncommunicable diseases (NCD) and injuries. In 1950 around 50% of all deaths in the country were due to common infections, reproductive events, and diseases related to undernutrition (Fig. 1) ([Secretaría de Salud 2001](#)). Today, these ailments concentrate less than 12% of total deaths, while NCDs and injuries are responsible for almost 90% of national mortality ([World Health Organization 2012](#)).

The contribution to mortality of the different age groups is also changing. In 1950, half of total deaths were concentrated in children under 5 and only 15% were concentrated in persons 65 years of age and older ([Secretaría de Salud 2007](#)). Nowadays, more than 50% of deaths are concentrated in older adults and less than 10% in children under 5 ([Zúñiga and García 2008](#)).

**Fig. 1** Evolution of the distribution of mortality by type of disease, México 1950–2010. NCD noncommunicable diseases, UN, RAP, and CD undernutrition, reproductive problems, and communicable diseases (Source: Ministry of Health, Mexico Secretaría de Salud (2001))



## History of the Mexican Health Care System

The origins of the modern Mexican health system date back to 1943, when the Ministry of Health (MoH) and the Mexican Institute for Social Security (IMSS) were created. IMSS would serve the industrial work force, while the MoH was assigned the responsibility of caring for the urban and rural poor (Frenk et al. 2003). In 1960, a social security institution for civil servants was created, the Institute for Social Security and Services for Government Employees (ISSSTE).

In order to extend access and improve the efficiency and quality of care, a health care reform was launched in 1983: a constitutional amendment establishing the right to the protection of health was introduced; a new health law was published; and health services for the uninsured population were decentralized to state governments (Soberón 1987). The force guiding this program was primary health care. However, universal access to comprehensive services would not be reached until the initial years of the new millennium.

In the 1990s several national health accounts studies revealed that more than half of total health expenditure in Mexico was out-of-pocket. This was due to the fact that half of the population lacked health insurance. This exposed Mexican households to financial crisis. Not surprisingly, Mexico performed poorly on the comparative analysis of fair financing developed by the WHO

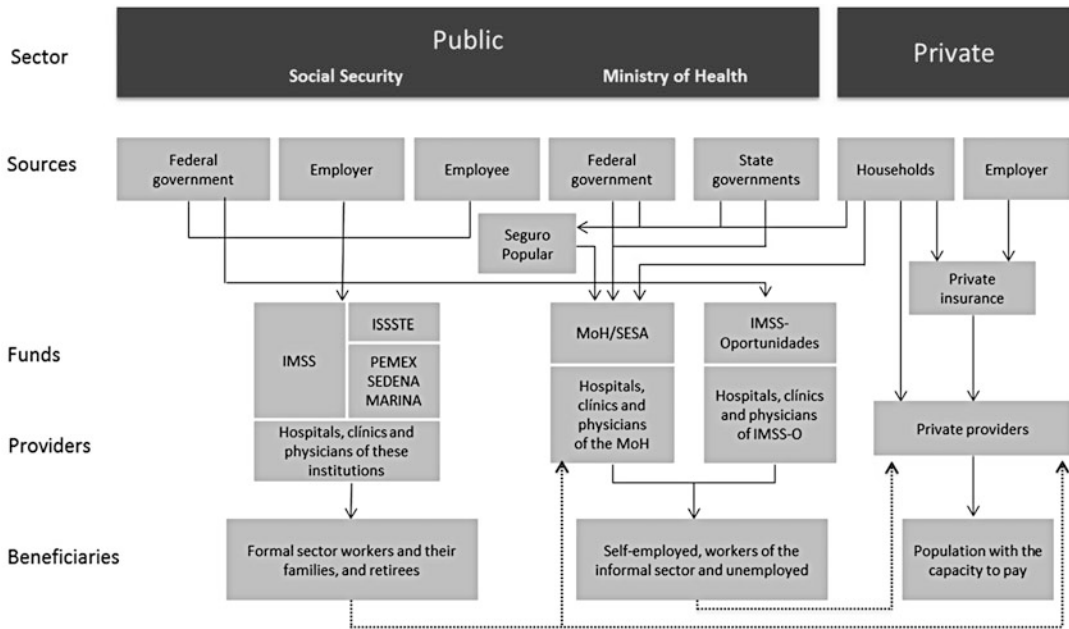
as part of the *World Health Report 2000* (World Health Organization 2000).

These results encouraged the development of further analysis that showed that catastrophic health expenditures were concentrated among the poor and uninsured. The products of these studies generated the advocacy tools to promote a legislative reform that established the System for Social Protection in Health (SSPH) in 2004 (Frenk et al. 2004). This system has mobilized public resources by a full percentage point of GDP over a period of 8 years to provide health insurance, through a public scheme called *Seguro Popular*, to all those ineligible for social security (those who are self-employed, unemployed, or altogether out of the labor force).

## Organization and Governance

### Organization

The Mexican health system includes a public and private sector. The public sector comprises the social security institutions [IMSS, ISSSTE, and the social security institutions for oil workers (PEMEX) and the armed forces (SEDENA and SEMAR)], *Seguro Popular*, and the institutions offering services to the uninsured population, including the Ministry of Health (MoH), the State Health Services (SESA), and the *IMSS-Oportunidades* Program (IMSS-O) (Fig. 2). These institutions run their own health facilities with their



**Fig. 2** The Mexican health system has a public and private sector providing services to overlapping population groups

own staff, except for *Seguro Popular*, which buys services for its affiliates from the MoH, SESA, and IMSS-O. The private sector includes facilities and providers offering services mostly on a for-profit basis financed either through insurance premiums or out-of-pocket payments.

**Planning and Regulation**

The MoH is in charge of most stewardship functions, including strategic planning, policy design, intra- and inter-sectoral coordination, regulation of personal health services, sanitary regulation, and evaluation of policies and programs. The regulation of personal health services includes the accreditation of medical and nursing schools, the certification of health professionals, and the accreditation of health facilities. These activities are developed in coordination with several professional bodies and NGOs, including the National Academy of Medicine and the National Association of Medical Schools and Faculties. The protection of health service users is in charge of the National Commission for Medical Arbitrage

(*CONAMED*) ([Comisión Nacional de Arbitraje Médico](#)).

Regulation is the responsibility of the Federal Commission for Health Risk Protection (*COFEPRIS*), charged with assuring food safety, defining environmental standards, promoting occupational health and safety, regulating the pharmaceutical industry, and controlling hazardous substances like alcohol and tobacco ([Comisión Federal de Protección contra Riesgos Sanitarios](#)).

The MoH also counts with an evaluation unit which evaluates the main policies and programs and publishes an annual report on the performance of the Mexican health system and its various components ([Dirección General de Evaluación del Desempeño, Secretaría de Salud, México](#)).

**Health Information Systems and Technology**

Health information is the responsibility of the General Directorate for Health Information based at the MoH ([Dirección General de](#)

**Table 1** Health care coverage, Mexico 2002 and 2010

Type of population	2000		2010	
	Number (million)	%	Number (million)	%
Population with social security	38.7	37.4	50.7	45.1
Population with private insurance <sup>a</sup>	2.5	2.4	2.8	2.5
Population enrolled in <i>Seguro Popular</i>	–	–	43.5	38.7
Population with health insurance	41.2	39.8	97.0	86.3
Uninsured population	62.2	60.2	15.3	16.6
Total population	103.4	100	112.3	100

Source: Refs. (Crónica; Comisión Nacional de Protección Social en Salud; Comisión Nacional de Protección Social en Salud 2012).

<sup>a</sup>Around half of the population with private health insurance is also covered by public insurance. In this figure we consider those with private health insurance only

**Información en Salud, Secretaría de Salud, México**). In collaboration with other public institutions, this office created the National Health Information System (SINAIS), which generates information on births, deaths, cases of disease, health infrastructure, health services, and financial and human resources (**Sistema Nacional de Información en Salud, México**). SINAIS counts with several subsystems including the Epidemiological Surveillance System, the Automatized Hospital Discharge System, and the National and State Health Accounts System.

The MoH has an area for the evaluation of medical technology, the National Center for Health Technology Excellence, whose main purpose is to produce and disseminate information on the appropriate selection, incorporation, and use of medical technologies based on evidence of their safety, effectiveness, and efficiency (**National Center for Health Technology Excellence**).

### Role of Patients

Patients in Mexico started playing a role in the operation of the Mexican health system until very recently through the “citizen endorsements groups,” created in 2001 as part of a quality program, the “National Crusade for Quality in Health Care.” The purpose of these groups is to train community volunteers to assess the responsiveness of health care facilities (Ruelas 2006). In 2006, there were 1764 active citizen groups that had endorsed over 1100 health units.

Besides these groups, citizens have traditionally played a limited role in the design and operation of health services, programs, and policies. The main exceptions are the HIV/AIDS and women’s health advocacy groups.

## Financing

### Coverage and Benefits

The Mexican health system is segmented along three broad categories of beneficiaries: (i) salaried workers and retired population, along with their families; (ii) self-employed workers and unemployed population, along with their families; and (iii) the population with the ability to pay.

As mentioned above, salaried workers are the beneficiaries of social security institutions, which in 2010 covered 50.7 million people (Table 1; Crónica). IMSS covered 80% of this population, and the rest was covered by ISSSTE and the social security institutions for oil workers and the armed forces.

The second category (self-employed and unemployed, and their families) was covered until 2003 by services of the MoH, SESA, and IMSS-O. The recently created *Seguro Popular* was covering 43.5 million individuals in this category by 2010 (**Comisión Nacional de Protección Social en Salud; Comisión Nacional de Protección Social en Salud**). By the end of 2011, affiliation to *Seguro Popular* reached 52 million. This means that Mexico is on



track to reach universal health coverage in the near future.

Finally, the third category includes the users of private health services, mostly upper and middle class individuals. However, the poor and those affiliated to social security institutions also use them on a regular basis. According to the National Health and Nutrition Survey 2012 (ENSANUT 2012), over 30% of the insured population regularly use private health services, mostly ambulatory care, for which they usually pay out-of-pocket (Instituto Nacional de Salud Pública 2013). The penetration of private insurance is low. Only six million people in Mexico are covered by private health insurance, half of which also are covered by public insurance (CNNExpansión).

Those affiliated to social security institutions have access to a broad, but not explicitly defined, package of health services that includes ambulatory and hospital care, including high specialty care. Coverage includes drugs as well. Those affiliated to *Seguro Popular* have access to a comprehensive and explicit package of 270 essential interventions and the respective drugs. They also have access to a package of over 60 high-cost interventions for the treatment of acute neonatal conditions, cancer in children, cervical and breast cancer, and HIV/AIDS, among other diseases. Finally, the uninsured population has access to a limited package of benefits that vary considerably depending on the type of population (urban or rural).

### Sources of Revenue, Collection, and Pooling

As shown in Fig. 2, social security institutions are financed with contributions from the government, the employer (which in the case of ISSSTE, PEMEX, SEDENA, and SEMAR is also the government in its role as employer), and the employee. The MoH and the SESA are financed mostly with federal and state government resources coming from general taxation. IMSS-O, which is directed to the rural poor of 17 states, is financed with federal resources but operated by

IMSS. Finally, *Seguro Popular* is financed with federal and state government contributions and family contributions, with total exemption for those families in the bottom 40% of the income distribution.

Private services are financed mostly out-of-pocket. A very small portion of private health expenditure comes from private insurance premiums.

### Health Expenditure

Total health expenditure as % GDP in Mexico in 2010 was 6.3%, well below the OECD average (9.3%) and below the Latin American average (6.8%), but up from 5.1% in 2000 (World Health Organization; Organization for Economic Cooperation and Development; World Health Organization). Health expenditure per capita in that same year was US\$<sub>ppp</sub> 603, up from US\$<sub>ppp</sub> 328 in 2000.

Mexico's public expenditure on health as a percentage of total health expenditure in 2010 was 49%, up from 46.6% in 2000 but still the third lowest of OECD countries (World Health Organization; Organization for Economic Cooperation and Development).

Private expenditure concentrates 51% of total health expenditure in Mexico, a much larger portion than the average OECD country (17%) and a larger portion than Argentina (35.6%), Colombia (25.4%), and Uruguay (34.7%) but lower than Brazil (53.0%) (World Health Organization; Organization for Economic Cooperation and Development; World Health Organization).

Ninety two percent of private health expenditure is out-of-pocket (World Health Organization). The remaining 8% corresponds to private insurance premiums (World Health Organization). In Argentina, Brazil, Colombia, and Uruguay, out-of-pocket expenditure concentrates 60%, 57.8%, 67.7%, and 39.6% of total private health expenditure, respectively (World Health Organization). This means that Mexico has the highest level of out-of-pocket expenditure of middle-income countries in Latin America. This exposes households to catastrophic financial events. In 2000, an estimated

three million Mexican families suffered catastrophic or impoverishing health expenditures (Frenk et al. 2006). However, several studies showed that by 2006 this figure began to decline due to the implementation both of several programs to combat poverty and *Seguro Popular* (Knaul et al. 2006, 2011).

---

## Physical and Human Resources

Excluding medical offices of the private sector, the Mexican health system has about 27,000 health units, 3976 of which are hospitals, for a rate of 3.5 hospitals per 100,000 population (Dirección General de Evaluación del Desempeño, Secretaría de Salud, México). Of the total number of hospitals, 1386 (33.6%) are public and 2590 are private (66.4%). Of the total number of public hospitals, 2147 (54%) belong to SESA and MoH and 1829 (44%) to social security institutions.

In 2010 the three main public institutions (MoH/SESA, IMSS, and ISSSTE) had 74,064 hospital beds and 2900 operating rooms for a rate of 6.5 beds per 10,000 population and 2.5 operating rooms per 100,000 population (Dirección General de Evaluación del Desempeño, Secretaría de Salud, México).

Private hospitals count with 34,000 hospital beds. Most of them are general hospitals and are concentrated in the largest cities of the country. Most of them have 20 beds or less. Some of these units, in fact, can hardly be considered hospitals at all since they have no laboratories, no radiology and imaging services, and no blood banks.

The Mexican health system also has over 20,000 public ambulatory units, most of which belong to SESA (Dirección General de Evaluación del Desempeño, Secretaría de Salud, México 2000).

Regarding high specialty medical equipment and procedures, Mexico has a rate of 3.9 computed tomography units (CTU) and 1.3 radiotherapy units (RTU) per million population, the lowest and second lowest figures for OECD countries, respectively, which on average have 8.2 CTU and 6.9 RT per million population (World

Health Organization 2013; OECD. OECD Health Data 2013).

Regarding human resources, there are 1.96 doctors per 1000 population, below the OECD average (3.0) and other Latin American countries, such as Argentina (3.0) and Uruguay (3.7) (World Health Organization 2013). The scarcity of these resources is particularly acute when it comes to human resources for mental health: in Mexico there are only 0.02 psychiatrists per 1000 population (World Health Organization 2013). The availability of nurses, 2.7 per 1000 population, is also below the OECD average of 8.6 (OECDiLibrary).

## Pharmaceuticals

The Mexican market of pharmaceutical products is the 12th largest market in the world and the second largest in Latin America, just below Brazil (Massachusetts Office of International Trade and Investment; Chhabara). Mexico spends 27% of its total expenditure on health in pharmaceuticals, the third highest figure for OECD countries (OECD). About 80% of total expenditure in pharmaceuticals is concentrated in generic drugs, a market that has shown important growth rates in the past decade.

Around 80% of total expenditure in pharmaceuticals is private and 90% is out-of-pocket, one of the highest figures in the world (Moise and Docteur 2008). The public sector concentrates 20% of the national expenditure in pharmaceuticals and 35% of its volume. This difference is due to the fact that most of the drugs purchased by public institutions are generics, which are considerably cheaper than patented drugs.

---

## Delivery of Personal and Public Health Services

Health care services in public institutions are provided at social security, MoH, SESA, and IMSS-O facilities. Those in the formal, private sector of the economy receive health services at IMSS clinics and hospitals. Those in the formal, public sector of the economy receive services at ISSSTE,

PEMEX, SEDENA or SEMAR facilities. Those affiliated to *Seguro Popular* receive health care at the MoH, SESA, and IMSS-O facilities. The latter institutions also provide services to the uninsured. All these public providers run their health care network with their own personnel.

Private providers offer services through a very heterogeneous networks that includes large hospitals offering high-quality but expensive care in a few metropolitan areas and a large amount of small hospital/clinics (general hospitals providing mostly obstetric care) offering services of poor quality.

Social security institutions and *Seguro Popular* are allowed to hire private providers to supply services for their affiliates when demand surpasses capacity or when there is a lack of personnel, equipment, or other inputs to provide any covered service. In 2012 IMSS contracted-out dialysis and hemodialysis services for almost US\$ 340 million ([Instituto Mexicano del Seguro Social](#)).

Furthermore, as mentioned above, due to problems of access and quality of public services, many individuals affiliated both to social security institutions and *Seguro Popular* make regular use of private out-patient services paying out-of-pocket. ENSANUT 2012 indicates that 39% of total out-patient services are offered by private providers.

The use of private hospital services by those affiliated to social security or *Seguro Popular* is less common for two reasons: the quality of services offered by public providers tends to increase with the level of care, and middle-class and poor households seldom have the resources needed to make use of private hospital facilities. ENSANUT 2012 indicates that only 17% of total hospitalizations in Mexico occur in private facilities, down from 23.9% in 2000 and 20.9% in 2006 ([Instituto Nacional de Salud Pública 2013](#)). This trend matches the upward trend in hospitalizations observed in units of the MoH which increased from 25.9% of total hospitalizations in Mexico in 2000 to 38.3% in 2012, a clear effect of the implementation of *Seguro Popular*.

Public health services are provided by MoH to all the population, regardless of its affiliation

status to any particular health institution. These services include health promotion, risk control, and disease prevention activities, including vaccination, and epidemiological surveillance.

## Quality of Care

Quality has been a concern of the Mexican health system for a long time. A quality assessment conducted at the end of the past century in more than 1900 public health centers and 214 general public hospitals documented problems with waiting times, drug supply, medical equipment, and use of medical records. Historically, public institutions have operated as monopolies with no choice, poor responsiveness to consumer needs, and lack of concern for quality. Furthermore, health care facilities were not subject to a formal accreditation process.

In the past decade two national quality programs were implemented: the National Crusade for Quality in Health Care and *Sicalidad*. These initiatives were designed to improve standards of personnel and technical quality in service delivery and enhance the capacity of citizens to demand accountability.

A central component of these initiatives was the strengthening of the certification process for public and private health units, which is now coordinated by the National Health Council (NHC), an institution created in 1917 as the highest policymaking body in the sector. This process was reinforced by a disposition incorporated to the General Health Law in 2003 requiring the accreditation of all units providing services to *Seguro Popular*.

Initiatives to monitor and improve the availability of drugs in public institutions were also implemented in the early 2000. External measurements have shown major improvements in drug availability in all public institutions, especially in ambulatory facilities.

A national system of indicators, *Indica*, was also put in place to monitor quality of care by state and institution. This monitoring system includes indicators for waiting times for ambulatory and emergency care, waiting times for elective

interventions, and distribution and dispensing of pharmaceuticals, among other indicators.

Several external surveys have measured the levels of satisfaction with health care in Mexico. Regarding overall satisfaction with hospital care, ENSANUT 2012 indicates that 80.6% of health service users consider health care services either “good” or “very good” (Instituto Nacional de Salud Pública 2013). Social security institutions providing services to oil workers and the armed forces show the highest satisfaction levels (97%), followed by private facilities (92%).

---

## Recent Reforms

The creation of the SSPH in 2004 allowed for the expansion of health care coverage for the non-salaried population while also improving the quality of the available services and the protection against health risks. This system was able to reorganize and increase public funding by a full percentage point of GDP over 8 years in order to provide universal health insurance. The vehicle for achieving this aim was *Seguro Popular*. By December of 2012, 52 million people were enrolled in it (Comisión Nacional de Protección Social en Salud). If we add to these figures those affiliated to social security institutions and those with private health insurance, we can reasonably state that Mexico is on track to achieve universal health coverage.

The reform also contemplated quality oriented initiatives including the organization of training programs on quality improvement tools for health professionals; the monitorization of quality indicators through the regular information systems and external satisfaction and responsiveness surveys; and the establishment of a compulsory accreditation for all units willing to provide services to those affiliated to *Seguro Popular*.

Regarding public health, the Mexican reform established a protected fund for community health services targeting health promotion and disease prevention interventions, which allowed, among other things, for a major expansion of the basic immunization scheme; additional public

health investments to enhance human security through epidemiological surveillance and improved preparedness to respond to emergencies, natural disasters, and the threats related to globalization, including potential pandemics; and a major reorganization leading to the establishment of a new public health agency (COFEPRIS) charged with protection against health risks.

Another crucial component of the health reform was an external evaluation that used a quasi-experimental design. This community trial, implemented in 2005–2006 in over 38,000 households taking advantage of the phase-in implementation of the intervention, showed that *Seguro Popular* was reducing out-of-pocket expenditures and providing protection against catastrophic health expenditures especially to the poorest households (King et al. 2009). Additional studies also showed improvements in health service utilization and effective coverage both of preventive and curative interventions, including interventions for the main causes of disease, such as diabetes and breast cancer (Lozano et al. 2006; Gakidou et al. 2006).

---

## Assessment

As shown in this chapter, Mexico has made progress in the three main objectives of health systems: improving health conditions, enhancing responsiveness to the legitimate expectations of the population, and providing financial protection (Murray and Frenk 2000). However, the country is facing emerging challenges.

Efforts to control pretransition ailments have yielded significant progress. However, as increased immunization coverage expanded and deaths due to diarrhea and acute respiratory infections declined, NCDs began to exercise an increasing pressure on the health of the population and the health system. Salient among these challenges is a critical need for additional public funding to extend access to costly interventions for NCDs, such as cardiovascular diseases, cancer, diabetes, and its complications, and mental health problems.

Another challenge facing the Mexican health system is to achieve a right balance between additional investments in health promotion, risk control, and disease prevention, urgently needed to address the health risks related to NCDs, on the one hand, and investments in personal curative health services on the other.

Finally, further progress in quality of health care is still expected. The most critical areas are technical quality of care; availability of drugs in hospital settings; availability of care during evenings and weekends; and waiting times for ambulatory emergency care and elective interventions.

Narrowing gaps in access to health care also remains a challenge that needs to be urgently addressed. These gaps affect mostly indigenous communities that concentrate almost 10% of the national population.

In general terms, the most pressing challenge of the Mexican health system is integration, which implies the creation of a national health fund that guarantees access to the same set of health benefits to all Mexicans, the reduction of transaction costs associated to a segmented system, and the universal and egalitarian exercise of the right to health care.

## References

- Central Intelligence Agency. The World Factbook. North America: Mexico. Available at: <https://www.cia.gov/library/publications/the-world-factbook/geos/mx.html>. Accessed 14 Oct 2013.
- Chhabara R. Making the most of the Mexican pharma market. Available at: <http://social.eyeforpharma.com/marketing/making-most-mexican-pharma-market>. Accessed 17 Oct 2013.
- CNNExpansión. Mexicanos adquieren pocos seguros: AMIS. Available at: <http://www.cnnexpansion.com/economia/2009/05/18/aseguradoras-suman-el-17-de-pib>. Accessed 15 Oct 2013.
- Comisión Federal de Protección contra Riesgos Sanitarios. Home Page. Available at: [http://www.salud.gob.mx/unidades/cofepris/notas\\_principal/rimonabant.html](http://www.salud.gob.mx/unidades/cofepris/notas_principal/rimonabant.html). Accessed 15 Oct 2013.
- Comisión Nacional de Arbitraje Médico. Home Page. Available at: [http://www.conamed.gob.mx/main\\_2010.php](http://www.conamed.gob.mx/main_2010.php). Accessed 15 Oct 2013.
- Comisión Nacional de Protección Social en Salud. Sistema de Protección Social en Salud. Informe de Resultados 2010. Available at: [http://www.seguro-popular.salud.gob.mx/images/pdf/informes/Informe\\_Resultados\\_SPSS\\_2010.pdf](http://www.seguro-popular.salud.gob.mx/images/pdf/informes/Informe_Resultados_SPSS_2010.pdf). Accessed 15 Oct 2013.
- Comisión Nacional de Protección Social en Salud. Sistema de Protección Social en Salud. Informe de Resultados 2012. Available at: <http://www.seguro-popular.salud.gob.mx/images/pdf/informes/InformeResultados-2-SPSS-2012.pdf>. Accessed 15 Oct 2013.
- Crónica. Tiene México el mayor número de beneficiarios en salud: FCH. Available at: <http://www.cronica.com.mx/notas/2010/541852.html>. Accessed 15 Oct 2013.
- Dirección General de Evaluación del Desempeño, Secretaría de Salud, México. Misión. Available at: [http://www.dged.salud.gob.mx/contenidos/dged/mision\\_vision.html](http://www.dged.salud.gob.mx/contenidos/dged/mision_vision.html). Accessed 15 Oct 2013.
- Dirección General de Evaluación del Desempeño, Secretaría de Salud, México. Observatorio del Desempeño Hospitalario 2011. Mexico City: Secretaría de Salud. pp. 1–28.
- Dirección General de Información en Salud, Secretaría de Salud, México. Misión, visión y objetivo. Available at: <http://www.dgis.salud.gob.mx/acercade/misionvision.html>. Accessed 15 Oct 2013.
- Frenk J, Sepúlveda J, Gómez-Dantés O. Evidence based health policy: three generations of reform in Mexico. *Lancet*. 2003;362(9396):1667–171.
- Frenk J, Knaul F, Gómez-Dantés O, et al. Fair financing and universal social protection. The structural reform of the Mexican health system. Mexico City: Secretaría de Salud; 2004.
- Frenk J, González-Pier E, Gómez-Dantés O, et al. Comprehensive reform to improve health system performance in Mexico. *Lancet*. 2006;368:1524–34.
- Gakidou E, Lozano R, González-Pier E, et al. Assessing the effect of the 2001–06 Mexican health reform: an interim report card. *Lancet*. 2006;368:1920–35.
- Ham-Chande R. Diagnóstico socio-demográfico del envejecimiento en México. In: Consejo Nacional de Población, México. Mexico City: CONAPO; 2012b. p. 141–55.
- Instituto Mexicano del Seguro Social. Gasto en subrogaciones y servicios integrales 2012 (unpublished report).
- Instituto Nacional de Salud Pública. Encuesta Nacional de Salud y Nutrición 2012. Cuernavaca: INSP; 2013.
- King G, Gakidou E, Imai K, et al. Public policy for the poor? A randomized assessment of the Mexican universal health insurance programme. *Lancet*. 2009;373:1447–54.
- Knaul FM, Arreola-Ornelas H, Méndez-Carniado O, et al. Evidence is good for your health system: policy reform to remedy catastrophic and impoverishing health spending in Mexico. *Lancet*. 2006;368:1828–41.
- Knaul FM, Arreola-Ornelas H, Méndez O, Wong R. Financiamiento y sistema de salud en México: evolución en la desigualdad en la carga financiera entre población afiliada a la seguridad social y afiliados al Seguro Popular. Mexico City: Fundación Mexicana para la Salud; 2011.

- Lozano R, Soliz P, Gakidou E, et al. Benchmarking of performance of Mexican states with effective coverage. *Lancet*. 2006;368:1729–41.
- Massachusetts Office of International Trade and Investment. Mexican pharmaceutical industry. Available at: <http://www.moiti.org/pdf/Mexican%20Pharmaceutical%20Industry.pdf>. Accessed 15 Oct 2013.
- Moïse P, Docteur E. Las políticas de precios y reembolsos farmacéuticos en México, OCDE, 2007. *Salud Publica Mex*. 2008;50(suplemento 4):s504–10.
- Murray CJL, Frenk J. A framework for assessing the performance of health systems. *Bull WHO*. 2000;78(6):717–31.
- National Center for Health Technology Excellence. Mission. Available at: <http://www.cenetec.salud.gob.mx/descargas/folletoingles.pdf>. Accessed 15 Oct 2013.
- OECD. OECD Health Data 2013. How does Mexico compare. Available at: <http://www.oecd.org/els/health-systems/Briefing-Note-MEXICO-2013.pdf>. Accessed 16 Oct 2013.
- OECD. OECDiLibrary. Pharmaceutical expenditure. Available at: [http://www.oecd-ilibrary.org/social-issues-migration-health/pharmaceutical-expenditure\\_pharmexp-table-en](http://www.oecd-ilibrary.org/social-issues-migration-health/pharmaceutical-expenditure_pharmexp-table-en). Accessed 17 Oct 2013.
- OECDiLibrary. Health: key tables from OECD. Practising nurses. Available at: [http://www.oecd-ilibrary.org/social-issues-migration-health/practising-nurses\\_nursepract-table-en](http://www.oecd-ilibrary.org/social-issues-migration-health/practising-nurses_nursepract-table-en). Accessed 16 Oct 2013.
- Organization for Economic Cooperation and Development., OECD StatExtracts. Available at: [http://stats.oecd.org/index.aspx?DataSetCode=HEALTH\\_STAT](http://stats.oecd.org/index.aspx?DataSetCode=HEALTH_STAT). Accessed 15 Oct 2013.
- Partida V. Veinticinco años de transición epidemiológica en México. In: CONAPO. La situación demográfica de México 1999. Mexico City: CONAPO; 1999.
- Reyna-Bernal A, Hernández-Esquível JC. Poblamiento, desarrollo rural y medio ambiente. Retos y prioridades de la política de población. In: CONAPO. La situación demográfica de México 2006. Mexico City: CONAPO; 2006.
- Ruelas E. Citizen's quality councils: an innovative mechanism for monitoring and providing social endorsement of healthcare providers' performance. *Healthcare Papers*. 2006;6(3):33–7.
- Secretaría de Salud. Programa Nacional de Salud 2001–2006. La democratización de la salud en México. Hacia un sistema universal de salud. Mexico City: Secretaría de Salud; 2001. p. 33.
- Secretaría de Salud. Programa Nacional de Salud 2007–2012. Mexico City: Secretaría de Salud; 2007.
- Sistema Nacional de Información en Salud, México. Información por temas. Available at: <http://sinais.salud.gob.mx/estadisticasportema.html>. Accessed 15 Oct 2013.
- Soberón G. El cambio estructural en la salud. *Salud Publica Mex*. 1987;29(2):127–40.
- The World Bank. Gini index. Available at: <http://data.worldbank.org/indicator/SI.POV.GINI>. Accessed 14 Oct 2013.
- The World Bank. Data. GDP growth (annual %). Available at: <http://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG>. Accessed 14 Oct 2013.
- UNDP. International human development indicators. Mexico. Available at: <http://hdrstats.undp.org/en/countries/profiles/MEX.html>. Accessed 14 Oct 2013.
- World Health Organization. World health report 2000. Health systems: improving performance. Geneva: WHO; 2000.
- World Health Organization. Non-communicable diseases. Country profiles 2011. Geneva: WHO; 2012. p. 124.
- World Health Organization. World Health Statistics 2013. Geneva: WHO; 2013.
- World Health Organization. National health accounts. Mexico. Available at: [http://apps.who.int/nha/database/StandardReport.aspx?ID=REP\\_WEB\\_MINI\\_TEMPLATE\\_WEB\\_VERSION&COUNTRYKEY=84027](http://apps.who.int/nha/database/StandardReport.aspx?ID=REP_WEB_MINI_TEMPLATE_WEB_VERSION&COUNTRYKEY=84027). Accessed 15 Oct 2013.
- World Health Organization. National health accounts. Available at: <http://www.who.int/nha/en>. Accessed 15 Oct 2013.
- Ham-Chande R. Diagnóstico socio-demográfico del envejecimiento en México. In: Consejo Nacional de Población, México. Mexico City: CONAPO; 2012a. p. 141–55

## Further Reading

### Three publications by 2000 the same authors were particularly useful for the development of this chapter:

- Frenk J, Gómez-Dantés O. Para entender el sistema de salud. Mexico City: Nostra Editores; 2008.
- Gómez-Dantés O. Mexico. In: Johnson JA, Stoskopf CH, editors. Comparative health systems. Global perspectives. Boston: Jones and Bartlett Publishers; 2009. p. 337–47.
- Gómez-Dantés O, Sesma S, Becerril V, et al. The health system of Mexico. *Salud Publica Mex*. 2011;53(suppl 2):S220–32.



Madelon Kroneman and Willemijn Schäfer

## Contents

<b>Introduction</b> .....	862
<b>Organization and Governance</b> .....	863
Organization of the System .....	863
Planning and Regulation .....	864
The Role of Patients and the Population .....	865
<b>Financing</b> .....	865
Dimensions of Coverage of Curative Care .....	865
Long-Term Care .....	866
Pooling of Funds .....	866
Purchasing Process .....	866
Health Spending and Cost Control .....	867
<b>Physical and Human Resources</b> .....	867
Physical Resources: Hospitals .....	867
Paying the Hospital .....	867
Medical Specialists .....	868
General Practitioners .....	868
Pharmacists .....	869
Nurses .....	869
Other Information on Health-Care Personnel .....	869
<b>Delivery of Health Services</b> .....	869
Public Health .....	869
Primary/Ambulatory Care .....	869
Specialized Ambulatory Care/Inpatient Care .....	870
Pharmaceutical Care .....	870
Long-Term Care .....	870
Mental Health Care .....	870
Dental Care .....	871
Out-of-Hour and Emergency Care .....	871

M. Kroneman (✉) · W. Schäfer  
Netherlands Institute of Health Services Research  
(NIVEL), Utrecht, The Netherlands  
e-mail: [m.kroneman@nivel.nl](mailto:m.kroneman@nivel.nl);  
[willemijn.schafer@gmail.com](mailto:willemijn.schafer@gmail.com)

Informal Care .....	871
Palliative Care .....	871
<b>Reforms</b> .....	871
<b>Assessing the Health System</b> .....	872
Some Indicators of Health and Health Care in the Netherlands .....	872
The Dutch Health-Care System in International Perspective .....	873
<b>References</b> .....	875

## Abstract

A lengthy process of policy efforts to reform the health-care system and to introduce managed competition into the system resulted in the new Health Insurance Act (*Zorgverzekeringswet*) in 2006. A single compulsory health insurance scheme was introduced, and managed competition for providers and insurers became a major driver in the health-care system. This has meant fundamental changes in the roles of patients, insurers, providers, and the government. Insurers negotiate with providers on price and quality, and patients choose the provider they prefer and join a health insurance policy of their choice. The system of managed competition is currently in place for the curative health-care sector and part of the mental health-care sector (ambulatory mental care and institutional mental health care up to 1 year). Since 2006, the role of the national government has changed from directly steering the system to safeguarding the proper functioning of the health-care markets. Long-term care (nursing care and long-term mental care) is regulated by the Long-term Care Act (*Wet landurige zorg*) and the Social Support Act (*Wet maatschappelijke ondersteuning*). During the past decade, social support for disabled and chronically ill and several forms of home care were already transferred to municipalities.

General practice plays a central role in the Dutch health-care system. All citizens are listed with a general practitioner (GP) or GP practice. GPs serve as gatekeepers: patients have to visit their GPs first for their health complaints and only upon referral they can go to a medical specialist. Furthermore, compared to other countries, the relative number of nurses is high.

Dutch citizens are on average very satisfied with their health-care providers, and the accessibility of the health-care system is excellent. However, so far, the Netherlands has not been successful in curbing the growth on health-care expenditure. The government tries to control costs in several ways, for instance, by increasing the compulsory deductible.

## Abbreviations

GDP	Gross domestic product
GP	General practitioner
OECD	Organisation for Economic Co-operation and Development

## Introduction

The Netherlands is situated in Western Europe and borders the North Sea, Germany and Belgium. It covers an area of 41,543 km<sup>2</sup> (Centraal Bureau voor de Statistiek 2009) and has a population of 16.8 million in 2013, the majority of whom (79%) are native Dutch (Statistics Netherlands 2013). The Netherlands has the tenth largest economy in the world and ranks 16 in GDP (Ministry of Foreign Affairs 2013). Between 1970 and 2011, the life expectancy at birth of the Dutch population has grown from 73.6 to 81.3 years (Statistics Netherlands 2013). The infant mortality declined from 4.9 per 1,000 live births in 2005 to 3.6 in 2011, which is slightly below the average rate for all OECD countries (4.4 per 1,000 live births) (OECD 2013). In 2011, most deaths are caused by malignant neoplasms (cancer), which is in contrast with other EU countries, where diseases of the circulatory system are the main cause of



death. The burden of disease is higher among immigrants than among native Dutch inhabitants. Important risk factors affecting the health of the Dutch population are smoking and overweight. Between 2000 and 2010, the average of regular daily smokers was slightly below the EU average (World Health Organisation 2013). According to self-reported data, almost half of the population is overweight (Statistics Netherlands 2013).

---

## Organization and Governance

### Organization of the System

A lengthy process of policy efforts to reform the health-care system and to introduce managed competition into the system finally resulted in the new Health Insurance Act (*Zorgverzekeringswet*) in 2006. With the introduction of a single compulsory health insurance scheme, the former dual system of public and private insurance for curative care was abandoned. Managed competition for providers and insurers became a major driver in the health-care system. The new system introduced new roles for patients, insurers, health-care providers, and the government. Three markets exist: the health insurance market, the health provision market, and the health purchasing market. Within the health-care purchasing market, insurers have to negotiate with providers on price, quality, and volume of care. In the health-care provision market, patients can choose the provider they prefer. In the health insurance market, citizens can join a health insurance policy which best fits their needs and requirements. The system of managed competition is currently in place for the curative health-care sector and part of the mental health-care sector (ambulatory mental care and institutional mental health care up to 1 year). Since 2006, the role of the national government has changed from directly steering the system to safeguarding the proper functioning of the health-care markets. With the introduction of market mechanisms in the health-care sector and the privatization of former public health

insurance funds (sickness funds), the Dutch system represents an innovative and unique variant of a social health insurance system.

The Dutch population aged 18 years and older is obliged to take a health-care insurance for the basic health-care package. Children under the age of 18 are included in the policy of one of their parents, and their premium is paid by the government. Health insurers are obliged to accept applicants without restrictions. Differentiation of premiums for different risk conditions (such as age, sex, and chronic diseases) is not allowed. Health insurers are free to set community-rated premium and to contract health-care providers, under the condition that they have to operate within the national health-care budget set by the government and that they have to contract sufficient providers to ensure good access to care for their insured population. Health insurers are compensated for their insured with high risk for health-care costs via a risk adjustment scheme. In addition to the basic insurance package, health insurers offer voluntary complementary insurance for care that is not covered by the Health Insurance Act. For instance, a (partly) coverage of glasses or dental care is often part of the voluntary health insurance.

General practice plays a central role in the Dutch health-care system. All citizens are listed with a general practitioner (GP) or GP practice. GPs serve as gatekeepers: patients have to visit their GPs first for their health complaints, and only upon referral they can go to a medical specialist. About 96% of all contacts are dealt with within primary care (Cardol et al. 2004). An important prerequisite is that GP care in the Netherlands is freely accessible and exempted from the compulsory deductible which is currently in place for other forms of care.

Long-term care (nursing care and long-term mental care) is regulated by the Exceptional Medical Expenses Act (*AWBZ*). This Act was intended originally (1968) to provide care for those with chronic conditions requiring continuous care that involves considerable financial consequences. Since the introduction of the Act, many types of care have been added resulting in a rapid growth in expenditure in such a way that

the affordability became at risk and thus the call for reform became urgent. During the past decade, social support for disabled and chronically ill and several forms of home care were already transferred to municipalities. In 2015 the long-term care in the Netherlands was completely reformed. Home nursing care for people who require 24 hours supervision per day is now regulated by the Long-term Care Act (Wet langdurige zorg). People in the need of care who live at home receive care through the Social Support Act (Wet maatschappelijke ondersteuning) which is the responsibility of municipalities. Home nursing care became part of the Health Insurance Act and is now the responsibility of health insurers.

## Planning and Regulation

The role of the Dutch government is steering from a distance. They define the framework in which health care can be developed. Responsibilities have been transferred to insurers, providers, and patients, and the government only supervises quality, accessibility, and affordability of health care. The establishment of new supervisory agencies in the health sector aims to avoid undesired market effects in the new system. Traditionally, self-regulation has been an important characteristic of the Dutch health-care system. Professional associations are responsible for reregistration schemes and are involved in quality improvement, for instance, by developing professional guidelines.

## Responsibilities of the National Government

The government should ensure that managed competition results in safe, accessible, and affordable health care of good quality. Only a few instruments have been left to the government to directly interfere in the health-care system. An essential competence of the government is setting the budget for health-care expenditures. Other important competences of the central government are taking decisions on the content of the basic health insurance package and on cost-sharing. Furthermore, in order to prevent preferred risk selection, the government

sets the rules for risk adjustment among health insurers. In the care sector, the central government has a number of explicit responsibilities. These include creating the preconditions for quality, accessibility, safety, and affordability of the care for people with chronic conditions; strengthening the position of citizens, in particular patients and their representatives; and stimulating innovation. To meet these responsibilities, the government has supervisory and advisory bodies in place. Furthermore, at national level, there is legislation which describes the conditions in which the markets have to operate.

## Supervisory Bodies

Independent supervisory bodies take care of safeguarding accessibility, affordability, and quality of care:

- The Dutch Healthcare Authority (NZa) supervises the compliance of actors with the Health Insurance Act (Zvw) and the Health Care Market Regulation Act (Wmg). NZa interferes with restrictions or obligations when an actor, that is a health insurer, health-care provider, or consumers, together or alone, hinders fair competition in (part of) the health-care market. Furthermore, the NZa establishes tariffs and performance directions for those health services that are not subject to free negotiations. Lastly, the NZa monitors health-care markets and promotes its transparent and fair operation. In addition, the NZa imposes on tariffs for health services that are not freely negotiable and on extending the share of freely negotiable services.
- The National Healthcare Institute (Zorginstituut Nederland) advises the Ministry of Health, Welfare and Sport on the content of the basic health insurance package. Furthermore, it supplies information to insurers (but also consumers and providers) on the nature, content, and scope of the basic health insurance. The Healthcare Institute also administers the Health Insurance Fund and operates the risk adjustment scheme.
- The Health and Youth Care Inspectorate (IGJ) supervises the health-care providers in the areas of quality and safety.

## The Role of Patients and the Population

Within the Dutch health-care system, the population is free to choose a health insurer. The idea is that people will choose those insurers with the best price/quality performance. In practice, the main reason for people to switch is the level of the premium for the basic insurance package and competition on quality of care seems to be absent (see [www.hspm.org](http://www.hspm.org)). Patients are expected to choose providers based on quality (for instance, through providers selected by health insurer and/or by comparing providers on quality on the website [www.kiesbeter.nl](http://www.kiesbeter.nl)). In practice, patients follow the recommendation of their GP in choosing a health-care provider (Dautzenberg et al. 2012; Reitsma et al. 2012).

---

## Financing

### Dimensions of Coverage of Curative Care

Basic health insurance is obligatory for all Dutch residents. Children under the age of 18 are insured free of charge but have to be included in one of the parents' policies. The nominal premium for children is paid by the government. For persons aged 18 or over, there is a compulsory deductible of €385 in 2018. Excluded from this deductible are GP care, maternity care, and dental care under the age of 18. In addition to the compulsory deductible, people can choose for a voluntary deductible. This voluntary deductible may range from €100 to maximum €500 in exchange for a reduction on the premium.

The basic health insurance covers all curative (somatic and mental) health care that is considered essential, effective, cost-effective, and unaffordable for individuals. "Essential" refers to its capacity to prevent loss of quality of life or to treat life-threatening conditions. The affordability criteria state that no services need to be included that are affordable for individual citizens and for which they can take responsibility (Brouwer 2004). The content of the benefit package is defined by the government and covers more or

less all primary and secondary curative care. Excluded are dental care for persons older than 18 years of age and some elective procedures such as plastic surgery without medical indication and, since 2013, simple walking aids. Partly covered are, for instance, allied health care, some medicines, and in vitro fertilization.

Citizens pay for their health insurance through a community-rated premium and an income-dependent contribution. For 2013, the community-rated premium varied from €92 to €112 per month. Health insurers are free to set the premium level. The insured persons pay these premiums directly to their health insurer. For children below the age of 18, the government covers the premium through a contribution into the Health Insurance Fund. Insurers are not allowed to differentiate the premium of one specific policy for the basic benefit package for different groups of people. There is one exemption: insurers may offer collective contracts. Collective contracts are established between groups of insured (e.g., a company with employees) and the insurance company. Insurance companies are allowed to offer a maximum of 10% reduction on the individual premium. Insured people are free to join a collective policy or buy an individual policy. The system of collective policies is established to give the insured more influence ("voice") on the insurance companies. The threat of the loss of a large number of insured persons may persuade insurers to satisfy the collectivity and compete on price and quality of care. In addition, successful negotiations may lead to more demand-driven care and care that is tailored to the need of the target group of the collective. In 2012, 67% of the insured persons participated in a collective insurance policy.

The income-dependent contribution is collected by the Tax Office, which levies the contribution from salary together with payroll taxes. After collecting all the contributions, the Tax Office transfers the money to the Health Insurance Fund (*Zorgverzekeringsfonds*), where the money is allocated after risk adjustment to the health insurers.

To ensure access to basic health insurance under a system with flat rate premiums and to

compensate for undesired income effects for lower-income groups, a “health-care allowance” funded from general tax was created. In 2011, six out of ten households received a health-care allowance of on average €85 per month. People with chronic diseases or a handicap receive a compensation of €99 per year for the compulsory deductible in 2013.

### **Voluntary Health Insurance (VHI)**

Most insurance companies offer voluntary packages in combination with the basic benefit basket. In 2012, 88% of the insured took out complementary VHI (Ten Hove et al. 2012). VHI covers for care that is not included in the basic package, for instance, dental care, glasses, or physical therapy (for persons without a chronic indication). In addition, some co-payments may be covered, for instance, for ambulatory mental care. Contrary to basic health insurance, health insurers are free to set premium levels and may apply preferred risk selection for complementary VHI based on medical criteria or other risk factors. Insurers are obliged to offer VHI independent from the basic health insurance, but some insurers discourage taking VHI without a basic insurance by increasing the premium or by stating that VHI can only be taken when a basic insurance is taken at the same insurer (Roos and Schut 2009).

### **Long-Term Care**

#### **Exceptional Medical Expenses Act (AWBZ)**

Long-term care is insured under the Long-term Care Act (Wlz). This is a social health insurance scheme that is intended to provide care for those with chronic conditions (physical and/or mental) requiring requiring 24 hour supervision (either physically, mentally or medically) per day. Everyone who is legally residing in the Netherlands or pays payroll tax in the Netherlands is compulsory insured. At present (2018), long-term care at home is provided by municipalities under the Social Support Act (Wmo). Home nursing is provided by health insurers under the Health

Insurance Act. Under certain conditions, people can receive a personal budget to buy the care they need.

To cover the expenses for the Wlz, a contribution of 9.65% is levied on the salary of the citizens, with a maximum of €3,280 per year (2018). The revenues are collected by the Tax Office and transferred to the Long-term Care Fund, administered by the Netherlands Healthcare Institute. The expenses for the Social Support Act are covered by general taxes and are transferred to the municipalities through the municipality fund. The budget is not earmarked.

### **Pooling of Funds**

In the Netherlands, administering and providing basic health insurance are delegated to private health insurers. These insurers are funded by the nominal premium directly received from clients and a contribution from the Health Insurance Fund, which pools the income-dependent employer contributions (collected by the Tax Office) and the state contribution (e.g., to cover children under 18). The allocation among the health insurers is based on the health risk profile of their insured population. The government sets the level of the income-dependent contribution, with the notion that, at national level, the total income-dependent contributions for adults should amount to approximately 50% of the total funding of basic health insurance, while the nominal premiums should account for the other 50%.

### **Purchasing Process**

Health insurers buy health care for their insured population (possibly by selective contracting). They negotiate contracts with hospitals (on volume and quality but also lump sum) and with committees that represent GPs. The negotiations with GPs are in practice hardly on tariff but more on activities aimed at increasing GP care and substitution of secondary care to primary care (modernization and innovation activities).

Purchasing of long-term institutional and home nursing care is delegated to health insurers. Health insurers negotiate with providers on price, volume, and quality of care.

## Health Spending and Cost Control

Initially, after the reform in 2006, the community-rated premiums were set too low, because health insurers tried to attract the population via these low premiums. This resulted in a loss for health insurers. Over the years, premiums have slowly increased. Since 2009, insurers have been able to realize a profit on their insurance policies (Nederlandse Zorgautoriteit [Dutch Healthcare Authority] 2013).

So far, the Netherlands has not been successful in curbing the growth on health-care expenditure. The government tries to control costs in several ways, for instance, by increasing the compulsory deductible. Initially, in 2008, the deductible was €150. This increased to €350 in 2013. Furthermore, some care is taken out of the basic package, such as simple walking aids in 2013. For long-term care, the government tries to increase the involvement of citizens, by stimulating them to take care of family and neighbors on a voluntary basis. Since 2006, the growth in expenditure varies from 6.8% (2008) to 2.3% (2011). In 2012, the growth was 3.7% (Centraal Bureau voor de Statistiek 2013a, b).

---

## Physical and Human Resources

### Physical Resources: Hospitals

The structure of health care in the Netherlands comprises a dense network of premises, equipment, and other physical resources. In 2010, there were 8 university hospitals and 84 acute care hospitals in the Netherlands, subdivided into 28 top clinical centers and 57 general hospitals (Nederlandse Vereniging van Ziekenhuizen 2012). In 2009, there were 2.8 beds per 100,000 inhabitants, which is among the lowest in Europe. In addition to general and

university hospitals, independent treatment centers have become part of the acute care hospital sector. These private centers provide selective non-emergent treatments for admissions up to 24 h.

Most hospitals are corporations. Hospitals are nonprofit institutions as a for-profit motive is not allowed. Since 2008, however, a few pilots have started that allowed paying out a part of the profit to shareholders. Attracting shareholders might give hospitals the opportunity to generate more investment for quality improvement and innovation. Whether or not hospitals should be allowed to generate profit and to have shareholders is still (2012) a topic of political debate.

Within hospitals, approximately 55% of medical specialists are self-employed and organized in partnerships (Nederlandse Zorgautoriteit [Dutch Healthcare Authority] and DBC-onderhoud 2012). These partnerships usually work in one hospital. In a few hospitals, especially university hospitals, the specialists are employed by the hospital. In 2012, there were 21,750 registered medical specialists. The largest categories were psychiatrists (3,299), internists (2,168), and anesthesiologists (1,805) (KNMG 2013).

### Paying the Hospital

Hospitals are paid through Diagnosis Treatment Combinations (*Diagnose Behandeling Combinaties, DBCs*) since 2005. The DBC system was based on the concept of DRGs (Diagnosis-Related Groups), but it constituted a newly developed classification system. The DRG system is based on the diagnosis of a patient, and there is one DRG per patient for each hospital episode. The DBC system provides a DBC for each diagnosis treatment combination, and thus, more than one DBC per patient is possible. The system was, however, considered too complex and error-prone. Therefore, by 2012, the system was updated. New DBCs were formulated, and the number of DBCs was reduced from 30,000 to 3,000. The DBC tariffs include the costs of medical specialist care, nursing care, and the use of medical equipment and diagnostic procedures.

Apart from these direct costs, also indirect costs such as education, research, and overhead are included. The reimbursement for each DBC is not influenced by longer hospital or shorter hospital stay or a deviant number of diagnostic procedures for a certain patient.

Since the introduction of the DBC system, there were two segments: the freely negotiable segment and the regulated segment. To get used to the new system, in which health insurers and hospitals had to negotiate prices for the DBCs, only a small part (10% in 2005) was freely negotiable, and the prices for the regulated part were based on the former system of paying the hospital. Gradually the freely negotiable part increased. In 2012, the former system of paying the hospital was abolished, with a transition model for the years 2012 and 2013. Now there is a freely negotiable part (about 70% of the DBC turnover) in which hospitals and insurers are free to set prices and a regulated part for which the Dutch Healthcare Authority (one of the supervisory organizations) establishes maximum prices. In practice, some insurers do not negotiate prices for the DBCs but negotiate a lump sum amount with the hospitals.

As compensation for investments is included in the tariffs, since 2008 for hospitals and since 2009 for long-term care institutions, health institutions are fully responsible for the realization of their (re)constructions and the purchase of equipment. No external approval of building plans applies, although the quality of premises is externally assessed every 5 years.

### Medical Specialists

Medical specialists are either independent professionals organized in partnerships working in a hospital (55% in 2010) (Nederlandse Zorgautoriteit [Dutch Healthcare Authority] & DBC-onderhoud 2012); or they are in salaried service of a hospital. Since 2008, medical specialists are paid through the DBC system. The independent partnerships have to negotiate their tariffs with the hospital they work in.

### General Practitioners

In 2012, there were 8,879 GPs (53 per 100,000 inhabitants), 43% of whom were female. GPs work in independent practices, either alone (26%) or with two or more other GPs (74%). Patients are listed with a GP practice. About 11% of the GPs work in salaried service for other GPs; the majority of these salaried GPs is female (87%) (Van Hassel and Kenens 2013).

GPs receive a capitation fee per patient per year. For older patients and patients from deprived areas, a higher fee is applicable, but this is only paid if there is an agreement with the health insurer (Nederlandse Zorgautoriteit [Dutch Healthcare Authority] 2011). Per patient contact the GP receives a fee, differentiated toward practice consultations, home visits, telephone consultations, and prescription refills. Practice nurses take part in the routine care for chronically ill persons in the general practice, like diabetes, hypertension, and COPD/asthma. Fees for practice nurses are freely negotiable or are part of integrated care agreements. Integrated care agreements are financed via bundled payments. Integrated care addresses the care for patients with the following chronic conditions, diabetes type II and COPD, and persons with high risk for cardiovascular diseases. According to the system of bundled payments, a care group organizes all care that is necessary for managing these diseases. Care groups are owned by GPs in a certain region; they vary in size from 4 to 150 GPs. The care group coordinates the care and pays the different care providers who are involved in the care. Patients are free to participate in integrated care or to organize the necessary care themselves. Besides the abovementioned payment methods, GPs may negotiate with insurers for the financing of activities for improvement of efficiency or substitution of care. These activities are only reimbursed if this is negotiated in a contract with the health insurer.

Out-of-hour services for GP care are mostly provided by GP out-of-hour cooperatives. GPs who participate in this system receive a per hour compensation. The majority of GPs participate in a GP out-of-hour cooperative (approximately 97% in 2013).

## Pharmacists

For pharmaceutical care, provided by pharmacists or dispensing GPs, remuneration is based on pre-defined activities that are described by the Dutch Healthcare Authority. Examples of such activities are dispensing the medicine to the patient and providing information about the medication. Health insurers and pharmacists can freely negotiate prices for these activities.

## Nurses

Compared to other countries, the relative number of nurses is particularly high. Most nurses are working in home care and in care for the elderly and disabled. Substitution and transfer of tasks from medical to nursing professionals is an important trend. For instance, practice nurses, who take care of chronic patients with certain diagnoses, are since 2011 allowed to prescribe medicines (Editorial Office Nursing 2011), and specialist nurses caring for pulmonary and diabetes patients are allowed to prescribe medicines since 2013 (Oelen 2013).

## Other Information on Health-Care Personnel

Medical education is provided at each of the eight Dutch universities, while nurses can either be educated at an intermediate, higher, or academic level, depending on the professional profile. The quality of health-care professionals is safeguarded by obligatory registration and by various licensing schemes. Workforce forecasting and careful planning of educational capacity seek to prevent shortages or oversupply of health professionals. In a small and densely populated country like the Netherlands, unequal distribution of providers is not a major issue, although in some parts of large cities, additional efforts need to be made to match demand and supply. In 2012, about 15% of the working age population was working in the health-care sector (including home care, child care, and social support).

## Delivery of Health Services

### Public Health

Disease prevention, health promotion, and health protection fall under the responsibility of municipalities. A number of uniform tasks are specified in the Public Health Act (Wpg) and include among others youth health care, public health for asylum seekers, medical screening, and community mental health.

Youth health care (*jeugdgezondheidszorg*) provides preventive and mental care for all children aged between 0 and 19 years. Until the age of 4, children visit child health centers (*consultatiebureaus*) for checkups. The most important tasks of preventive health care are the monitoring of growth and development, early detection of health or social problems (or risks), screening and vaccination, and providing advice and information concerning health. This care is provided by specialized physicians and nurses. When treatment is necessary, the child health center will refer to other primary health-care providers, mostly GPs. Youth mental care is the responsibility of municipalities.

The National Vaccination Programme (*Rijksvaccinatieprogramma*, RVP) consists of childhood vaccinations (DTP-Hib-HepB, MMR, MenC, pneumococci, and HPV for girls of the age of 12). Other national screening programs are screening for cervical cancer, breast cancer, and vaccination against influenza. The heel prick for newborns screens for 17 diseases.

### Primary/Ambulatory Care

In the majority of cases, the first point of contact for people with a medical complaint will be their GP. The GP has a central role in the health-care system and acts as gatekeeper of the system. This means that for “prescription-only medicines” or medical specialist care, a prescription or referral from a GP is required. For specific problems, patients can also directly access allied health professionals, such as physiotherapists and remedial therapists. However, these professionals are not

qualified to prescribe medication or to refer to secondary care. Two other directly accessible primary care professionals are midwives and dentists. These disciplines are also qualified to refer to some forms of secondary care, such as gynecologists in case of midwives and dental surgeons in case of dentists. They are also allowed to prescribe some types of medication.

Patients register with a GP of their choice and can switch to a new one without restriction. However, GPs have the right to refuse a patient. Reasons to refuse patients can be that the patient lives too far from the practice or because GPs have too many patients on their list. Almost 100% of the population can reach a GP within 15 min from their home. GPs can usually be visited within 2 days. A full-time GP has a practice list of approximately 2,350 persons.

### **Specialized Ambulatory Care/Inpatient Care**

Dutch hospitals provide practically all forms of outpatient as well as inpatient secondary care. Except in cases of emergency, patients only consult a specialist upon referral from a GP. Most hospitals also have 24-h emergency wards.

### **Pharmaceutical Care**

The supply of prescription-only pharmaceuticals is exclusively reserved to pharmacists and dispensing GPs (in some rural areas). Over-the-counter (OTC) pharmaceuticals for self-medication are available both at pharmacies and chemists. There are three types of pharmacies: public pharmacies, hospital pharmacies, and dispensing general practices. Public pharmacies should be reachable within 4.5 km from the patient's home. If this is not the case, a local GP can ask for a dispensing license. In 2011, about 6% of the GP practices are dispensing medicines. A new development in pharmaceutical care is the emergence of Internet pharmacies. In 2013, there are eight Internet pharmacies active. Most of them do not have a physical location and deliver medicines by courier services.

### **Long-Term Care**

Long-term care is provided both in institutions (residential care) and in communities (home care). Long-term institutional care is financed by the Long-term Care Act (Wlz). The Centre for Needs Assessment (CIZ) has been commissioned by the government to carry out assessment for eligibility under the Wlz. Patients, their relatives, or their health-care providers can file a request with the CIZ for long-term care. The CIZ assesses the patient's situation and decides what care is required. Patients can choose between receiving a personal care budget (only in the case that they need care for more than 10 h per week) to purchase care themselves and receiving the care in kind. Personal budgets are subject to discussion because of a number of fraud cases.

Nursing homes are especially for people with severe conditions who require constant nursing care. All others in need of care receive this care at home. The majority of the residents in nursing homes and residential homes are older than 80 years.

Home care is provided by home care organizations. Besides care for the elderly and people with disabilities, home care organizations provide maternity care. Since the long-term care reform of 2015, the number of people who are eligible for nursing homes have decreased drastically. It is the policy of the Dutch government to keep people at home as long as possible.

### **Mental Health Care**

Mental health care is provided both in primary and in secondary health care. Primary health-care professionals in mental health care include GPs, psychologists, and psychotherapists. When more specialist care is required, the GP refers the patient to a psychologist, an independent psychotherapist, or a specialized mental health-care institution. When the mental problems can be handled within general practice, the GP may refer to a mental care practice nurse (praktijkondersteuner GGZ), who is working within the practice.



The first three years of mental health treatment are part of the basic health insurance and are thus financed under the Health Insurance Act (Zvw). The funding of preventive mental health care and youth mental care is part of the Social Support Act (WMO), which means that the responsibility for organizing this care lies with the municipalities.

## Dental Care

Oral health care is provided in primary care by private dentists and dental hygienists. Most citizens register with a dentist. Most dentists work in small independent practices (about 70%). Dental hygienists are specialized in preventive care and can be visited directly or upon referral from the dentist. Preventive tasks and relatively simple dental care are increasingly being substituted to dental hygienists. Nine out of ten dentists regularly refer to a dental hygienist either in their own practice, to the practice of a colleague, or to an independent dental hygienist practice.

## Out-of-Hour and Emergency Care

Patient with nonlife-threatening conditions goes to the special GP cooperatives for out-of-hour care. For life-threatening conditions or upon referral of the GP in the GP post, patients can go to the 24-h emergency department of the hospital.

## Informal Care

The estimates of the number of people who provide informal care vary from approximately 1.7 million people (Oudijk et al. 2010) to 3.7 million (Houben-van Herten and Te Riele 2011). Informal carers (60% women, about half in the age of 45–65 years old) provided care (emotional support, household work, accompanying during visits to family) to ill or disabled people, mostly to parents (40%) or spouses (18%). It is the policy of the government to stimulate informal care, in order to keep healthcare affordable.

## Palliative Care

Palliative care is provided by general practitioners, home care, nursing homes, specialists, and voluntary workers at home. Furthermore, there are growing numbers of hospices and palliative units (e.g., in nursing homes). Most palliative care is integrated into the regular health-care system.

## Reforms

The main reform in the Dutch health-care system took place in 2006. The dual system in which two third of the population (earning an income below a certain threshold) was insured publicly and one third privately was abolished. Since 2006, there is one insurance system for all citizens, with a community-rated premium that cannot be differentiated toward different risk groups. Insurers are obliged to accept citizens who apply for a health insurance policy. Together with this reform, the financing system changed. Although some aspects of market forces were already incorporated into the system before the reform, since 2006, market mechanisms became officially introduced into the system. This imposed a new role for especially health insurers and health-care providers. They had to learn to negotiate on price, volume, and quality. To ensure a smooth transition, in the first years, only a small part of the provided care was freely negotiable. This share increased over the years, and in 2012, about 70% of the hospital care expenditure was freely negotiable, with the remaining 30% being regulated covering care that is too difficult or not suitable for free market negotiations, such as intensive care in hospitals. The Dutch Healthcare Authority defines the care activities that are subject to remuneration. The prices for these activities in the free segment can be negotiated by the market parties, although for some issues, maximum prices are set. For instance, for the remuneration of independent medical specialists, a maximum hourly tariff is set. Selective contracting by insurers is allowed, as long as insurers can assure sufficient care for their clients. However, until recently, none of the large insurers opted for selective contracting. There has been one

attempt by a large insurer to refrain to contract a large hospital in the Dutch capital in 2012, which got a lot of attention in the Dutch newspapers. The hospital finally agreed with the lower budget and thus can still provide care to their patients.

Another important reform is found in the Exceptional Medical Expenses Act (AWBZ). This Act regulated long-term care in the Netherlands up to 2015. However, over the years, the act encompassed more and more care activities, leading to a strong increase in expenditure. The main target of the reform is to reduce the care insured under the act to care where it initially was meant for: care that is unaffordable for individual citizens and their insurers. This is, for instance, care in a medical home for the elderly. The following care was transferred from the AWBZ to other acts. Home help and social support became a responsibility of municipalities under the Social Support Act (WMO). Curative mental care was transferred to the Health Insurance Act and became part of the basic insurance package (for the first three years of care). Youth care was transferred to municipalities under the Youth Act. Home nursing care is transferred to the Health Insurance Act. The most important consequence of this choice is that under the Health Insurance Act, citizens have a right on certain care whereas under the municipalities, the emphasis will be on individual responsibility. Municipalities have the obligation to compensate citizens in such a way that they can participate in the society. The individual circumstances of the citizen may be taken into account. This is called the compensation principle: tailor-made measures instead of rules. The reform came with a major reduction in the budget, since municipalities were considered to be closer to their citizens and thus better able to efficiently organize the care.

---

## Assessing the Health System

### Some Indicators of Health and Health Care in the Netherlands

The Dutch government stipulated in the explanatory note accompanying the health-care budget in 2013 that essential care of good quality should

be available and affordable for all citizens. The increasing demand for care and increasing costs as a result of technological and demographic developments may result in fundamental changes in health care. People are encouraged to stay at home as long as possible, with the aid of informal carers and volunteers. Examples of new initiatives are institutional care providers, who aim to agree by contract with informal carers to provide a minimum of 4 h of informal care per month. This led to a lot of societal commotion. Furthermore, mild forms of institutional care are no longer provided, and new patients needing this type of care will receive this care at home.

Dutch citizens are on average very satisfied with their health-care providers (they give a score of 7.7–7.9 on a scale of 1–10) (Statistics Netherlands 2012). Healthy persons are slightly more satisfied than persons with ill health, and lower-educated people are more satisfied than young people and higher-educated people.

In 2011, life expectancy for males was 79.2 years and for females 82.9 years. In the past decade, the life expectancy for men increased with 3.4 years and for women with 2.2 years. Healthy life expectancy increased significantly for men (from 9.2 to 10.9 healthy years for 65-year-olds) but not for women (Statistics Netherlands 2012).

Mortality from cardiovascular diseases has steadily decreased over the past decade. Several factors have contributed to this decrease, such as a better treatment of high cholesterol and high blood pressure and more attention for a healthy lifestyle. Furthermore, more people are aware of the fact that they have a high blood pressure, making treatment possible. Besides, the development in technological options to treat cardiovascular diseases resulted in more patients surviving the disease (Statistics Netherlands 2012). Mortality due to cancer increased lightly in the past decade. In 2008, cancer got ahead of cardiovascular diseases as most important cause of mortality.

Affordability of health care is still a cause of debate in the Netherlands. Expenditure on health care continues to increase over the years, both due

to increasing prices and an increase in volume of care. The government wishes to diminish especially the increase in volume of care. From 2006 to 2011, the expenditure on care under the Health Insurance Act and the Exceptional Medical Expenses Act increased with on average 4.4% per year. In 2011, the expenditure increased with 3.6%.

Citizens find accessibility of and solidarity in health care important. However, citizens appear to have little insight in health-care expenditure. They are aware of the compulsory deductible and of the community-rated premium for the Health Insurance Act, but they are hardly aware of the income-related premiums for the Health Insurance Act and the Exceptional Medical Expenses Act that is paid by their employer directly to the government (Kooiker et al. 2012). Competition in care is not popular among Dutch citizens, it is associated with a profit orientation, expensive managers, and a large overhead (Kooiker et al. 2012).

The accessibility of the Dutch health-care system is excellent. Nearly all citizens are insured, and waiting times are on average acceptable. There are a few specialisms that have a larger waiting time than the norm of 4 weeks for a first appointment and only for a few treatments the waiting time exceeds what is seen as acceptable.

Competition in the health insurance market seems to be present. In 2012, 6% of the citizens switched insurers, and in 2013, this was 8.3%, which can be seen as an indicator that competition in this market is present. In the health-care purchasing market, nearly all general practitioners are contracted by the health insurers for the maximum tariff (Nederlandse Zorgautoriteit [Dutch Healthcare Authority] 2012). Health insurers managed to contract 90% of the hospitals for the year 2011 before the end of that year, which is rather late, considering that health insurers have to publish their premiums in November of the year before. To evaluate quality of care, several indicators have been developed by the Dutch Healthcare Authority, but these are not yet published due to the fact that they cannot yet be corrected for casemix. The development

in prices for the freely negotiable part of hospital care showed a decrease in 2010 of 3% and in 2011 of 1.3%. However, these decreases are mainly due to the tariff caps for medical specialists that were the result of the large increase in medical specialist's income in the years before. Selective contracting in the health purchasing market is currently still in its infancy. In 2012, a large insurer decided to not contract a large hospital, but later that year, the hospital accepted the lower tariffs proposed by the insurer. In the health-care provision market, patients mainly go to the medical specialist who is advised by their general practitioner. There is information available on the Internet on quality of care, but consumers find it difficult to use this information (Damman et al. 2012).

## The Dutch Health-Care System in International Perspective

When looking at health-care supply, the Netherlands has a low number of acute care hospital beds with 301 beds per 100,000 inhabitants in 2010, below the EU average, but 10 countries have a lower number of beds, with Finland on top with about 180 beds per 100,000 inhabitants. The supply of long-term beds (in nursing and elderly homes) is large compared to most European countries, with 1,036 beds per 100,000 inhabitants. For those countries where information is available, only Finland and Malta have a higher supply of long-term care beds in 2011. The Netherlands has nearly the lowest number of physicians in Europe (58 physicians per 100,000 inhabitants), with only Denmark and Ireland having even lower numbers. The number of general practitioners is also below the EU average, with 72 GPs per 100,000 inhabitants, the EU average being 82 GPs per 100,000 inhabitants (World Health Organisation 2013).

Acute care hospital admission rates are among the lowest in Europe with 11.4 admissions per 100 inhabitants in 2009. Since 2001, with 8.8 admissions per 100 inhabitants, the number of admissions is increasing. The average

length of stay had decreased considerably from above the EU average in 2000 with 9 days to 5.6 days in 2009, which is below the EU average. The number of doctor's consultations is slightly below the EU average, with 5.8 consultations per person in 2009. Health-care expenditure as percentage of GDP is the highest among Europe with almost 12% in 2010 (World Health Organisation 2013).

This chapter is mainly based on the Health System Review of the Netherlands (Schäfer et al. 2010) and the publications in [www.hspm.org](http://www.hspm.org): The Netherlands, with updates where necessary.

**Box 1 Main features of the most important acts that regulate the Dutch health care system**

*Health Insurance Act (Zorgverzekeringswet)*

Regulates the compulsory basic health insurance for citizens, the voluntary and compulsory deductible, the obligation for health insurers to accept every person who applies for a policy, the risk adjustment system to compensate health insurers for persons with high health-care consumption, and the supervision of the system.

*Long-term Care Act*

This act is a social health insurance scheme and regulates the admission of people in nursing homes. People should need 24 hours supervision for being eligible for this type of care.

*Youth Care*

This act regulates mental care and educational support for children under the age of 18 and their parents. The care is organized by municipalities.

*Health Care Allowance Act (Wet op de Zorgtoeslag)*

Regulates that people with low incomes are partly compensated for the community-rated premium, in order to keep health insurance affordable for this group.

*Social Support Act (Wet Maatschappelijke Ondersteuning)*

**Box 1 (continued)**

Introduces the right for each citizen to be able to fully participate in society; municipalities should help to overcome barriers. For instance, home help, transportation, home adaptations, sheltered housing, and wheelchairs can be applied for by the municipality.

*Health Care Market Regulation Act (Wet marktordening gezondheidszorg)*

This act regulates the development, structuring, and supervision of the health-care markets. The act regulates the establishment of the Dutch Healthcare Authority as an independent administrative organization that supervises the health-care markets.

*Health Care Admission Act (Wet Toelating Zorginstellingen)*

Health-care institutes need an admission if they provide care under the Health Insurance Act or the Exceptional Medical Expenses Act. A request is handled by the Central Information point Professions in Health Care (Centraal Informatiepunt Beroepen Gezondheidszorg).

*Public Health Act (Wet publieke gezondheid)*

The act regulates collective prevention, infectious diseases control, and youth care.

*Individual Health Care Professions Act (Wet op de Beroepen in de Gezondheidszorg)*

Regulates the care provision by health-care professionals and the quality of care. A second aim is protection of patients. Professionals have to register in the BIG registry.

*Medical Treatment Agreement Act (Wet op de Geneeskundige Behandelovereenkomst)*

Regulates the right to information, consent for medical treatment, and access to medical files. The Act further regulates the requirement of confidentiality and the right to privacy during medical treatment.

## References

- Brouwer WBF. Met het oog op gepaste zorg. Deel I: Over-, onder en gepaste consumptie in de zorg vanuit economisch perspectief [With a view to suitable care, part I: over-, under- and suitable consumption of care from economic perspective]. Zoetermeer: Council for Public Health and Health Care (RVZ); 2004.
- Cardol M, Van Dijk L, De Jong JD, De Bakker D, Westert GP. Tweede Nationale Studie naar ziekten en verrichtingen in de huisartspraktijk: huisartsenzorg: wat doet de poortwachter? [Dutch National Study of General Practice: GP care: activities of the gatekeeper]. Utrecht: NIVEL; 2004.
- Centraal Bureau voor de Statistiek. Regionale Kerncijfers Nederland [Regional core statistics the Netherlands]. <http://statline.cbs.nl/StatWeb/publication/?DM=SLNL&PA=70072NED&D1=286-288&D2=0,2,10,31,62,84,135&D3=11-13&VW=T>. 26 Aug 2009. Centraal Bureau voor de Statistiek. 28 Aug 2009.
- Centraal Bureau voor de Statistiek. Stijging zorguitgaven vooral door toename volume [Increase healthcare expenditure mainly due to increase in volume]. 3 Jan 2013a and 15 Oct 2013a.
- Centraal Bureau voor de Statistiek. Uitgaven aan zorg met 3,7 procent gestegen [Health expenditure increased with 3.7 percent]. 16 May 2013b and 15 Oct 2013b.
- Damman OC, Hendriks M, Rademakers J, Spreeuwenberg P, Delnoij DMJ, Groenewegen PP. Consumers' interpretation and use of comparative information on the quality of health care: the effect of presentation approaches. *Health Expect*. 2012;15(2):\$32#197–211.
- Dautzenberg M, Weenink J-W, Faber M, Ouwens M. Kiezen borstkankerpatienten voor kwaliteit? [Do breast cancer patients choose for quality]. Nijmegen: IQ Scientific Institute for Quality of Healthcare; 2012.
- Editorial Office Nursing. Verpleegkundig specialist mag medicatie voorschrijven [Specialist nurses allowed to prescribe medicines]. *Nursing, Tijdschrift voor verpleegkundigen*. 2 Nov 2011.
- Houben-van Herten M, Te Riele S. Vrijwillige inzet [Volunteers]. Den Haag/Heerlen: Centraal Bureau voor de Statistiek; 2011.
- KNMG. Aantal geregistreerde specialisten/proefielartsen op 31 december van het jaar [Number of registered medical specialists per December 31]. 11 Jan 2013 and 10 Oct 2013.
- Koiker J, De Klerk M, Ter Berg J, Schothorst Y. Meebetalen aan de zorg. Nederlanders over solidariteit en betaalbaarheid van de zorg [Co-paying in health care. Dutch citizens about solidarity and affordability of health care]. Den Haag: Sociaal Cultureel Planbureau; 2012.
- Ministry of Foreign Affairs. About the Netherlands. 2013. 10 Oct 2013.
- Nederlandse Vereniging van Ziekenhuizen. Gezonde zorg. Brancherapport algemene ziekenhuizen 2012 [Healthy care. Branche report general acute care hospitals 2012]. Utrecht/Den Haag: Nederlandse Vereniging van Ziekenhuizen/SIRM; 2012.
- Nederlandse Zorgautoriteit [Dutch Healthcare Authority]. Tariefbeschikking [Tariff decision]. TB/CU-7023-01; volgnr. 29. 16 Dec 2011.
- Nederlandse Zorgautoriteit [Dutch Healthcare Authority]. Marktscan huisartsenzorg. Weergave van de markt tot en met 2011 [Market scan GP care. Overview of the market up and until 2011]. Utrecht: Nederlandse Zorgautoriteit; 2012.
- Nederlandse Zorgautoriteit [Dutch Healthcare Authority]. Marktscan en beleidsbrief Zorgverzekeringsmarkt. Weergave van de markt 2009–2013 [Market scan and policy brief Health insurance market. Overview of the market 2009–2013]. Utrecht: Nederlandse Zorgautoriteit; 2013.
- Nederlandse Zorgautoriteit [Dutch Healthcare Authority] & DBC-onderhoud. Toelichting op de honorariumberkening DBC-zorgproducten [Explanatory note to the calculation of the tariffs for medical specialists in DBC care products]. Utrecht: Nederlandse Zorgautoriteit; 2012.
- OECD. OECD health data. 2013. 9 Sept 2013.
- Oelen M. Meer verpleegkundigen willen medicatie voorschrijven [More nurses would like to prescribe medicines]. *Nursing, Tijdschrift voor verpleegkundigen*. 25 Apr 2013 and 10 Oct 2013.
- Oudijk D, De Boer A, Woittiez I, Timmermans J, De Klerk M. Mantelzorg uit de doeken [Informal care explained]. Den Haag: Sociaal Cultureel Planbureau; 2010.
- Reitsma M, Brabers A, Masman W, De Jong J. De kiezende burger [The choosing citizen]. Utrecht: NIVEL; 2012.
- Roos AF, Schut FT. Evaluatie aanvullende en collectieve ziektekostenverzekeringen 2009 [Evaluation of VHI and collective health insurances 2009]. Rotterdam: Instituut Beleid en Management Gezondheidszorg (Erasmus MC) Erasmus Universiteit Rotterdam; 2009.
- Schäfer W, Kroneman M, Boerma W, Van den Berg M, Westert G, Devillé W, Van Ginneken E. The Netherlands health system review. *Health Syst Transit*. 2010;12(1):1–228.
- Statistics Netherlands. Gezondheid en zorg in cijfers 2012 [Health and healthcare in figures 2012]. Heerlen: Centraal Bureau voor de Statistiek; 2012.
- Statistics Netherlands. Statline. 2013. 15 Sept 2013.
- Ten Hove M, Van Hilten O, Berger-van Sijl M, Mets-op't Land JM. Zorgthermometer. Verzekerden in beweging [Care monitor. Insured on the move]. Utrecht: Vektis; 2012.
- Van Hassel DTP, Kenens RJ. Cijfers uit de registratie van huisartsen. Peiling 2012 [Figures from the GP registry 2012]. Nivel: Utrecht; 2013.
- World Health Organisation. European health for all database. 2013. World Health Organisation.



William A. Haseltine and Chang Liu

## Contents

<b>Introduction</b> .....	878
<b>Organization and Governance</b> .....	878
Organization and Planning .....	878
Regulation .....	879
Health Information Systems and Technology .....	879
The Role of Patients .....	880
<b>Financing</b> .....	880
Funding .....	880
Coverage and Subsidies .....	880
Sources of Revenue .....	881
Cost Control .....	882
Pooling of Funds and Purchasing .....	884
<b>Physical and Human Resources</b> .....	884
Healthcare Infrastructure .....	884
2012 Singapore Healthcare Professional Workforce .....	884
Workforce Trends .....	884
Paying Healthcare Professionals .....	884
<b>Delivery of Health Services</b> .....	885
Primary Care .....	885
Community Health Assist Scheme .....	885
Care Coordination .....	885
Long-Term Care .....	886
Breakdown of operators for various long-term care services .....	886
Mental Healthcare .....	886
Pharmaceutical Care .....	887
The Private Hospitals .....	887
<b>Reforms</b> .....	887
Main Reforms .....	887
Recent Reforms .....	888
Planned Reforms .....	888

---

W. A. Haseltine (✉) · C. Liu  
ACCESS Health International, New York, NY, USA  
e-mail: [wahaseltine@me.com](mailto:wahaseltine@me.com); [chang.liu@accessh.org](mailto:chang.liu@accessh.org)

<b>Assessment</b> .....	888
User Experience .....	888
Health Outcomes .....	888
Transparency and Accountability .....	889
<b>References</b> .....	890

### Abstract

Singapore is a small island nation located off the southern tip of the Malay Peninsula in Southeast Asia. The country has a population of 5.31 million, of which 3.82 million are citizens and permanent residents. With a land area of 715 square kilometers, Singapore's population density is 7,422 per square kilometer, making it one of the most densely populated sovereign states in the world. Ethnically, the population is overwhelmingly Chinese – almost 75%, followed by Malays at just over 13%, and Indians at 9%.

## Introduction

Among the citizens and permanent residents (i.e., excluding nonresidents in the country), approximately 23% fall under the age of 20, 67% are between 20 and 64, and 10% are 65 or older. The median age is 38.4 (Department of Statistics, Singapore 2013).

Until 1959, the year it achieved internal self-government, Singapore was a colonial outpost of the British Empire. At the time of the British withdrawal, the country was impoverished, with no industrial base or natural resources upon which to build its economic future. After a brief and unsuccessful merger (1963–1965) with Malaysia – its much larger neighbor to the north – Singapore became a fully independent nation under a government controlled by the People's Action Party or PAP.

The People's Action Party has been the majority party ever since, and its longevity in power has provided Singapore with a remarkable era of political stability. This stability has over the years nurtured a consistent political vision, a constancy of purpose and action, and a culture of cooperation among all government ministries. As a result, its policymakers have been able to develop and implement extremely long-range plans that reflect the nation's desire for collective

well-being and social harmony. Stable, astute political leadership and long-term economic policy planning has turned the once impoverished country into an economic powerhouse allowing it to build its world-class healthcare system.

Earliest steps leading up to creation of the system involved improving the general state of public health through proper sanitation, control of infectious diseases, and development of clean water and food supplies. Once satisfied that it had reached its goals, health policy planners began to build the health system's infrastructure, including primary care centers at the community level as well as regional hospitals.

## Organization and Governance

### Organization and Planning

Singapore's Ministry of Health has overall government responsibility for addressing the healthcare needs of the people. Key ongoing activities include: assessment of needs and planning for services and for manpower, governance, and financing.

**Assessing needs:** The ministry makes regular projections of the disease burden and determines whether the current levels of service are sufficient. Service gaps that are detected are prioritized at the national and the regional levels.

**Services planning:** The ministry projects facility requirements for primary care locations, acute and community hospitals, nursing homes, and other services. Local care models are assessed to ensure they remain up to date with the latest medical advances as well as local developments. The ministry is also responsible for planning and developing the systems IT capability.

**Manpower planning:** The ministry projects manpower demand and responds with training and education, attracting talent, and overseas recruitment as necessary to meet demand. It is

also responsible for workforce management including retention and upgrading of skills.

**Governance and financing of the system:** The ministry is also responsible for financing policies and governance, including a performance management system. It also creates feedback mechanisms to drive continual improvement in all areas of responsibility.

## Regulation

The healthcare system is regulated by the Ministry of Health through legislation, regulation, and enforcement. One of its agencies, the Health Sciences Authority, regulates health products, including medicines. Professional bodies, including the Singapore Medical Council, Singapore Dental Council, Singapore Nursing Board, and Singapore Pharmacy Board, self-regulate their healthcare professionals through codes of ethics and conduct, practices, and guidelines.

One of the core regulatory functions of the ministry is the licensing of healthcare institutions under the Private Hospitals and Medical Clinics Act and conducting regular inspections and audits. These institutions provide services that aid in or provide medical diagnosis, treatment, rehabilitation, and management of patients. Laboratory and radiology services are two examples. Public and private hospitals, clinics, laboratories, and nursing homes are required to submit applications to the ministry for the license to operate. Pre-licensing inspections are conducted to ensure standards. Complaints, surveillance, and analysis of advertisements are used to identify potential problems, and they are followed up with compliance audits and possible prosecutions. Marketing by these licensed facilities is also regulated in order to safeguard the public against false or unsubstantiated claims and to prevent inducement to use nonessential services such as aesthetics medicine.

The ministry also works closely with professional bodies such as the Academy of Medicine and the College of Family Physicians and with union-associations such as the Singapore Medical Association as well as industry groups to discuss a wide range of issues such as their

practice, ethics, and standards of care and to consult on policy and operational matters. The ministry also engages them to explain policy rationale and garner their support in implementing various initiatives.

The Health Sciences Authority regulates the manufacture, import, supply, presentation, and advertisement of health products – including medicines, complementary medicines (traditional medicine and health supplements), cosmetic products, medical devices, tobacco products, and medicinal products for clinical trials. Its mission is to ensure that all meet internationally benchmarked standards of safety, quality, and efficacy.

The insurance industry is regulated by the Monetary Authority of Singapore as part of its role as the financial regulatory authority of Singapore. The Ministry of Health regulates the segment of the health insurance market for plans that are paid by Medisave.

## Health Information Systems and Technology

Singapore benefits from an information management system that collects, reports, and analyzes information to aid in the formulation of policy as well as the monitoring of implementation. Sources of information include administrative data and survey-based data, articles, and reports from professional journals and reports and from external organizations.

The Singapore healthcare system is heavily invested in IT infrastructure and in the development of information systems for processing and storing large volumes of data in support of policy research, planning, operations, and monitoring. High-quality data standards, IT security, and audits are utilized to ensure accuracy and reliability of all information collected. In addition, external data is carefully screened to ensure that sources are reputable and trustworthy.

Both public and private healthcare providers are required to report their service statistics to the Ministry of Health, including two types of information: inpatient capacity and utilization, such as number of inpatient beds, beds in service, bed



occupancy rates, inpatient discharges, and average lengths of stay, and surgical procedures, including inpatient and day surgeries, and deliveries.

In addition, public providers are required to report on their polyclinic, specialist outpatient, and emergency department attendances.

## The Role of Patients

The needs of the nation's patients and stakeholders are taken into account through various means. Public consultation takes place before policies are enacted to ensure that public sentiment, concerns, and feedback are added to the discussion; that diverse views, testing, and refinement of ideas take place; and that public understanding and support are cultivated in order to facilitate implementation.

The Ministry of Health conducts an annual patient satisfaction survey for patients of the public sector healthcare institutions. The survey focuses on key service areas such as overall satisfaction and expectations, care coordination, facilities, care and concern shown by medical professionals, as well as their knowledge and skills.

---

## Financing

### Funding

Funding of the system comes from a combination of government subsidies, individual savings, insurance, and other third-party payers, such as employer benefits, etc. The philosophy at the heart of Singapore's system is the requirement that consumers of healthcare must share in the costs of their care. Thus, private expenditure on care (including Medisave, MediShield, and Medisave-approved insurance plans) is high compared to countries with comparable systems – almost 70 (68.6)% of the total national expense of healthcare. As a result, while government subsidies reduce the cost of services provided for those who opt for subsidized care, patients

approach their healthcare choices knowing that they will pay a part of the bill. Still, national saving accounts, insurance programs, and a safety net help to ameliorate the financial burden.

## Coverage and Subsidies

Subsidies flow to and through the healthcare system in this way: government pays subsidies directly to public hospitals, polyclinics, and other healthcare providers reimbursing them for a portion of their costs for treating patients. The funding system is a hybrid mix of block grants and Casemix, a methodology for classifying and describing providers "output." Approximately 70 medical conditions are financed through Casemix.

Hybrid block grants are allocated to public hospitals. A portion of the hospitals' annual budgets are provided as a block grant, with the remainder provided on a piece-rate basis for 70 common conditions based on Diagnosis-Related Groups (DRG). DRG is a system for classifying inpatient and day surgery cases, according to the patients' diagnosis and treatment, into one of more than 600 groupings. Hospitals can reallocate their savings for use in the broad areas that the Ministry of Health has identified, such as teaching and research. The hybrid block budgets are reviewed every 3 to 5 years against the actual workload of the care providers.

Patients receive the benefits of the government system of subsidies in a number of ways, including acute and inpatient care in specific ward classes in the public hospitals, for outpatient care in the public polyclinics as well as the specialist outpatient clinics at public hospitals, and emergency care at all public hospitals. Eligible low- and middle-income patients may also receive subsidies for intermediate- and long-term care at facilities managed by voluntary welfare and private organizations, outpatient treatment for chronic and or acute conditions, and also certain dental procedures, at *private sector* primary care providers.

Subsidies are closely linked to the ward classes in Singapore's public hospitals, which range from private rooms to dormitory-style accommodations

with a corresponding range of amenities, choices, and prices but access to the same doctors and assurance of the same quality of care. There are four classes: A, B1, B2, and C. A is the most costly, with C the least costly. A-level wards contain private rooms with bath, air conditioning, and access to private doctors of the patient's choice. C patients are in open wards, with eight or nine patients in a room, sharing a bath, and usually without air conditioning. Doctors are assigned to these patients.

As amenities increase, subsidies decrease. Patients in the A wards receive no subsidy, while C-ward patients receive subsidies of up to 80% – depending on their income – of their ward charges, drugs, and medical treatment. C-ward patients also receive subsidies on surgical procedures and physicians' fees. In the wards between A and C, subsidies increase as amenities and choices decrease.

Class ward	Subsidy level
A	0%
B1	20%
B2	65–50% <sup>a</sup>
C	80–65% <sup>a</sup>

<sup>a</sup>Financial means testing determines eligibility for subsidy for patients in C and B2 wards

Means testing in public hospitals as of 1 January 2009		
Average monthly income of patient (SGD) <sup>a</sup>	Citizens subsidy <sup>d</sup>	
	Class C ward(%)	Class B2 ward(%)
\$3,200 and below <sup>b</sup>	80	65
\$3,201–\$3,350	79	64
\$3,351–\$3,500	78	63
\$3,501–\$3,650	77	62
\$3,651–\$3,800	76	61
\$3,801–\$3,950	75	60
\$3,951–\$4,100	74	59
\$4,101–\$4,250	73	58
\$4,251–\$4,400	72	57
\$4,401–\$4,550	71	56
\$4,551–\$4,700	70	55
\$4,701–\$4,850	69	54
\$4,851–\$5,000	68	53
\$5,001–\$5,100	67	52
\$5,101–\$5,200	66	51

(continued)

Means testing in public hospitals as of 1 January 2009

Average monthly income of patient (SGD) <sup>a</sup>	Citizens subsidy <sup>d</sup>	
	Class C ward(%)	Class B2 ward(%)
\$5,201 and above <sup>c</sup>	65	50

<sup>a</sup>Monthly income is defined as average monthly wage based on last available 12 month data (including bonuses)

<sup>b</sup>No income and property with annual value (estimated value of a property if it were rented out) \$13,000 and below

<sup>c</sup>No income and property with annual value exceeding \$13,000

<sup>d</sup>Subsidies for Singapore permanent residents in most income bands will receive half the corresponding subsidy that citizens receive (Ministry of Health, Singapore)

Patients do have a choice in the matter of ward classes. Individuals with high incomes can choose the C ward, but their subsidy would be much lower than what a low-income individual receives. Conversely, low-income patients can choose to stay in a class A ward if they can pay for it.

## Sources of Revenue

### Government Healthcare Budget

Funding of the healthcare system takes place through the Ministry of Health. The ministry's budget for fiscal year 2013 is \$5.7 billion. The ministry's budget is used for healthcare subsidies, promoting good health practices in the population, developing manpower, training of healthcare professionals, and infrastructure. A total of \$4 billion is allocated for subsidies to Singaporeans receiving medical care at the public hospitals, polyclinics, community hospitals, and institutions providing intermediate and long-term care. A sampling of other budget allocations include: \$177 million for initiatives addressing obesity prevention, tobacco control, childhood preventive health services, chronic disease management, and public education and \$70 million for Medisave grants to newborn Singapore citizens (Ministry of Health, Singapore 2013c).

### Private Expenditure on Healthcare

The other major source of funding for the system is private financing and expenditure on healthcare. Singaporeans pay co-payments and deductibles that are often higher than in other nations.

According to the World Health Organization, private expenditure amounts to almost 70 (68.6)% of the nation's total expenditure on care. This statistic reflects the government's guiding philosophy that healthcare is not free and, as stated earlier, that consumers of care must pay a portion of the cost their care. Of the private expenditure, 74.2% represent out-of-pocket expenditure versus 8% from Medisave and 6% from MediShield and Integrated Shield Plans (World Health Organization 2013; Ministry of Health, Singapore).

At the heart of Singapore's system of private financing and expenditure are mandated savings and insurance programs that help consumers pay for care. They are known as the "3Ms" – Medisave, MediShield, and Medifund. They play a critical role in maintaining the health and welfare of Singapore's people and the success of the healthcare system itself. The most critical component of the trio is *Medisave*, a mandatory, individual medical savings account to which workers contribute a percentage of their wages which employers match. Medisave grew out of the nation's Central Provident Fund, a mandatory savings program originally created by the British during their rule of Singapore to help workers pay for their retirement. Contributions to the accounts are tax exempt, as are withdrawals. The account is used to pay for health services and health insurance for the account's owner as well as for family members.

*MediShield*, the second of the 3Ms, is a low-cost insurance program paid for by the insured for coverage against catastrophic inpatient bills and selected outpatient care. MediShield premiums can be paid for from the individual's Medisave account. Singaporeans are automatically enrolled in the program but are able to opt out if they so desire. Soon to be introduced is an extension of this program called Medishield Life which will cover all Singaporeans.

Private health insurance is also available. While affordable, the plans also include deductibles and co-payments in accordance with the healthcare systems requirement that consumers of care must contribute to the cost of their care. Catastrophic insurance is widely held and covers partial costs of

expensive or long-term treatment. Insured patients must usually pay 20% of the cost of such care.

Private, Medisave-approved insurance, called Integrated Shield Plans, are meshed together with MediShield to form an integrated plan for users. Such private plans give patients additional benefits and coverage for paying the costs of private hospitals or Class A and B1 wards in the public hospitals. Policyholders keep the benefits and coverage afforded then by their basic MediShield plans. In addition, Medisave can be used to pay the premiums of the approved, private plans, subject to a limit. Like MediShield, they also include deductibles and co-payments in accordance with the healthcare systems requirement that consumers of care must contribute to the cost of their care. Catastrophic insurance is widely held and covers partial costs of expensive or long-term treatment. Insured patients must usually pay 10–20% of the cost of such care (Ministry of Health, Singapore).

*Medifund*, the third "M" is an endowment program funded by the government as a healthcare safety net that aids the poor pay in paying for their care. Medifund was set up in 1993 to assist Singaporeans who could not pay their medical bills. Needy citizens can apply for assistance and are means tested before their applications are approved.

In addition to the 3Ms, another program, labeled ElderShield was introduced in 2002 to provide insurance coverage for the costs of long-term care necessitated by very serious disabilities in the elderly. ElderShield is an opt-out program that commences for individuals when they turn 40 years of age. The insurance is offered by private insurers only, who are selected through competitive bidding that takes place every 5 years. Premiums are fixed at a flat rate based on the age of the individual joining the program and are paid by the insured until age 65. Benefits are set at fixed monthly payouts of \$400 per month.

## Cost Control

Singapore is a leader in keeping costs under control, and it does so while providing world-

class healthcare. The nation spends 4.5% of GDP on care versus, for example, 17.9% of GDP in the United States and 9.3% in the United Kingdom. Here are some examples of private and public spending on healthcare for several nations. All data as of 2010.

	Singapore	United States	India	China
Total expenditure on health as % of GDP	4.5	17.6	3.7	5
General government expenditure on health as % of total expenditure on health	31.4	48.2	28.2	54.3
Private expenditure on health as % of total expenditure on health (World Health Organization 2013)	68.6	51.8	71.8	45.7

Singapore controls the costs of healthcare in a number of ways, perhaps first and foremost in the manner by which it both fosters and controls competition. The nation approaches healthcare as a quasi-capitalist market. Amid concerns in the early 1990s of soaring health costs, the government issued a white paper entitled “Affordable Health Care” that, among other issues, set the goal of engaging competition and market forces to improve service and raise efficiency. It was established that government would intervene directly in the healthcare sector when the market failed to keep costs down. This became the guiding policy of the system. Public and private hospitals exist side by side in this market, with the public sector having the advantage of patient incentives and subsidies. Because it can regulate the number of public hospitals and beds, the government is able to shape the environment of the marketplace. Within that environment, market forces regulate the private sector, which must be careful to not price itself out of the

market. At the same time, government sets subsidy and cost-recovery targets for each ward class, which indirectly keeps the public sector hospitals from producing excess profits. Hospitals are also given annual budgets for patient subsidies, so they can plan accordingly, knowing in advance the levels of reimbursement they will receive for patient care. They are required to break even within this budget. The entire system functions successfully because the quality of care in the public hospitals is extremely high and is scrupulously maintained.

Singapore also regulates the number of medical students studying in the country, as well as the number of foreign medical schools’ degrees recognized in the country. In this way, the number of practicing physicians is controlled, preventing an oversupply of medical services and avoiding induced demand. The medical savings programs, the insurance programs, and the subsidies to public hospitals are continually adjusted. The numbers of beds in the public hospitals are carefully controlled. Government regulates and limits the private insurance programs available to Singaporeans. Wages of doctors in the public sector are kept reasonable and not sky-high and are periodically reviewed with the goal of keeping them competitive with the private sector.

The private sector operates and thrives in this quasi-capitalist environment, serving patients who wish to pay more for certain services or amenities and competing with public sector facilities on price and quality.

### Price Transparency

Another factor controlling costs is price and outcome transparency. The Ministry of Health makes available on its website the hospital bills for common illnesses (arranged alphabetically from anemia to urinary stone), treatments, and ward classes: [http://www.moh.gov.sg/content/moh\\_web/home/costs\\_and\\_financing/HospitalBillSize.html](http://www.moh.gov.sg/content/moh_web/home/costs_and_financing/HospitalBillSize.html).

Patients can look up the costs of specific surgeries, the number of cases treated in each hospital, tests, and more. The data is complete for public sector hospitals while private hospital data is voluntary and may not carry the detail of the public sector information. Armed with pricing

information, consumers of care can better shop for the services they require.

### Pooling of Funds and Purchasing

Currently, there is no framework to pool funds to purchase provider services and goods, although a system does exist that aggregates demand for bulk purchasing pricing. The Group Purchasing Office (GPO Pharma) consolidates drug purchases at national level. One goal of this system is to keep drug prices affordable for the elderly and lower-income groups and contain the costs of pharmaceutical-related expenditure. GPO also purchases medical supplies, equipment, and IT services for the healthcare system.

### Physical and Human Resources

#### Healthcare Infrastructure

The data below provide a clear snapshot of the main components of Singapore’s healthcare infrastructure as of December, 2012:

- Number of public acute hospitals (beds): 7 (6,985)
- Number of public specialty centers (beds): 8
- Number of private acute hospitals (beds): 9 (1,555)
- Number of private other hospitals (beds): 1 (20)
- Number of public polyclinics: 18
- Number of private medical clinics for primary care: about 2,400
- Number of community health centers: 2
- Number of nursing homes: 66
- Number of hospices: 4

#### 2012 Singapore Healthcare Professional Workforce

	Public	Private	Total (in active practice)
<b>Total no. of doctors</b>	6,131	3,515	9,646
<b>Specialists</b>	2,342	1,293	3,635

(continued)

	Public	Private	Total (in active practice)
<b>Nurses</b>	20,911	8,348	29,259
<b>Midwives</b>	89	65	154
<b>Dentists</b>	357	1,215	1,572
<b>Optometrists/opticians</b>	155	2,124	2,279
<b>Pharmacists</b>	934	1,048	1,982

Information on the number of occupational therapists, psychologists, and medical lab technicians are not available at this time (Ministry of Health, Singapore).

#### Workforce Trends

Anticipating growth in demand, Singapore will expand its healthcare professional workforce by 20,000 by the year 2020. This increase covers doctors, nurses, dentists, pharmacists, and allied health professionals, representing a 50% increase from 2011. The nation is also expanding training pipeline, encouraging mid-career professionals to join the healthcare sector, and supporting older healthcare staff who wish to continue working for as long as they can.

Singapore is also looking to greater use of technology, such as tele-consultations, and equipment such as patient mobility aids to raise the productivity levels of its professional workforce.

#### Paying Healthcare Professionals

The data below show the gross monthly wage of healthcare professionals in Singapore in 2012

	2012 Gross monthly wage <sup>a</sup>		
	25th percentile (\$)	Median <sup>b</sup> (\$)	75th percentile (\$)
Healthcare professionals			
(Primary care doctors):	9,058	11,398	16,358
General practitioners/physicians			
Specialist doctors:			
Specialist medical	9,919	20,516	30,300

(continued)

Healthcare professionals	2012 Gross monthly wage <sup>a</sup>		
	25th percentile (\$)	Median <sup>b</sup> (\$)	75th percentile (\$)
practitioners (medical)			
Specialist medical practitioners (surgical)	15,610	21,595	26,516
Nurses: (registered nurses)	2,583	3,061	3,837
Pharmacists	3,743	4,387	5,574
Psychologists	2,841	3,537	4,498
Occupational therapists	2,880	3,139	3,723
Medical technicians: Medical and Pathology Lab. Tech	2,700	3,268	4,492

<sup>a</sup>Data on monthly gross wages was collected from the Occupational Wage Survey, 2012. Monthly gross wage refers to the sum of the basic wage, overtime payments, commissions, allowances, and other regular cash payments. It is before deduction of employee CPF contributions and personal income tax and excludes employer CPF contributions, bonuses, stock options, other lump-sum payments, and payments in kind. Detailed information on the Survey Coverage and Methodology is available online

<sup>b</sup>Median wage refers to the wage level at the middle of the wage distribution which divides the bottom half of wage earners from the upper half (Ministry of Manpower, Singapore 2012)

## Delivery of Health Services

### Primary Care

Primary care is provided mainly by private rather than public providers. There are 2,400 private medical clinics offering primary care. Many belong to private general practitioner chains, and they are located throughout the neighborhoods of Singapore. There are 18 public polyclinics – multi-doctor facilities providing outpatient care, immunization services, health screening, pharmacy services, and more. Some also offer dental services. The polyclinics are meant to serve lower-income patients who might not be able to afford the fees of the private clinics.

Specialized ambulatory surgical services are provided at the Singapore General Hospital and National University Hospital.

### Community Health Assist Scheme

Singapore's public healthcare sector has been strengthening its ties with the large networks of private general practitioners. They are being enlisted into the Community Health Assist Scheme, a program that provides basic care, treats certain chronic illnesses, and offers dental care. Lower- and middle-income patients can receive subsidized outpatient services, including dental services, at the private clinics just as they would at a government polyclinic. The program also covers treatment of common chronic diseases.

### Care Coordination

Singapore's Agency for Integrated Care (AIC) was set up under MOH Holdings in 2009 and operates at the patient, provider, and system levels for the benefit of patients and families. The Agency provides hospitals with teams – called Aged Care Transition Teams or ACTION Teams – that coordinate discharge planning and facilitate the transition of elderly patients from hospitals to the intermediate and long-term care sector. It is also the central referral agency for aged care services in the community, such as nursing homes, community hospitals, day rehabilitation, day dementia care, homecare services, and home hospice services. Another key function of the Agency is working with primary and intermediate and long-term care providers to expand service capacity and improve healthcare capabilities. The Agency also supports caregivers through "AIC @ City Square Mall," a program that provides information on community support resources and referral services for health and social care issues. In addition, the Agency administers financial assistance schemes such as the Caregivers Training Grant and the Foreign Domestic Worker Grant, to help families offset the costs of hiring and training foreign domestic workers.

### Long-Term Care

Long-term care services for the elderly are managed by voluntary welfare organizations or by private operators. These include both residential and nonresidential care options. Government subsidies are available for seniors who utilize these services, subject to a means test.

Residential facilities cater to the convalescent sick or elderly individuals who do not require hospital care but are too ill or frail to care for themselves or to be cared for at home. Examples of residential care facilities include nursing homes and inpatient hospices. Respite care services are also available to caregivers.

Nonresidential services such as home and community-based care are also available to support the elderly. Home care services involve the care staff visiting the homes of the homebound elderly to provide medical, nursing, social care, and/or palliative care services. There are also eldercare services, such as maintenance day care, day rehabilitation, and dementia day care, provided within centers in the community. The elderly attend these centers during the day but go back to their homes in the evening. Such nonresidential services are important in providing alternative care options to institutionalization and facilitate seniors to age gracefully in the community.

Below is a listing of the various long-term care services available in Singapore.

#### Breakdown of operators for various long-term care services

Facility		Total no.	VWO-run	Private operators
Residential facilities	Nursing homes	66	32	34
	Inpatient hospices	4	4	0
Nonresidential facilities	Eldercare day centers	69	63	6
	Home palliative care providers	9	8	1

(continued)

Facility		Total no.	VWO-run	Private operators
	Home care providers (Including home healthcare and social services)	38	20	18

**Note:** All figures are dated as of Dec 2013 except for the number of Eldercare day centers and nursing homes, where figures are as of current (Ministry of Health, Singapore)

### Mental Healthcare

The National Mental Health Blueprint, formulated in 2007, guides Singapore agencies in providing mental health services, including active mental health education and prevention as well as early detection and treatment for people at risk or facing emotional difficulties. The Community Mental Health Master Plan, developed in 2012, lays the groundwork for building a network of care and supporting systems to enable integrated community living.

In addition, resources and workshops developed by Singapore’s Health Promotion Board promotes mental well-being. Programs are targeted at young and the old as well as their family members/caregivers.

Singapore has one acute tertiary psychiatric hospital – the Institute of Mental Health. Services offered there include psychiatric, rehabilitative, and counseling services for children, adolescents, adults, and the elderly, long-term care, and forensic services. The Institute also houses the National Addictions Management Services to treat patients with addictions.

Psychiatric services are also embedded in the other public hospitals, which offer general as well as more specialized services such as eating, sleep, addiction disorders, and geriatric psychiatry.

### Community Care

As of this writing, Singapore is rolling out a series of community-based mental health services to complement those offered in tertiary mental health facilities. Balanced development of tertiary

and community-based services has been shown to improve health and social outcomes while reducing system cost. Components of the community care program include: multidisciplinary shared care teams that provide treatment and care to the mentally ill through service networks in the community, support for caregivers to cope with care giving, and community safety network for people with dementia and depression and their caregivers. There are also community-based, targeted mental health programs for youths, adults, and the elderly.

### **Psychiatric Intermediate and Long-Term Care**

The majority of psychiatric long-term care services, where individuals require residential care or a period of transition and close supervision after discharge, are provided by the Institute of Mental Health and voluntary welfare organizations – supported by Ministry of Health and Ministry of Social and Family Development – such as Singapore Association for Mental Health and Singapore Anglican Community Services. Types of long-term care facilities include psychiatric nursing homes, rehabilitation homes, and day care centers.

### **Pharmaceutical Care**

In Singapore, pharmacists are now involved in providing more direct patient care as members of multidisciplinary healthcare teams. In the public sector, pharmacy services and pharmaceutical care by pharmacists are provided through the Departments of Pharmacy at each public hospital/institution.

Pharmacists dispense and review medications, conduct medication counseling to patients upon discharge, and perform specialized clinical pharmacy services in hospitals, such as a dedicated ICU pharmacist.

In the outpatient and community setting, pharmacists also undertake health management and disease prevention counseling, provide patient medication management and adherence services

as well as run specialized pharmacy clinics, such as an anticoagulation clinic.

In the intermediate- and long-term care setting as well as in the home, programs have been introduced where pharmacists visit nursing homes and aid in managing residents' medication needs more effectively. With the Pharmacist Outreach Program, pharmacists visit the homes of referred patients to check medication compliance and identify and address drug-related problems in consultation with the primary physician.

Pharmacists are also involved in supply of medicines and medication safety, at the institutional level through reviewing drug formularies and monitoring the use of drugs. Pharmacists are also involved in medication safety initiatives at the institutional or national level, medication error reporting and monitoring frameworks, monitoring and reporting of adverse drug events.

### **The Private Hospitals**

Private hospitals account for approximately 20% of inpatient beds. Patients may use either the public or private system, as long as they can pay the costs of their preferred provider. Luxury amenities are available in some of the private hospitals. Private hospitals are also more involved in medical tourism than are the public facilities. Parkway Pantai is the main private hospital group in Singapore.

There is a trend toward tapping private hospitals' spare capacity for treating public system, subsidized patients. Private hospitals' bed occupancy rate averages about 55% (MOH 2012 Committee of Supply Speech).

---

## **Reforms**

### **Main Reforms**

Several main reforms in the Singapore system are aimed at making healthcare more affordable for consumers.

The Community Health Assist Scheme, which provides subsidized healthcare services at private (as



opposed to public) general practitioner clinics, will no longer have age restrictions, opening up subsidized medical and dental care at over 900 private clinics for lower- and middle-income Singaporeans.

Currently, Medisave can be used for treatment of ten chronic diseases in the outpatient setting. The government is also expanding Medisave use for five more chronic conditions – osteoarthritis, benign prostatic hyperplasia, anxiety, Parkinson’s disease, and nephritis/nephrosis (chronic kidney disease) – and bringing the total number up to 15. These will also be subsidized through the Community Health Assist Scheme, again, giving patients the opportunity to be treated at the private clinics.

High-risk groups will also benefit from expanded Medisave use for pneumococcal and influenza vaccinations. Over the years, Medisave use has been expanded gradually to cover chronic conditions such as diabetes and high blood pressure as well as health screenings and vaccinations for selected groups. The Medisave Contribution Ceiling was increased in 2016, and there is no longer a Medisave Minimum Sum.

## Recent Reforms

### MediShield Life

MediShield Life is the recent reform and transformation of the national health insurance program. The reform initiated in November 2015 aim to address the growing need for chronic disease care and long-term care. Coverage is now universal and compulsory and includes individuals with preexisting conditions. Previously ending at age 90, coverage is now for life. The lifetime cap on benefits has been removed and the annual limit increased to SGD100,000.

Another recent change provides better protection from large hospital bills, by reducing co-insurance payments below 10 percent for the portion of the bill exceeding SGD5,000. Less than 1 percent of Singaporeans will need to pay additional premiums.

## Planned Reforms

### Better Care for The Aged

In 2015, the Ministerial Committee on Ageing unveiled new features of an SGD3 billion national plan to help Singaporeans age with confidence, lead active lives, and maintain strong bonds with family and community. The plan encompasses about 60 initiatives covering 12 areas: health and wellness, learning, volunteerism, employment, housing, transport, public spaces, respect and social inclusion, retirement adequacy, health care and aged care, protection for vulnerable seniors, and research.

Patient costs at specialist outpatient clinics in public hospitals will be lowered through increased subsidies for lower- and middle-income groups. As of this writing, complete details of the plan have not been announced.

---

## Assessment

### User Experience

Singapore’s Ministry of Health conducts an annual patient satisfaction survey that helps it understand patients’ levels of satisfaction and expectations for the public healthcare institutions. The survey includes patient satisfaction for certain service attributes such as waiting times, facilities, and care coordination.

The results of the 2012 survey showed that 77.1% of respondents indicated they were satisfied; 77.7% of patients would “strongly recommend” or “likely recommend” the healthcare institutions to others based on their own experience.

### Health Outcomes

Singapore’s healthcare system delivers very high-quality care with outcomes that are usually better than those found in most high-income countries. It is ranked sixth globally by the World Health Organization – far ahead of the United States at number 37 and the United Kingdom at 18.

- Life expectancy for women – currently 84.5 years versus 65 years in 1960
- Life expectancy for men – currently 79.9 years versus 61.2 years in 1960

Singapore also has a vastly improved survival rate among newborns and infants, a rate better than most developed countries:

- Neonatal mortality rate per 1,000 births is now 1.1 versus 17.7 in the 1960.
- Infant mortality rate per 1,000 births is now 1.8 versus 34.9 in 1960.

Other outcomes:

- Under 5 mortality rate (per 1,000 live births – both sexes) is 2.8 versus 7.5 in 1990

Source: Singapore registry of births and death report (2012).

In addition, its cancer survival rates are similar to Europe's, and its cardiovascular disease death rate is half that of the rest of the Asia/Pacific region.

### Efficiency

Singapore uses a performance measurement and management process to help healthcare providers assess and benchmark their performance against their peers. The National Health System Scorecard uses internationally established performance indicators to compare performance in Singapore. The Public Acute Hospital Scorecard is used to measure institutional-level performance. Its indicators cover clinical quality and patient perspectives. Similar scorecards for providers are being rolled out in primary care facilities and in community hospitals.

The scorecards lay out the standards of service and key deliverables required of the public healthcare institutions, and they are monitored to ensure compliance. They incorporate internationally accepted indicators and definitions where possible, such as the Centers for Medicare & Medicaid Services Joint Commission-aligned measures for acute myocardial infarction and

stroke. The inclusion of these evidence-based and validated indicators allow for comprehensive benchmarking, enabling identification of areas of strong performance as well as areas where improvements are needed.

In 2008 Singapore introduced a set of National Standards for Healthcare which is used to set priorities for improvement efforts and alignment with planning initiatives. It focuses on key areas of concern and promotes a culture of continuous quality improvement.

National Standards for Healthcare is implemented through a network of Healthcare Performance Offices each chaired by a senior clinical leader who reports directly to the institutions chief executive officer/chairman medical board. Resulting quality improvement outputs can then be incorporated into the National Health System Scorecard and the Public Acute Hospital Scorecard for performance analysis and monitoring.

### Transparency and Accountability

Regarding policy development and implementation, Singapore's Ministry of Health uses public consultation with stakeholders and the public before policies are enacted. Stakeholders are engaged through dialogue and the public through public consultation. A set of principles and processes guide the public consultation ensuring that public sentiment, concerns, feedback, and diverse views are taken into account.

The Ministry of Health also gathers data on consumer needs and determines actionable insights that might improve healthcare policies. It also engages in extensive face-to-face conversations through visits to private and public sector institutions, town halls, and feedback sessions. The ministry also identifies potential issues and concerns from the complaint and appeal letters it receives from customers or their Members of Parliament. Quarterly Customer Feedback reports are brought to senior management meetings for discussion. The corporate planning cycle incorporates the review of customer feedback as a key process to guarantee policy responsiveness.

Some concrete actions taken as a result of public consultation include: extension of Medisave use for pneumococcal vaccination, treatment of schizophrenia and major depression, and expanded coverage for major chronic diseases; raised withdrawal limits for community hospital stays and day rehabilitation center visits; and Medisave use for mammograms and colonoscopies. Directly as a result of customer feedback, Medisave withdrawals were also extended to palliative care, including palliative care in the home.

## References

- Department of Statistics, Singapore. [http://www.singstat.gov.sg/statistics/latest\\_data.html#8/](http://www.singstat.gov.sg/statistics/latest_data.html#8/) [http://www.singstat.gov.sg/publications/publications\\_and\\_papers/cop2010/cop2010adr.html](http://www.singstat.gov.sg/publications/publications_and_papers/cop2010/cop2010adr.html). Accessed Oct 2013.
- Yong Gan Kim. *Straitsimes.com* <http://www.straitstimes.com/mnt/html/parliament/mar6-GanKimYong-pt1.pdf>
- Ministry of Health, Singapore. MOH (APO). 2012.
- Ministry of Health, Singapore. 2013. [http://www.moh.gov.sg/content/dam/moh\\_web/Publications/Educational%20Resources/2009/MT%20pamphlet%20%28English%29.pdf](http://www.moh.gov.sg/content/dam/moh_web/Publications/Educational%20Resources/2009/MT%20pamphlet%20%28English%29.pdf). Accessed Oct 2013.
- Ministry of Health, Singapore. 2013. [http://www.moh.gov.sg/content/moh\\_web/home/costs\\_and\\_financing/schemes\\_subsidies/Medishield/Medisave-approved\\_Insurance.html](http://www.moh.gov.sg/content/moh_web/home/costs_and_financing/schemes_subsidies/Medishield/Medisave-approved_Insurance.html). Accessed Oct 2013.
- Ministry of Health, Singapore. Expenditure overview. 2013c. [http://www.singaporebudget.gov.sg/budget\\_2013/expenditure\\_overview/moh.html](http://www.singaporebudget.gov.sg/budget_2013/expenditure_overview/moh.html). Accessed Oct 2013.
- Ministry of Health. All Singapore residents to enjoy universal coverage under MediShield Life, with no exclusions. 2015a. [https://www.google.com.sg/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwikkpWSzPDRAhVJM08KHWjXAhQFggMAA&url=https%3A%2F%2Fwww.moh.gov.sg%2Fcontent%2Fmoh\\_web%2Fhome%2FpressRoom%2FpressRoomItemRelease%2F2015%2Fall-singapore-residents-to-enjoy-universal-coverage-under-medish0.html&usg=AFQjCNFJfEHv-OmUHPEjw-RiELml8a1PFw](https://www.google.com.sg/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwikkpWSzPDRAhVJM08KHWjXAhQFggMAA&url=https%3A%2F%2Fwww.moh.gov.sg%2Fcontent%2Fmoh_web%2Fhome%2FpressRoom%2FpressRoomItemRelease%2F2015%2Fall-singapore-residents-to-enjoy-universal-coverage-under-medish0.html&usg=AFQjCNFJfEHv-OmUHPEjw-RiELml8a1PFw)
- Ministry of Health. Better health, better future for all. Ministry of Health Initiatives for 2015. 2015. [https://www.google.com.sg/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwiq4P-GzPDRAhXCry8KHZr9BQYQFggMAA&url=https%3A%2F%2Fwww.moh.gov.sg%2Fcontent%2Fdam%2Fmoh\\_web%2FpressRoom%2Fresources%2FMOH%2520Factsheet.pdf&usg=AFQjCNFMoyT0VGwMdxLjebX-UVSJs5EOg](https://www.google.com.sg/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwiq4P-GzPDRAhXCry8KHZr9BQYQFggMAA&url=https%3A%2F%2Fwww.moh.gov.sg%2Fcontent%2Fdam%2Fmoh_web%2FpressRoom%2Fresources%2FMOH%2520Factsheet.pdf&usg=AFQjCNFMoyT0VGwMdxLjebX-UVSJs5EOg)
- Ministry of Manpower, Singapore. Median, 25th and 75th percentile of monthly gross wages of Common Occupations in Health Industry. 2012. Ministry of Manpower (Table 4.9, <http://stats.mom.gov.sg/Pages/Occupational-Wages-Tables-2012.aspx#>)
- Weizhen T. Today Online. Thursday Oct 17 2013. <http://www.todayonline.com/singapore/medishield-life-more-sustainable-private-medical-schemes-gan>
- World Health Organization. World health statistics 2013. [http://www.who.int/gho/publications/world\\_health\\_statistics/2013/en/](http://www.who.int/gho/publications/world_health_statistics/2013/en/). Accessed Oct 2013.



# Health System in the USA

# 40

Andrew J. Barnes, Lynn Y. Unruh, Pauline Rosenau, and  
Thomas Rice

## Contents

<b>Introduction</b> .....	893
<b>Organization and Governance</b> .....	893
Public and Private Organizations .....	893
<b>Financing of Major Insurance Programs</b> .....	896
Coverage .....	896
Sources of Revenue .....	896
Financing and Financial Flows .....	896
Medicare .....	898
Medicaid .....	899
Private Insurance .....	901
<b>Physical and Human Resources</b> .....	903
Physical Resources .....	903
Human Resources .....	905

---

A. J. Barnes (✉)  
Department of Health Behavior and Policy, School of  
Medicine, Virginia Commonwealth University, Richmond,  
VA, USA  
e-mail: [andrew.barnes@vcuhealth.org](mailto:andrew.barnes@vcuhealth.org)

L. Y. Unruh  
Department of Health Management and Informatics,  
College of Health and Public Affairs, University of Central  
Florida, Orlando, FL, USA  
e-mail: [lynn.unruh@ucf.edu](mailto:lynn.unruh@ucf.edu)

P. Rosenau  
Division of Management, Policy and Community Health,  
School of Public Health, University of Texas Health  
Science Center at Houston, Houston, TX, USA  
e-mail: [pauline.rosenau@uth.tmc.edu](mailto:pauline.rosenau@uth.tmc.edu)

T. Rice  
Department of Health Policy and Management, Fielding  
School of Public Health, University of California, Los  
Angeles, CA, USA  
e-mail: [trice@ucla.edu](mailto:trice@ucla.edu)

<b>Provision of Health-Care Services</b> .....	907
Public Health .....	907
Outpatient Services .....	907
Acute Inpatient Care .....	908
Mental Health Care .....	909
Pharmaceutical Care .....	909
Long-Term Care .....	910
Palliative Care .....	910
<b>Reforms</b> .....	910
<b>Assessment</b> .....	912
Overview .....	912
Access .....	912
US Data .....	913
International Comparisons .....	913
Outcomes and Quality .....	914
Expenditures .....	918
<b>Conclusions</b> .....	921
<b>References</b> .....	922

### Abstract

This analysis of the US health system reviews its organization and governance, health financing, health-care provision, health reforms, and health system performance. The US health system has both considerable strengths and notable weaknesses. It has a large and well-trained health workforce, a wide range of high-quality medical specialists as well as secondary and tertiary institutions, and a robust health research program and, for selected services, has among the best medical outcomes in the world. But it also suffers from incomplete coverage of its citizenry, health expenditure levels per person far exceeding all other countries, poor health indicators on many objective and subjective measures of quality and outcomes, an unequal distribution of resources and outcomes across the country and among different population groups, and lagging efforts to introduce health information technology. It is difficult to determine the extent to which deficiencies are health system related, though it seems that at least some of the problems are a result of poor access to care. Because of the adoption of the Affordable Care Act (ACA) in 2010, the USA is facing a period of enormous potential change. The major provisions of the ACA were

implemented in 2014, although judicial setbacks, delays, and legislative repeals to its core provisions have reduced its overall impact. Improving coverage was a central aim, envisaged through mandates that certain individuals purchase, and employers offer, private health insurance as well as subsidies for lower-income uninsured citizens to purchase private insurance. However, in late 2017, the individual mandate to purchase insurance was repealed by Congress, with an effective date of January 2019. Eligibility for Medicaid, which provides public coverage for low-income individuals and families, is also expanded, and greater protections for insured persons have been instituted. Furthermore, primary care and public health are receiving increased funding, and improving quality and controlling expenditures are addressed through a range of policies. Early assessments of the ACA suggest coverage rates have expanded, particularly for low-income adults in some states. Whether the ACA will be effective in addressing the US health-care system's historic challenges can only be determined over time.

The material used in this chapter was adapted or taken directly from our book on the US health-care system – Rice T, Rosenau P,

Unruh LY, Barnes AJ, Saltman RB, van Ginneken E, *Health Syst Transit* 15(3):1–431, 2013.

## Introduction

The US is a large, wealthy country, with double the gross domestic product of any other in the world. It is a federal, constitutional democracy, with decision-making authority divided between the federal and state governments. In 2016 nearly one-fifth (17.9%) of its economy was spent on health care (\$3.3 trillion), amounting to \$10,348 per capita (Hartman et al. 2017). As with many such national averages in this report, there are wide variations across the states, with spending per capita in 2014 ranging from about \$5,982 per person in Utah to more than \$11,944 in the District of Columbia (Kaiser Family Foundation 2014a). Tax rates are lower than in almost all other high-income countries, consistent with the fact that its public sector provides fewer social services. Tax rates are lower than in almost all other high-income countries, consistent with the fact that its public sector provides fewer social services. Despite being a high-income nation, the US ranks poorly, compared to other high-income countries, on measures of income equality. Because the US birth rate is higher than that of most developed countries, its dependency ratio – those too young or too old to work, divided by the working age population – is expected to grow more slowly than in most other countries.

The racial and ethnic makeup of the US population is quite varied, with approximately 61.3% non-Hispanic White, 17.8% Hispanic or Latino, 13.3% non-Hispanic Black or African American, and the remainder other and/or mixed racial and ethnic groups (US Census Bureau 2017). Hispanics and Latinos are the fastest-growing group, with a 49% population increase between 2000 and 2010, compared to just 5% for others (Ennis et al. 2011). This proportional relationship also continues to change: Asians have replaced Hispanics and Latinos to be the fastest-growing group, with a total population of 21 million as of 2015, representing a 3.4 % increase compared

with 2014 (U.S. Census Bureau 2016). Moreover, in California, there are now almost twice as many Hispanics and Latinos age 18 and younger than there are whites (Kidsdata.org 2015).

Historically, the US has resisted central planning or control at both the federal and state levels. The US health-care system reflects this wider context, having developed largely through the private sector and combining high levels of spending with distinctively low levels of government regulation. The US spends far more money on health care per person than any other country.

International comparison shows a varied picture with respect to access to health care, health behaviors, and outcomes. The US is unusual among high-income OECD countries in that most Americans still receive their coverage from private health insurance, and more than 12% of non-elderly adults are uninsured, although this proportion has been reduced significantly through implementation of the Affordable Care Act (Kaiser Family Foundation, 2016a). With regard to health behaviors, the picture is again varied; the USA has been notably effective in reducing smoking rates and has one of the lowest smoking rates internationally. But it has been less effective in grappling with nutritional health and obesity. The US does well on some disease indicators (e.g., certain cancers) but poorly on others (e.g., asthma). Compared to other developed countries, life expectancy is lower and mortality is higher (World Bank 2017).

---

## Organization and Governance

### Public and Private Organizations

In the US health-care system, public and private payers purchase health-care services from providers subject to regulations imposed by federal, state, and local governments as well as by private regulatory organizations. Figure 1 illustrates the interplay between four main actors: (1) government, (2) private insurance, (3) providers, and (4) regulators, as well as the types of relationships that connect them.

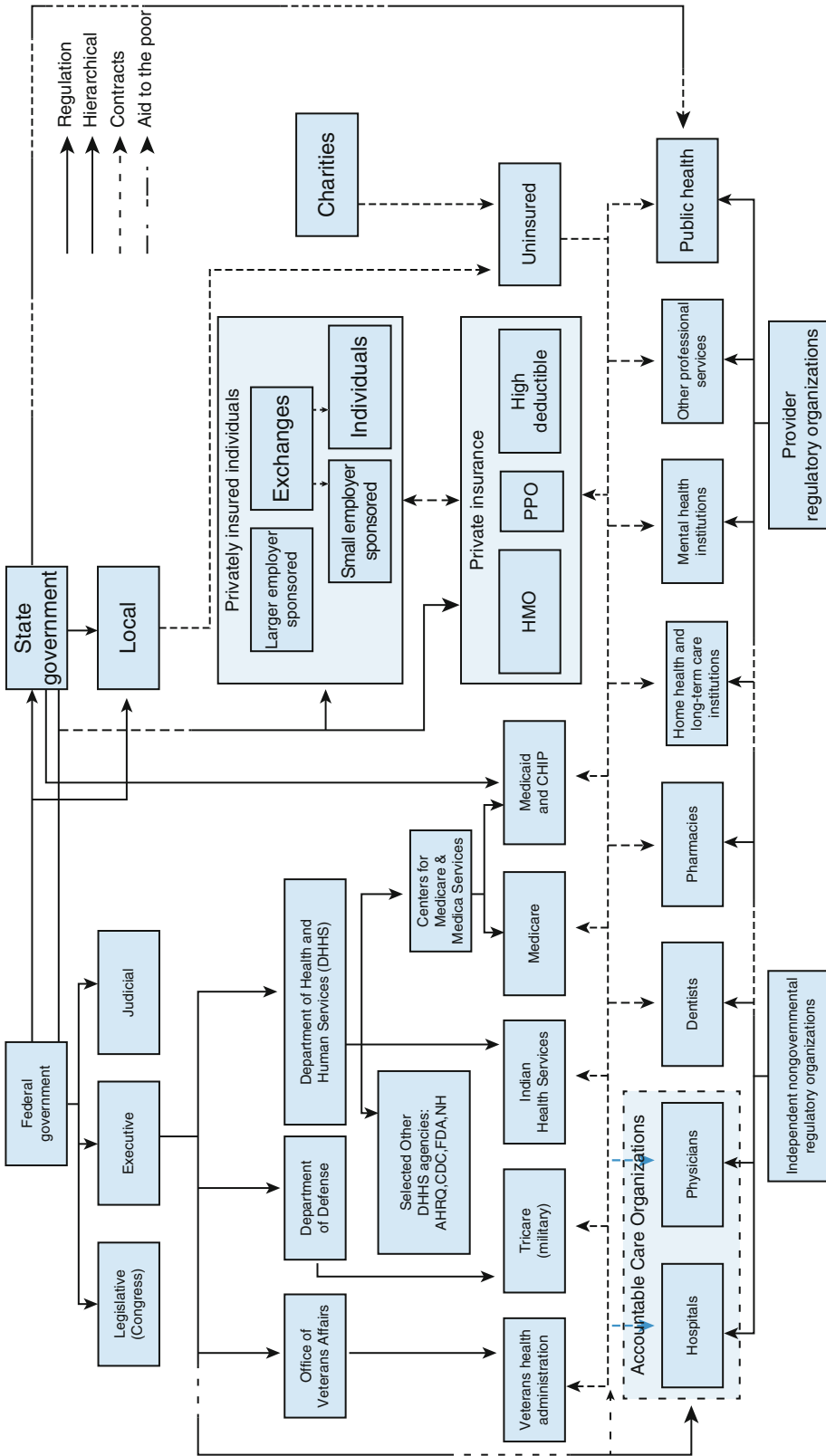


Fig. 1 Organization of the US health system

Government actors include those at the federal, state, and local levels. Both the federal and state governments have executive, legislative, and judicial branches (although the figure only shows this for the federal government). Under the executive branch of the federal government, the Department of Health and Human Services (HHS) plays the largest administrative role in the US health-care system. HHS includes agencies such as the Centers for Medicare and Medicaid Services (CMS) that administer the two major public health insurance programs: (1) Medicare, which provides near-universal coverage for those 65 and older as well as the disabled and those with end-stage renal disease, and (2) Medicaid and the Children's Health Insurance Programs (CHIP), which primarily provide insurance for some low-income families and those with disabilities. Medicaid also covers long-term care services after individuals have used up all their own income and assets and, along with Medicare, low-income seniors (referred to as "dual eligibles"). Other agencies within HHS include research and regulatory agencies such as the Agency for Healthcare Research and Quality (AHRQ), the Centers for Disease Control and Prevention (CDC), the Food and Drug Administration (FDA), and the National Institutes of Health (NIH). The Office of Veterans Affairs, which oversees the Veterans Health Administration to provide care to military veterans, is a federal agency independent of HHS.

Public purchasers include federal and state agencies. Medicare is the largest public purchaser. State governments, along with funds provided by the federal government, purchase health-care services through Medicaid and CHIP, although both programs are state-administered. Both state and local governments are also involved in providing health care in a number of ways making it possible for low-income and other disadvantaged individuals and families to obtain care. These include such things as operating public hospitals as well as providing medical and preventive services through state and local health departments and their associated clinics and community health centers.

In addition to government purchasers, private insurers and individuals also purchase health care

in the US. Private insurance plans have historically been categorized into three types: health maintenance organization (HMO) plans that provide or contract to provide managed care, preferred provider organization (PPO) plans that contract with a preferred network of providers to provide care at lower costs, and high-deductible plans (HDHPs) that typically offer lower premiums but higher deductibles than HMOs and PPOs. The vast majority of Americans with private insurance obtain it through an employer. The Patient Protection and Affordable Care Act (ACA), signed into law on March 23, 2010, is resulting in significant changes in the US health-care system. As shown in Fig. 1, these include the establishment of federal and state-based insurance exchanges for individuals without access to public or employer-based insurance to purchase private coverage as mandated by law. The ACA also allows providers that organize into Accountable Care Organizations (ACOs) to share in savings they achieve in the Medicare program.

### Planning

There is a range of public and private organizations that undertake health system planning in the US. In spite of this, coordinated health planning by various actors as outlined in Fig. 1 is not highly developed. In part this reflects the pluralist and market-oriented nature of the US health-care system. Planning for emergencies and natural disasters, however, is given serious consideration in both the government and private sector. For example, the CDC plans for national and international response to public health emergencies.

### Regulation

All actors in the health-care system are subject to regulation, often from multiple government and nongovernment agencies. Major federal regulatory organizations fall under the umbrella of HHS and include CMS, which regulates public payments to private providers and provider quality; the CDC, which focuses on prevention and control of communicable and noncommunicable diseases; and the FDA, which regulates food and drug safety. State regulatory bodies include public health departments, provider licensing boards,



and insurance commissioners. Local counties and cities also regulate health care through their public health and health service departments including regulating communicable diseases and restaurant safety. Independent nongovernment and provider organizations such as the American Medical Association (for physicians) and the Joint Commission (for hospitals) also play a regulatory role in the US health-care system.

### **Patient Rights**

The US does not have a national comprehensive Patient Bill of Rights (WHO August 2007). The right to health care is not in the US Constitution, and it remains controversial though some states have enacted a Patient Bill of Rights. Some patient rights in the US have been initiated by the court system. For example, the Supreme Court ruled that individuals with disabilities have the right to receive services in non-institutional settings whenever possible. Since the 1990 passage of the Americans with Disabilities Act (ADA), those in the US with physical and/or mental disabilities have been granted additional civil rights. The Health Insurance Portability and Accountability Act (HIPAA) of 1996 governs the security and confidentiality of patient information. As a result of this legislation, how patient information is collected, stored, and transferred is subject to careful protection.

---

## **Financing of Major Insurance Programs**

### **Coverage**

Public purchasers – primarily Medicare and Medicaid – cover more than 30% of the population (Kaiser Family Foundation 2016b). The remainder of the US population – including those with employer-sponsored health insurance, individual private insurance, and the uninsured – are considered private purchasers. More than half of Americans obtain health insurance from their employer. Employer-sponsored coverage is funded by a combination of employer and employee premiums and employee out-of-pocket costs. After

implementation of the ACA and the expansion of the individual private insurance market through income-based subsidies, nearly 16 million Americans have individually purchased coverage, at least half of whom purchased private insurance through one of the federal or state-based exchanges. In 2016, 2 years after the implementation of the ACA's major coverage expansion efforts, approximately 9% of all Americans were uninsured (28 million) including many young adults, minorities, and low-income households (Kaiser Family Foundation, 2017a; Kaiser Family Foundation, 2018a).

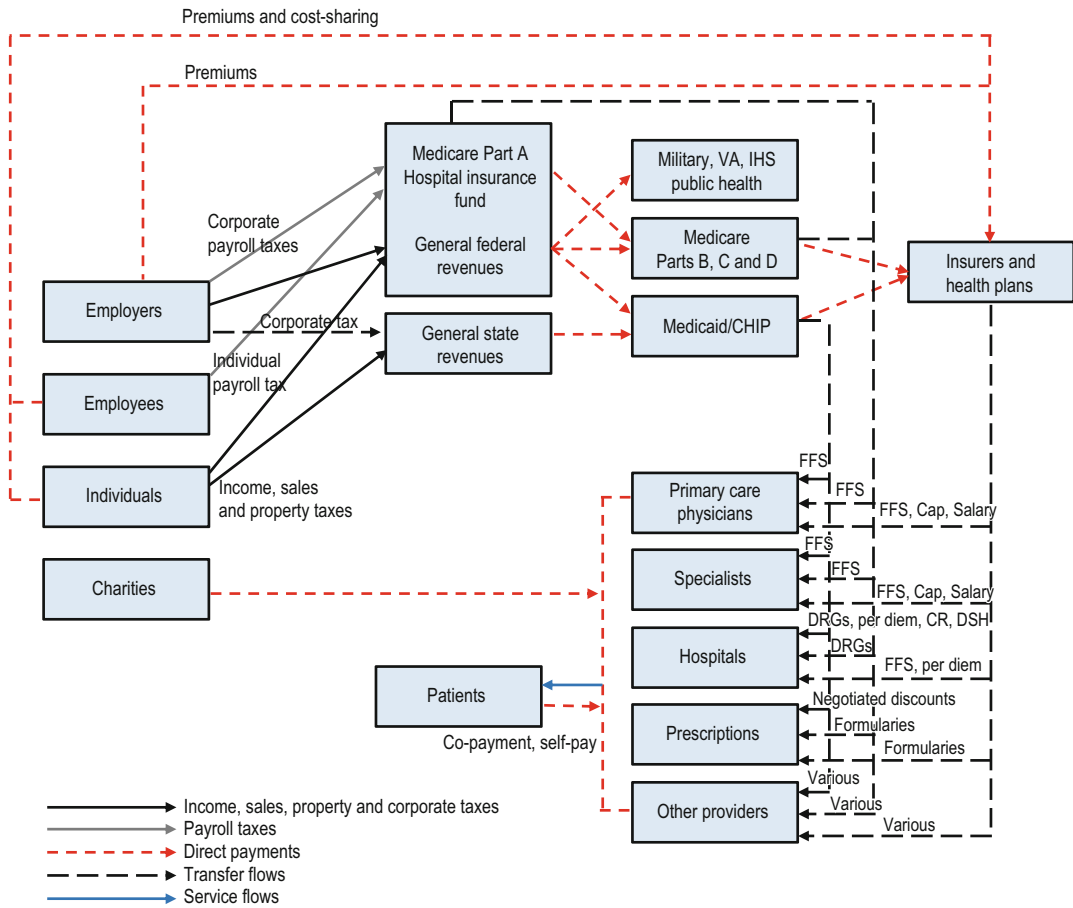
### **Sources of Revenue**

The sources of revenues in the US health-care system have changed considerably over the past 40 years. In 1970, one-third of funding was from out-of-pocket payments. Currently, public sources constitute 37% of spending and private sources 34%, with the remaining 11% out-of-pocket (CMS 2016). While out-of-pocket payments have fallen as a percentage of the total, real out-of-pocket spending per person has actually risen considerably. This is because the size of the health-care system has grown so rapidly.

### **Financing and Financial Flows**

Broadly speaking, financing in the US health-care system originates from employers, employees, and individuals. From them, it flows to private insurers and health plans as well as state and federal governments. Private and public purchasers then transfer dollars to providers through a variety of payment mechanisms. Figure 2 depicts financial flows in the US health-care system.

Beginning with the left-hand side of the figure, employers, employees, individuals, and charities pay into the health-care system through various taxes, premiums and other out-of-pocket expenses, and donations. Employed persons and their families contribute to private employer-sponsored insurance through premiums and cost



**Fig. 2** Sources of revenue, financing, and financial flows

sharing. Individuals may purchase non-group coverage outside of the employment market. In addition to payroll taxes, individuals contribute to general federal and state revenue funds to finance public health-care coverage through income, sales, and property taxes. There is no value-added tax (VAT) in the US.

In the past care for low-income and uninsured individuals has been financed through private charities, a safety net system of public and community clinics, as well as by hospitals and physicians. Additional funding came from general tax revenues, but in many cases the care received was uncompensated and therefore is borne by providers. Prior to the ACA, it was estimated that of the \$57 billion in uncompensated care expenditures, hospitals contribute 61% and physicians

14%, with the remainder coming from a variety of community organizations (Kaiser Family Foundation 2013). In 2011, the federal government, through the Medicaid Disproportionate Share Hospital (DSH) program, allotted \$11.2 billion to hospitals serving a disproportionate number of uninsured and Medicaid patients (Kaiser Family Foundation 2013). These payments were expected to decrease as the ACA was fully implemented and many of the uninsured and those with preexisting conditions acquired health insurance. However, many states have not expanded Medicaid leaving a number of uninsured continuing to require uncompensated hospital care and subsequent legislation delayed reducing DSH payments to hospitals (Kaiser Family Foundation 2016c).

**Table 1** Payment mechanisms for health services

	Payers				
	Medicare	Medicaid/CHIP	Insurers and health plans	Insured individuals	Uninsured individuals
<b>Services</b>					
Inpatient hospital care	DRG	DRG, per diem, CR	FFS, per diem	Co-payment, coinsurance	Direct
Physicians and other health professionals	FFS	FFS, capitation	FFS, capitation, salary	Co-payment, coinsurance	Direct
Prescription drugs	Subsidies for premiums	DAWP	Formularies	Co-payment, coinsurance	Direct
Long-term care and home health	PPS for limited duration	PPS, CR	Per diem for limited duration	Direct	Direct

*Notes:* CR cost reimbursement, DAWP discounted average wholesale price, DRG diagnosis-related group, FFS fee-for-service, PPS prospective payment system

In the US, how health services are paid for depends on the service provided, the type of health worker providing it, the funder, as well as where the service is provided (e.g., hospital or ambulatory care center, California or New York). Given this complexity, the payment mechanisms for each type of health service is shown according to the payer involved (e.g., Medicare, insurers, and health plans) in Table 1.

**Medicare**

The Medicare program provides health insurance coverage to nearly all Americans age 65 and older as well as to many disabled Americans and people with end-stage renal disease – a total of about 55 million people. It covers medically necessary care with the exception of extended long-term care and dental care. Medicare is divided into four parts, labeled Parts A, B, C, and D. Part A, hospital coverage, includes not only hospital care but also some post-acute nursing home, home health care, and hospice care. Part B, supplemental medical insurance, is a voluntary program with essentially the same eligibility requirements as Part A. It covers physicians’ services (both inpatient and outpatient); outpatient care; medical equipment, tests, and X-rays; home health care; some preventive care; and a variety of other medical services. Despite its voluntary nature, about 95% of those eligible enroll in it because it is heavily subsidized.

Part C, Medicare Advantage, is an alternative to Parts A and B. Enrollment is voluntary. It provides coverage for the same services and, at the discretion of the organization offering coverage, sometimes additional benefits such as vision or hearing. One of the main differences between Part C and the preceding two parts which are sometimes called “traditional Medicare,” is that Part C coverage is offered through private organizations (e.g., insurers and HMOs). In 2017, 33% of Medicare beneficiaries were enrolled in Medicare Advantage plans, but aspects of the ACA could lead to reductions in enrollment in the future (Kaiser Family Foundation 2017b).

Part D, prescription drug coverage, began in 2006 and is also voluntary. Like Part C, Part D benefits are provided through private insurers. There are dozens of Part D plans in each state – in addition to dozens of Medicare Advantage plans providing drug coverage in many urban areas. Also like Part C, premiums and benefits vary by plan, with competition occurring based not only on premium differences but also on differences in benefits and, in particular, the drugs that are included on a plan’s formulary that are listed as “preferred” drugs and which therefore are subject to lower patient co-payments. Over 70% of Medicare beneficiaries are covered under Part D. Most other beneficiaries have drug coverage from another source, such as coverage from a former employer, but 12% do not have any drug coverage (Kaiser Family Foundation 2017c).

In addition to services not covered, there are substantial patient cost-sharing requirements. As a result, about 90% of all beneficiaries obtain some form of supplemental insurance coverage, mainly through Medicare Advantage plans (which usually cover additional services), Medicaid, or private policies called “Medigap.” Coverage for hospital care under Part A contains two significant gaps. First, there is a deductible for each inpatient hospital stay. In 2018, that amount was \$1,340 (Medicare.gov 2018a). Second, for those rare stays that exceed 60 days, there are substantial *daily* co-payments. Part A’s nursing home coverage is limited because it is only for short-term skilled care following a hospital admission, rather than extended long-term care. For eligible stays, up to 100 days are covered. During the first 20 days, there are no co-payments, but there is a substantial *daily* co-payment for days 21–100 of a stay of \$167.50. In contrast, there is no co-payment for home health-care services.

Coverage for physicians’ and other medical services under Part B is also subject to patient cost sharing. The patient is responsible for 20% of all covered expenses (with no maximum) after meeting an annual deductible of \$183 (all figures are for 2018) (Medicare.gov 2018b). The 20% coinsurance requirement is perhaps the main reason why the vast majority of Medicare beneficiaries seek some form of supplemental insurance coverage. It is difficult to generalize about the depth of coverage under Part C because each plan has its own benefit structure. Federal minimum requirements are that coverage be at least as comprehensive as under Parts A and B. As noted, most Part C plans offer additional services. About 80% offer prescription drug coverage. It is also difficult to generalize about Part D (stand-alone prescription drug coverage) because benefits vary by insurance plan. The main characteristic is a feature called the “donut hole.” Insurers provide coverage (with cost sharing) up to a certain amount of drug spending per year, at which point there is a period of no coverage at all. When total drug spending reaches a “catastrophic” level, almost all drug costs are covered. As part of the ACA, the donut hole will shrink and is scheduled to be eliminated by 2020.

## Medicaid

Unlike Medicare, which is available to nearly all individuals age 65 and older, Medicaid is a means-tested program. It is designed to provide health insurance for those with the lowest-income levels and fewest assets, the disabled, and to poor seniors with Medicare coverage, as well as the disabled and seniors who have exhausted their financial resources, often as a result of very high long-term care expenses. Medicaid is a key resource for some of the poorest and sickest Americans.

Medicaid programs are state-based, but they are funded jointly by the states and the federal government. In return for federal dollars, states are required to meet certain federal government standards. Participation by the states is voluntary though historically all of the states have chosen to participate. Services are largely purchased from the private sector. Until 2014, the federal government paid between 50 and 74% of Medicaid costs proportional to each state’s income, with the states paying the remainder. Beginning in 2014, federal contributions changed for those states that expanded Medicaid, with the federal government paying 100% of costs for those newly eligible, gradually falling to 90% by 2020.

Medicaid covers several distinct population groups. The breadth of coverage varies across states according to these population groups and by state.

Prior to the ACA, the main groups typically covered by Medicaid were as follows:

- Low-income children
- Low-income pregnant women
- Low-income disabled persons
- Low-income senior citizens
- Low-income parents of dependent children

For adults, in some states that have not expanded Medicaid coverage, not only are there income restrictions but also asset limitations that can preclude eligibility.

Medicaid covers roughly 17 million more Americans (a total of 74 million) than Medicare. As noted, the breadth of coverage varies considerably by eligibility group and by state. As of

February 2018, 33 states and the District of Columbia had expanded their Medicaid coverage in accordance with the ACA, and 18 had not (Kaiser Family Foundation 2018b). In those states that have chosen to expand, all adults and children below 138% of the federal poverty level (FPL) are now eligible for Medicaid. (In 2017, the federal poverty level was \$12,060 for a single individual and \$24,600 for a family of four.) (Healthcare.gov, 2018).

In the other states, children and pregnant women have the most liberal eligibility requirements. States are required to cover pregnant women and children up to age six if their incomes are at or below 138% of the federal poverty level (FPL) and children ages 6–18 up to 100% of the FPL. Many states employ even higher, or more generous, income eligibility thresholds. When combined with CHIP coverage, the median state provides coverage to children up to 235% of the FPL and pregnant women up to 185%. To illustrate the critical role that Medicaid plays for pregnant women, the program pays for 45% of all births in the US. Coverage is somewhat narrower for seniors and the disabled, however, with eligibility mandated up to 75% of the FPL.

In the 18 states that have not expanded coverage, low-income parents of dependent children face the most stringent eligibility requirements. Nine states cover them only if their incomes are below 40% of the FPL – with Alabama and Texas providing such coverage only up to 18% of the FPL (i.e., an annual income even as low as \$2,200 would disqualify an individual from coverage in that state). In contrast, Connecticut and the District of Columbia cover these adults at in excess of 200% of the FPL or higher, taking advantage of the joint funding by the federal government. Recently, several states have either considered or passed legislation that would also impose work requirements on many Medicaid recipients of working age (Kaiser Family Foundation 2018b). This illustrates the large variation in breadth of coverage that currently exists between states, although this variation has been reduced considerably as a result of the ACA.

Beginning in 2014, states that choose to expand their Medicaid coverage will receive

100% of the costs from the federal government to add all poor people and the near poor up to 138% of the poverty level to Medicaid rolls for 4 years. The federal contribution will gradually decrease to 90%.

Several states have petitioned the federal government for special arrangements in their Medicaid expansion, and they have received approval to proceed. These are called “1115 demonstration waivers” and typically involve exceptions to the usual Medicaid rules that are budget neutral for CMS. Examples include charging a co-pay or premium to recipients for services, imposing a penalty for nonpayment of premiums, including work requirements, offering “wellness incentive” programs, and structuring the program like a health savings account (HSA). As of February 2018, 35 states have received waivers from CMS to tailor their own Medicaid programs (Kaiser Family Foundation 2018c).

The initial evidence on the effectiveness of these innovations to save money, improve the quality of care, and/or improve population health is limited. However, states are required by CMS to report such evidence during the demonstration waiver. Almost all of the waivers add to the complexity of the Medicaid program and could increase the cost of administration. This will be evaluated by CMS going forward. In the tradition of American federalism, successful innovations could spread to other states.

The scope of coverage under Medicaid is generally wide but varies by state. Federal law requires that states provide the following services: inpatient and outpatient hospital, physician, nurse practitioner, laboratory and radiology, nursing home and home health care for those age 21 and older, health screening for those under age 21, family planning, and transportation. Other services are optional for states. This designation means that if a state chooses to cover the service, it will receive matching funds from the federal government. Optional services include some major services such as prescription drugs and dental care but also such things as care provided by professionals besides physicians and nurse practitioners, durable medical equipment, eyeglasses, rehabilitation, various types of

institutional care, home- and community-based services, personal care services, and hospice.

In general, those eligible for Medicaid receive services at little or no cost. However, states sometimes put restrictions on the number of services that are covered per year. Moreover, payments to physicians are usually low. In 2013, about 30% of physicians reported that they would not take new Medicaid patients (Decker 2013). Psychiatrists were the most likely to reject new Medicaid patients (56%), and cardiovascular disease specialists see the most, with only 9% rejecting such patients (Decker 2013).

One development with the potential to provide more mainstream access to physician office care is the movement toward the use of managed care in the Medicaid program. Over 70% of Medicaid beneficiaries are in managed care plans. The exact nature of these arrangements varies from state to state. Some include capitation (rather than fee-for-service) for providers and/or primary care case management. States often prefer managed care both as a means of enhancing quality and controlling costs and are likely to rely on it as the program expands through provisions in the ACA.

## Private Insurance

In 2016, 179 million Americans were covered by private insurance; 157 million of these had employer-sponsored coverage (Kaiser Family Foundation 2016d). While having employer-sponsored insurance is almost always advantageous – employers generally subsidize premiums – it is not available to everyone. First, it is necessary to be employed or be a family member of someone employed. Second, the employer has to offer coverage; until 2015 or 2016, it was completely voluntary on the part of the employer. Third, if coverage is offered, the employee has to be eligible for it. And fourth, even if eligible, the employee has to be willing to pay the employee's share of the premiums, which can be considerable. It is the people who are better-off economically who are able to meet the four conditions mentioned above. Individuals and families without

an entry into the employer insurance market, and who are not eligible for Medicare and Medicaid, often seek coverage individually. Historically, individual coverage has had several disadvantages over employer group coverage and therefore was normally purchased only if the employer-sponsored coverage was unavailable. Prior to the ACA, plans purchased in the individual private market were usually unsubsidized; administrative costs tended to be high (25–40%); health examinations were often necessary; cost-sharing requirements were, on average, higher; and fewer types of services tended to be covered. However, the individual market is changing substantially with the creation of the health insurance exchanges under the ACA.

Some employers, particularly larger ones, offer a choice of health insurance products to their employees. Among firms offering a choice, only about 20% of employees nationally can choose among three or more plans (California HealthCare Foundation 2009). For federal government employees, there can be dozens of choices. Employees with a choice can generally switch to a different plan irrespective of their health history or status once per year.

Historically the most common arrangement offered by employers was a PPO. Among all covered workers, in 2017 48% were enrolled in PPOs, 14% in HMOs, 10% in point of service plans (POS – a blend of HMO and PPO arrangements that allow members to seek care from non-network providers at a higher cost), 28% in high-deductible plans (note that some of these may be PPOs or HMOs), and less than 1% in conventional insurance (traditional fee-for-service) plans (Kaiser Family Foundation 2017d). The biggest change in recent years has been the relatively rapid rise of high-deductible plans with a savings option, many of which are classified as health savings accounts (HSAs). In HSAs, the policy holder agrees to purchase insurance with a high deductible (currently averaging about \$2,200 annually for individual coverage and twice that for family coverage). Premium contributions can be made by the individual and/or employer. These contributions are tax deductible, can accumulate year to year if unspent, and therefore can be used for future medical

expense. They can be withdrawn to pay for eligible medical care.

Market share in health insurance is dominated by larger firms that generally market nationally. (Blue Cross Blue Shield plans, while having a national presence, usually market in individual states.) In 2013, three of the largest insurers covered 80% of people enrolled in individual, small group, and large group private insurance markets in at least 37 states (US Government Accountability Office 2014).

Prior to January 2014, insurers priced their productions in two ways: experience rating and community rating. Under experience rating, the most common technique used, insurers charged employers (or individuals) on the basis of the past cost experiences or, when data is lacking, on predicted expenditures. In contrast, community rating entailed charging the same amount to all groups (or even individuals). In the individual insurance market, premiums were generally experience-rated. Each individual went through medical underwriting in which their risks are assessed.

Under the ACA, state-based exchanges combined with the individual mandate to purchase insurance are intended to reduce adverse selection problems in the individual and small group market by requiring plans selling in exchanges use community rating (older individuals can be charged more than the younger, but differences within age cohorts will be prohibited), rather than experience rating, and by increasing risk pooling to a far greater extent than has been the case in the past in the US. Exchanges will also reduce or eliminate the need for individuals to purchase insurance through agents or brokers, whose fees can absorb 20% of the total premium during the first year of enrollment (Whitmore et al. 2011). One of the key requirements of the ACA is that individuals purchase coverage or pay a penalty. Similarly, firms with more than 50 employees will also have to provide coverage or pay a penalty. These “sticks,” combined with the “carrots” of subsidies for individuals to purchase coverage, will, it is hoped, lead to a system where community rating will be viable.

There are significant user charges associated with private insurance. Beginning with premiums,

the average cost of employer-based single coverage was \$6,690 in 2017, 18% of which was paid by the employee. For family coverage, it was 31% of the total cost of \$18,764. The percentage of family coverage paid by the employee has risen considerably over the past decade – by 6.8% per year compared to 4.8% for the share paid by the employer (Kaiser Family Foundation 2017a). This is one of several examples of how employers have shifted more costs onto employees as health-care costs have risen.

As is the case in many high-income countries, there are often substantial co-payments for prescription drugs. In most employer-sponsored plans, there are multiple “tiers,” each of which has its own cost-sharing requirements. Their purpose is mainly to encourage the use of cheaper drugs, particularly generics, the use of which has grown substantially in recent years. One way in which employer coverage tends to be more generous than Medicare’s is that there is usually a limit on annual out-of-pocket expenditures. Over 80% of employer-sponsored health plans establish such a maximum. In 2014 the median out-of-pocket maximum for an employee with individual coverage was approximately \$6,000 (Kaiser Family Foundation 2014a).

Administrative costs tend to be higher in private insurance than government-sponsored programs like Medicare and Medicaid. This is a result of several factors in addition to the need for profits. Private insurers engage in “underwriting” activities, which involve examining past claim expenses to determine a competitive, yet still profitable premium to charge. They also need to market and advertise since, unlike government programs, they do not have a captive audience. Finally, to protect themselves against unexpectedly high claims, insurers often need to factor in a risk premium. Estimates vary on the size of administrative costs (including profits and taxes). Most agree, however, that administrative costs are much higher for insurance policies covering individuals and small firms. One study, conducted by a US actuarial firm, estimated that in 2003, private insurers spent 16.7% on administrative costs. Among the latter, administrative costs were estimated to be 30% in the individual

market, 23% in the small employer market, and 12.5% for large employers (Milliman 2006). In contrast, Medicare administrative costs for the overall program were 1.4% (Centers for Medicare and Medicaid Services, 2016).

---

## Physical and Human Resources

A health-care system requires adequate physical and human resources for the delivery of health care. Physical resources encompass capital stock, infrastructure, medical equipment, and information technology. Human resources are practitioners who diagnose and treat patients, technologists, technicians, and support occupations (Bureau of Labor Statistics (BLS) 2011a, b).

### Physical Resources

#### Capital Stock

Table 2 presents trends in the number of several types of health-care facilities in the US for selected years through 2012. The total number of ambulatory care facilities increased by 24% from 1997 to 2012. All types of ambulatory facilities, such as physician and dentist offices, ambulatory surgical centers, and rural health clinics, experienced this growth. Ambulatory surgical centers and rural health clinics grew tenfold or more between 1980 and 2012.

In contrast to the growth in ambulatory care, the number of hospitals decreased significantly from 1975 to 2009. The consolidations and closings of hospitals that contributed to this decline are related to changes in hospital payment from retrospective to prospective and the rise of managed care practices promoting reduced lengths of stay and competition between hospitals (Harrison 2007).

The total number of nursing homes also decreased, but the number of skilled nursing homes increased threefold. The number of Medicare-certified home health and hospice agencies increased fivefold or more, most likely in response to changes in Medicare reimbursement and shifts from inpatient to outpatient care.

## Institutional Infrastructure

A number of changes have occurred in the infrastructure of health-care institutions in the past decades. Figure 3 shows that between 1970 and 1990, the number of community hospital beds per 1,000 population declined by 14%. From 1990 to 2012, the decline was even greater, at 30%. The number of beds in psychiatric institutions fell 58% from 1970 to 1990 and another 36% from 1990 to 2000, leveling off in 2000. The number of skilled nursing home beds fell nearly 15% from 1990 to 2012.

## Medical Equipment

The use of medical equipment has skyrocketed over the past decades. Reductions in hospital length of stay and the provision of more acute care on an outpatient basis require a greater use of medical equipment (Danzon and Pauly 2001). Medicare, Medicaid, and private insurance companies indirectly cover the costs of medical equipment in medical facilities as part of the overall reimbursement for care and directly cover the costs of medical equipment to individuals (Tunis and Kang 2001).

## Information Technology

Health information technology (HIT) has become an important part of health care (Hersh 2009). On the provider side, medical record-keeping, decision-making, imaging, and prescribing can now be aided by computer and Internet data storage, organization, and retrieval. On the consumer side, the Internet has become a source of information (and misinformation) on health care, and patients may be able to communicate with physicians through email. HIT is slowly integrating the provider and consumer sides so that patients can view and add to their medical record online (Hogan and Kissam 2010).

The adoption of health information systems has been slow in the US. In 2013, 78% of office-based physicians used some kind of electronic health record (EHR) in their practice, while 59% of hospitals had a basic EHR system (Adler-Milstein et al. 2014; Hsiao and Hing 2014).

The US government has put significant funding into the expansion of HIT. In 2009 the



**Table 2** Number of selected types of health-care facilities in the US, 1975–2012

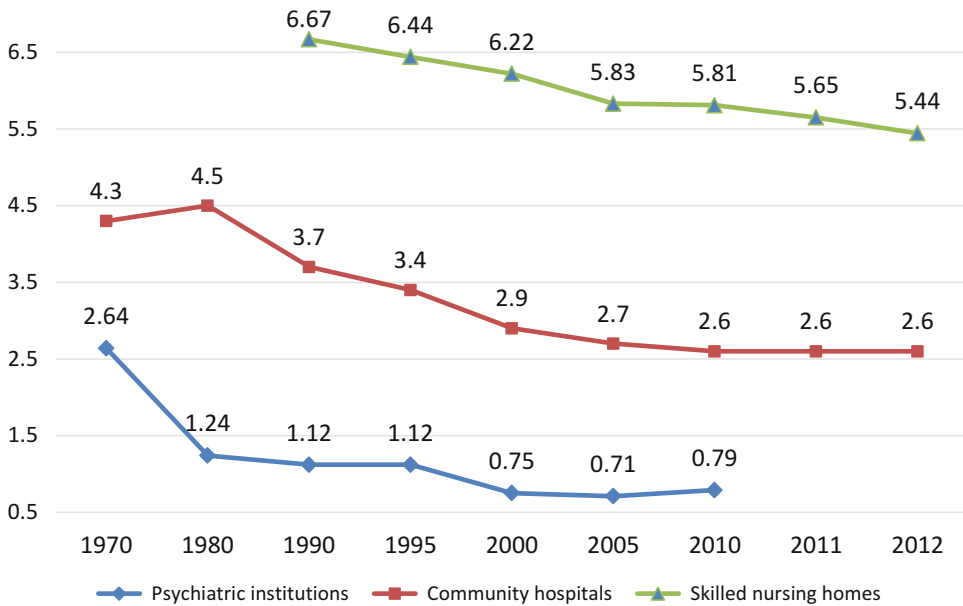
Type of facility	Number of facilities											% chng
	1975	1980	1985	1990	1995	2000	2005	2010	2012			
Ambulatory care (all facilities) <sup>a</sup>	—	—	—	—	455,381 <sup>a</sup>	489,038 <sup>a</sup>	547,709 <sup>a</sup>	—	582,733	24.53		
Physicians' offices <sup>a</sup>	—	—	—	—	195,449 <sup>a</sup>	203,118 <sup>a</sup>	209,730 <sup>a</sup>	—	221,470	12.48		
Dentists' offices <sup>a</sup>	—	—	—	—	114,178 <sup>a</sup>	118,305 <sup>a</sup>	127,033 <sup>a</sup>	—	133,194	15.37		
Ambulatory surgical centers (Medicare certified)	—	—	336	1,165	2,112	2,894	4,445	5,316	5,335	176.3		
Rural health clinics (Medicare certified)	—	391	428	517	2,775	3,453	3,661	3,845	3,940	163.89		
Hospitals (all)	7,156	6,965	—	6,649	6,291	5,810	5,756	5,794	5,723	-22.25		
Nursing homes (all)	—	—	—	—	16,389	16,886	—	15,700	15,673	-4.46		
Skilled nursing homes (Medicare certified)	—	5,052	6,451	8,937	—	14,841	15,006	15,084	15,132	99.88		
Home health agencies (Medicare certified)	2,242	2,924	5,679	5,661	8,437	7,857	8,090	10,914	11,930	136.72		
Hospices (Medicare certified)	—	—	164	772	1,927	2,326	2,872	3,405	3,509	182.14		
End-stage renal disease facilities (Medicare certified)	—	999	1,393	1,987	2,876	3,787	4,755	5,631	5,766	140.93		

Sources: For ambulatory care facilities (all, physicians' offices, dentists' offices), Census Bureau 2010 (NAICS data). Obtained from <http://factfinder.census.gov/>. For hospitals: *Health, United States, 2013*, Table 107; *Health, United States, 2014*, Table 98. For nursing homes (all): *Health, United States, 2014*, Table 101. For the Medicare-certified facilities of all types: *Health, United States, 2013*, Table 111. Column for 2012 in table uses 2011 data

Notes: — Data not available

Information is not available about the methods for counting the number of facilities. We assume that each stand-alone facility is counted whether it is part of a larger organization or not. In that case if a merger results in the closing of one facility, the number of facilities will decrease, but if a merger does not result in the closing of a facility, the number will be unchanged

<sup>a</sup>Years for these numbers are 1997, 2002, and 2007, respectively. The numbers for 2007 are estimations



**Fig. 3** Number of beds in US community hospitals, psychiatric institutions, and nursing homes per 1,000 population, 1970–2012 (Notes: Community hospitals are defined as nonfederal, short-term general, and other specialized hospitals. The types of facilities included in the category of community hospitals have changed over time. Psychiatric institutions are defined as all 24-h psychiatric hospitals and residential treatment organizations. Skilled nursing homes are those that are certified with the Centers

for Medicare and Medicaid Services. Sources: (1) For community hospitals: *Health United States, 2006, 2007, 2008, 2009, 2011*. (2) For psychiatric hospitals: Foley et al. (2004), DHHS pub. no. (SMA)-06-4195, chap. 19; *Health, United States, 2009*, Table 119; *Health, United States, 2011*, Table 117. (3) For skilled nursing homes: *Health, United States, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2011*)

Health Information Technology for Economic and Clinical Health (HITECH) Act was passed. It provides \$30 billion to hospitals to adopt EHRs. Hospitals must build systems that have “meaningful use” in stages of increasingly advanced requirements (Adler-Milstein et al. 2014). In addition, the ACA has incentivized physicians and hospitals to adopt EHRs by encouraging innovations such as ACOs, which are difficult to run without an EHR (Adler-Milstein et al. 2014).

**Human Resources**

**Health-Care Workforce**

Table 3 presents the numbers of workers employed in several health-care occupations between 1990 and 2014. Increases in employment occurred with most health-care diagnosing and treating practitioners, such as physicians,

chiropractors, registered nurses (RNs), and therapist occupations. Employment also increased with most of the technologist and technician occupations and all of the support occupations. Employment fell for dentists, physician assistants, and clinical laboratory personnel.

**International Mobility**

The numbers of US health-care professionals include immigrants to the US and exclude emigrants from the US. In 2014, 26% of physicians and 24% of residents in specialty programs in the US were international medical graduates (Ranasinghe 2015). Over 8% of the US nursing workforce in 2004 consisted of international nursing graduates (US DHHS 2010).

Although immigrants add to the health-care workforce supply, there is no evidence that they improve distributional issues. Furthermore, a reliance on immigration reduces the incentive to

**Table 3** Employed US health-care personnel per 1,000 population, 1990–2014 (selected occupations)

	1990	1995	2000	2005	2010	2011	2012	2013	2014	% chng
<b>Health-care diagnosing and treating practitioners</b>										
Chiropractors	–	–	0.15	0.28	0.18	0.18	0.19	0.18	0.21	0.56
Dentists	0.64	0.59	0.61	0.55	0.57	0.58	0.53	0.58	0.60	–0.01
Optometrists	0.09	0.13	0.12	0.14	0.12	0.09	0.11	0.13	0.15	0.27
Pharmacists	0.69	0.65	0.80	0.84	0.83	0.88	0.91	0.88	0.92	0.15
Physicians and surgeons	2.32	2.64	2.62	2.81	2.82	2.64	2.91	2.95	3.18	0.20
Physician assistants	–	–	0.15	0.25	0.32	0.26	0.35	0.41	0.26	–0.84
Podiatrists	0.06	0.04	0.02	0.04	0.04	0.02	0.03	0.04	0.03	0.01
Registered nurses	6.70	7.52	7.79	8.17	9.21	8.68	9.19	9.15	9.06	0.16
Occupational therapists	0.15	0.20	0.20	0.29	0.35	0.36	0.38	0.35	0.35	0.59
Physical therapists	0.37	0.49	0.51	0.60	0.61	0.71	0.67	0.71	0.77	0.45
Respiratory therapists	0.25	0.36	0.27	0.32	0.42	0.43	0.35	0.35	0.35	0.27
Speech-language therapists (pathologists)	0.25	0.35	0.31	0.33	0.43	0.40	0.47	0.43	0.43	0.35
<b>Health-care technologists and technicians</b>										
Clinical laboratory technologists and technicians	1.20	1.42	1.02	1.13	1.11	1.03	1.02	1.08	0.92	–0.10
Dental hygienists	0.35	0.36	0.39	0.45	0.46	0.47	0.52	0.58	0.55	0.35
Licensed practical and licensed vocational nurses	1.77	1.52	1.81	1.72	1.86	1.80	1.70	1.77	2.01	0.11
Medical records and health information technicians	0.28	0.08	0.31	0.41	0.38	0.37	0.29	0.28	0.43	0.36
<b>Health-care support occupations</b>										
Nursing, psychiatric, and home health aides	5.87	6.69	5.24	6.42	6.24	6.36	6.77	6.75	6.21	0.16
Dental assistants	0.76	0.80	0.76	0.88	0.97	0.98	0.88	0.88	0.86	0.12

Sources: Current Population Survey (CPS), Bureau of Labor Statistics, HRSA, DHHS; US Census Bureau, Census 1990, 2000, 2010, and population estimates 2011–2014

Notes: Dashes indicate data are not available. % change is from 1990 to 2014 or from the earliest year. A new occupational classification system for occupational employment (SOC) was introduced by the CPS in 2003. The 1990 and 1995 data are based on the old classification system and may not be fully comparable to later data. The table reports numbers employed rather than full-time equivalents (FTEs), so the actual amount of human resources employed may be less than that reflected in the table due to part-time employment. On the other hand, since these are employment numbers, the total number of individuals in each occupation would be larger if unemployed individuals were counted

Calculations: Employment and population were rounded to three decimal places

expand educational capacity, raise wages, and improve working conditions (Flynn and Aiken 2002). Finally, migration from low-income countries is a “brain drain” for those countries (Aiken 2007).

**Distribution**

The US has a high proportion of specialists to primary care physicians (around 1.5 times as many in 2012) (Hing and Hsiao 2014). Further, the primary care physician to population ratio in

rural areas is only 4/5 that of urban areas (Hing and Hsiao, 2014). In nursing, the biggest distributional issue is the low number of RN faculty (AACN 2017). This creates bottlenecks in the educational process and contributes to nursing shortages (AACN 2017). The ACA includes policies aimed at improving supply and distribution issues related to primary care including scholarships and loan repayment programs for primary care physicians, short-term increases in primary care payment rates for Medicaid, and additional

support for Federally Qualified Health Centers to provide essential health services to more uninsured and low-income patients.

### **Adequacy**

Projections of the adequacy of physicians using several forecasting models indicate a future shortage of physicians of 5–20% by 2020 (COGME 2005; BHPPr 2008). Other projections indicate that a smaller increase in supply would be needed if distributional issues were improved or if there was an increased use of nonphysician providers and osteopaths (Weiner 2007). In nursing, forecasters unanimously predict a large future shortage (BHPPr 2010).

---

## **Provision of Health-Care Services**

The US has several major health-care sectors, including public health, primary, specialty, acute inpatient, dental, mental health, pharmaceutical, post-acute, long-term, and palliative care. Access to these services and navigation through the US health-care system differs depending upon the care that is needed and whether an individual is insured or uninsured. Insured individuals tend to enter the health-care system through a primary care or specialty provider. Uninsured individuals often do not have a regular primary care provider but instead may visit community health centers and emergency departments. Due to out-of-pocket costs, they may be reluctant or unable to seek care unless they are experiencing an emergency.

### **Public Health**

Public health focuses on promoting health at the population level through investigating and intervening in the environmental, social, and behavioral factors in health and disease. It emphasizes prevention and health promotion (Shi and Singh 2012). Public health is promoted mostly through public agencies. At the federal level, public health services are headed by the US Public Health Service (USPHS), a division of HHS. There are several subdivisions within the USPHS, such as the

CDC. Federal laws allow state health agencies to determine the scope and amount of services and to establish the vehicles for providing those services. As a result, the services vary significantly across the states. Local public health agencies at the county or city levels (“health departments”) carry out many public health functions (Salinsky 2010).

Public health services include communicable disease control, environmental hazard prevention, emergency terrorism preparedness and response, occupational health, health promotion and screening, and licensing, regulation, and planning of health-care facilities and providers.

## **Outpatient Services**

### **Primary Care**

In 2010 55% of the visits to physicians in the US were to a primary care physician (US Department of Health and Human Services 2014). Primary care practitioners are physicians, nurse practitioners, physician assistants, and nurse midwives who are generalists or who specialize in family medicine, internal medicine, pediatrics, obstetrics, and gynecology (Bodenheimer and Pham 2010).

Access to primary care requires that patients have the ability to pay for care, adequate transportation to care, and the health literacy to demand and use the care; it also requires that the supply, distribution, and time of providers are adequate (Shi and Singh 2012). For these reasons, the uninsured and those with insurance but unable to afford high out-of-pocket costs due to inadequate coverage have difficulty accessing primary care. Additionally, those covered by Medicaid may experience problems accessing primary care due to their inability to find a private physician that accepts Medicaid patients (Shi and Singh 2012).

### **Specialty Care**

Forty-five percent of visits to physicians in the US in 2010 were to specialists (US Department of Health and Human Services 2014, Tables 91, 92). Many of the issues with access to primary

care are even more of a concern with specialty care. Care coordination among primary care and specialist providers is a growing issue in the US, where the typical Medicare beneficiary sees two primary care physicians and five specialists a year, and patients with multiple conditions may see up to sixteen physicians (Bodenheimer 2008). This can lead to over-, under-, and conflicting treatment and polypharmacy. Two initiatives to improve care coordination in the US are patient-centered medical homes (PCMHs) and ACOs (Phillips and Bazemore 2010; CMS 2012). In PCMHs each patient has an ongoing relationship with a primary care provider, who directs the medical team, and the patient's care is coordinated across all health-care settings, with patients actively participating in decision-making (Rittenhouse et al. 2011). In ACOs payment from Medicare is tied to the performance of the provider organization, thus conferring financial risks and rewards for care management and patient outcomes to providers.

### Emergency Care

Emergency departments (EDs) are a major part of the US health-care safety net (Shen and Hsia 2010). EDs in hospitals that receive payment from Medicare are required by the Emergency Medical Treatment and Active Labor Act (EMTALA) to provide care to anyone needing emergency treatment. Hospitals must care for the individuals until they are stable. This allows under- and uninsured persons access to the ED for emergency conditions.

EDs tend to be overused for nonurgent problems and for serious problems that could have been prevented with better primary and specialty care. ED overcrowding, long wait times, hospital diversions, the lack of ED space and staff, and patient boarding have been problems for many years (GAO 2009).

### Urgent Care

Urgent care is walk-in care provided outside the ED setting in centers that are open in the evening on weekdays and at least 1 day over the weekend (Weinick et al. 2009). Services focus on acute episodic care for minor illnesses and emergencies such as upper respiratory infections, lacerations,

and fractures. Medical care is typically performed by family physicians, nurse practitioners, and physician assistants (Weinick et al. 2009).

In 2011 there were more than 9,000 urgent care centers (UCCs) in the US (Yee et al. 2013). Urgent care services have expanded in response to difficulties in seeing primary care practitioners on an urgent basis and after-hours, high ED costs, and long ED wait times (Yee et al. 2013). Some individuals use UCCs because they do not have a regular source of primary care. An individual must have insurance or pay out-of-pocket for care.

### Retail Clinics

Located in pharmacies, grocery stores, and department stores, retail clinics are emerging as places to go for treatment of minor medical conditions (RAND 2010). They tend to be staffed by non-physician practitioners, such as nurse practitioners or physician assistants, and they treat a limited number of conditions and needs, such as skin conditions, sore throats, pregnancy testing, infections, diabetes screening, and immunizations (Mehrotra et al. 2008).

### Acute Inpatient Care

Individuals who are acutely ill and need to have round-the-clock care require inpatient care provided in hospitals. The availability of hospital services depends upon the insurance status of the individual seeking care, the type of hospital, and the geographic area. For those who have private or public insurance, care is accessed through a physician referral to a hospital that the physician recommends and that is in the insurance provider network. For those without insurance, access to care depends upon how sick they are.

When an uninsured patient's condition is not an emergency (such as planned surgery), access to hospital care becomes dependent upon hospital ownership. Government-owned hospitals must provide charity care to those who do not have insurance or cannot pay for out-of-pocket portions of their care (Weissman et al. 2003). These hospitals provide the majority of charity care in the US (Weiner et al. 2008). Charity care is also provided

by nonprofit private hospitals. It is financed through federal payments for treating Medicaid patients for DSH hospitals, tax exemptions, and cross-subsidies from other payers (Weissman et al. 2003). For-profit hospitals also provide charity care, but they do not receive tax exemptions for this, and it is unclear whether they provide as much charity care as nonprofit hospitals (Cram et al. 2010; Schlesinger et al. 2003). The expansion of health insurance, as being undertaken through the ACA, is expected to improve access to inpatient care in the US and reduce hospitals' uncompensated care costs, cost shifting, and other irrationalities of the system.

## Mental Health Care

The mental health-care landscape has changed significantly over the past decades. Long-term institutionalization, which was a major treatment strategy for many mental health problems, is no longer the preferred way to treat those problems. Instead, treatment occurs through outpatient care, accompanied by the increased use of pharmaceuticals which can be managed on an outpatient basis, and short-term inpatient stays (US Department of Health and Human Services 2014, Table 106; Ling et al. 2008).

Only about one-third of Americans with mental health problems actually receive treatment for their problem (Cunningham 2009). Insured patients generally receive mental health care in the outpatient settings of offices of private psychiatrists, psychologists, and licensed social workers and inpatient settings of private psychiatric and general hospitals (Shi and Singh 2012). Patients without insurance who cannot pay out-of-pocket expenses are treated in state and county mental health hospitals, community health centers, EDs, and hospitals (Shi and Singh 2012). Other access issues include shortages of mental health providers and the stigma that is attached to mental illness (Cunningham 2009).

A goal of the ACA is to improve access to mental health care by promoting mental health parity and expanding insurance coverage for mental health. Insurance regulation will prohibit

discrimination against those with preexisting mental health conditions, increasing rates, or canceling insurance for those who develop mental health conditions.

## Pharmaceutical Care

Spending on prescription drugs has been the fastest-growing component of US health costs until just recently. Since 1970 spending increased rapidly until 2001 (CMS 2014). From the 1990s to 2015 US spending on retail prescription drugs increased from 7% to 12% of total health expenditures (GAO, 2017). Pharmaceutical production and marketing in the US are completely privatized but are regulated by the Food and Drug Administration (FDA). Prices are not regulated, although the government negotiates payment discounts in some of its programs such as Medicaid (but not Medicare where a provision in the Part D legislation prohibits Medicare from negotiating bulk discounts on drugs).

Many pharmaceuticals are overused, inappropriately used, and underused in the US. Overuse and inappropriate use occur with certain medications such as antibiotics and antidepressants and with the practice of polypharmacy among the elderly (Conti et al. 2011; Misurski et al. 2011; van der Hooft et al. 2005). Underuse is associated with financial barriers. In 2011, 23% of individuals in the National Health Interview Survey reported cost-related medication underuse (Berkowitz et al. 2014).

Overuse of medications has been cited as result of aggressive marketing by pharmaceutical companies to both physicians and consumers (Brody and Light 2011; Budetti 2008; Williams et al. 2011). Pharmaceutical companies sometimes market their drugs by taking advantage of new diseases, literally promoting the existence of the disease in their advertisements (also known as "disease mongering") (Brody and Light 2011). A health problem is reframed and promoted in the media and popular culture as having a pharmaceutical solution (Williams et al. 2011). These strategies have been termed "pharmaceuticalization." Whether a condition is a true health

problem and is best treated with pharmaceuticals or other products, or has been pharmaceuticalized, is controversial (Metzl and Herzig 2007).

## Long-Term Care

Long-term care consists of a number of different health-care services for individuals with conditions that are not expected to significantly improve and that need ongoing care.

Through a complex financial web, essentially all Americans have access to nursing homes. The financial options are as follows: If an elderly person is admitted to a nursing home post hospitalization, Medicare will cover a limited amount of skilled nursing days, contingent upon rehabilitation progress. If the individual needs to stay beyond Medicare-covered days, or was never hospitalized, she must pay out-of-pocket or through Medicaid, if an individual has used up (“spent down”) her own assets first (not including a family home and other exclusions). A private room in a nursing home averaged \$90,000 a year in 2016 (Longtermcare.gov, 2018), so those paying out-of-pocket soon run out of money. Long-term care insurance covers nursing home care, but few Americans have this insurance (Kovner and Knickman 2011) because it is expensive and only rarely subsidized.

## Palliative Care

Palliative care is the care of persons with a terminal illness. It entails the relief of pain and other symptoms to make the person comfortable and psychosocial and spiritual support (Field and Cassel 1997). Hospice services are an integral part of palliative care and were delivered to 1.6 million persons in 2009, mostly older persons and those with cancer (Shi and Singh 2012; NHPCO 2010). In 2010, 32% of Medicare decedents older than 65 years received care from a Medicare-certified hospice (Aldridge et al. 2015).

Medicare, Medicaid in most states, and most private insurance plans cover hospice. Due to the fact that most hospice care is for the elderly, and

the elderly are fully covered by Medicare, the number of uninsured individuals needing hospice care is quite small (Lorenz et al. 2003). For the small number of individuals without insurance coverage, hospices may provide care regardless of ability to pay (Pietroburgo 2006).

---

## Reforms

The Patient Protection and Affordable Care Act (ACA) constitutes one of the most important reforms to the US health system to date. The ACA was signed into law in 2010 and was implemented over several years. Its scope is very broad, and while its principle goal was to increase access to health services through the expansion of both private and public insurance, it also included measures to improve quality and to control costs. In the version of the ACA signed into law, almost everyone was required to have insurance; this is called the “individual mandate.” There were penalties for failure to have insurance, but exemptions apply (e.g., religious objection, inability to pay). However, in 2017, the individual mandate to purchase insurance was repealed by Congress—individuals will no longer be required to purchase coverage beginning in 2019. Sliding scale subsidies help individuals and families purchase required private health insurance coverage through health-care exchanges. For example, a family of four (all nonsmokers) with a very-low-income level of \$23,550 in 2014 received a tax credit to cover 95–100% of its insurance premiums if purchased on a government-sponsored health insurance exchange officially called the Marketplace. The same with an income of \$40,000 per year received a tax credit worth 77% of the total cost of their health insurance. They had to pay \$161 per month or about 5% of their annual income for health insurance. If this family’s income reached 400% of the FPL or around \$95,000 per year, they had to purchase insurance without any subsidy. They paid about 9% of their annual income for health insurance. For a given amount of coverage offered by a particular private insurer, premiums can vary by rating area (i.e., geographical location), age,

family size, and tobacco use. A calculator available on the health insurance exchange website allows those seeking insurance to determine the approximate of subsidy they will receive (Kaiser Family Foundation 2018d).

Health insurance exchanges have been set up by states or the federal government to make it easier for consumers to compare and choose health insurance policies by providing information in a standardized form. Policies are regulated as to what they must cover. Insurers selling through the exchanges cannot reject an applicant due to health status nor can they charge more to those with a history of preexisting medical conditions. Premiums can, however, vary based on age, smoking status, and geographic location. No annual or lifetime limits can be placed on the value of insurance coverage. There are also limits on the percent of premiums insurers must use for the health benefits of those who purchase policies.

The ACA also sets Medicaid eligibility standards which were more generous than those in effect in many states. The law made the federal government responsible for most of the cost of this expansion of Medicaid (90–100%) in states that were below the new national standard. However, as a result of the Supreme Court ruling in 2011, states were given the option of not expanding Medicaid. As of early-2018, 32 states and D.C. have expanded Medicaid with the others working on waivers or not taking action at this time (Kaiser Family Foundation 2018c, 2018d). They may, however, choose to participate in subsequent years. In June of 2015, the Supreme Court ruled on the *King v. Burwell* case. King challenged the constitutionality of federal subsidies awarded to those purchasing health insurance on federal insurance exchanges. When the ACA was drafted and adopted into law, wording indicated that subsidies would be available to those who enrolled in an exchange “established by the state,” and King argued that the federal exchanges were not established by a state and therefore they could not offer subsidies. The case was critical to the survival of the ACA because initially most states (34) failed to establish their own exchange. The federal government had stepped in to set one up in each of these states. In some cases the

federal government was invited to do this by the state itself, but in other cases the state refused to set up their own exchange as a means to protest against the ACA. The Supreme Court sided with the Obama administration (Burwell) and ruled that the intent of Congress had been to provide subsidies on all exchanges across the USA.

Medicare benefits were enhanced by the ACA. Preventive services are covered without a co-payment from the patient. Over time, the coverage gap (“doughnut hole”) for prescription drug coverage is being removed. Medicare Advantage plans (private out-sourced forms of managed care Medicare) are experiencing reductions in how much they are paid by the federal government to take care of Medicare patients because of evidence that they have been paid much more than their costs in the past. Those achieving higher-quality scores for care receive bonuses and those with lower scores, financial penalties.

Employers with 50 or more employees must offer health insurance, or face a penalty. This mandate became effective in 2015. Employers with fewer employees do not have to provide coverage. Some small employers receive tax credits to offer coverage.

Providers who choose to organize into ACOs have the opportunity to share in any savings they accrue, initially from Medicare but eventually other payers may participate as well. The ACA includes experiments with innovative payment systems that avoid the problems inherent in fee-for-service reimbursement. Bundled service payments are an example. Scholarships and loans included in the ACA are intended to encourage more primary care physicians to work in underserved rural and urban areas. Cost control policies in the ACA included the formation of an Independent Payment Advisory Board to keep Medicare spending in-line with economic growth. Additionally, while the ACA forbids the use of cost-effectiveness research in determining service coverage and reimbursement under Medicare, the law established the Patient-Centered Outcomes Research Institute to spur comparative effectiveness research in the health-care sector.

The ACA was designed to be budget neutral. To help pay for the ACA, high-income individuals



and families pay higher taxes on unearned and investment income, and they pay higher payroll taxes to finance Medicare. A tax was added to some medical devices and to services offered by tanning salons. There is also a tax on “Cadillac” or high-benefit health insurance plans offered by employers, although numerous postponements in Congress have delayed levying the tax until at least 2020. In the end the ACA is redistributive from the healthier to the sicker and from the wealthier to the poorer.

The ACA was adopted by a small margin in the Congress and opposition to this reform remains strong. But today it is the law and it is unlikely that it will be completely reversed. Voters and stakeholders become accustomed to the benefits they receive and removing them is increasingly difficult as time passes. Revisions to the ACA will be ongoing; health system reform is never final. New legislation may be necessary to resolve dilemmas that were overlooked or impossible to resolve at the time the ACA was adopted by Congress. While the current Republican President Donald J. Trump made repealing and replacing the ACA a central focus of his 2016 presidential campaign, widespread opposition to repealing the benefits of the ACA undermined efforts to remove some of its protections. Nonetheless, Congress repealed the individual mandate to purchase health insurance (effective in 2019) in addition to other legislative strategies to reduce ACA protections, including a 2017 Executive Order by President Trump for agencies to explore options that would expand short-term health insurance and other less-comprehensive forms of health coverage, relax rules about associations offering less comprehensive coverage to members, shorten the sign-up period for individual coverage, reduce outreach for enrollment for individual coverage, and attempt to cut spending on federal subsidies offered to help individuals purchase health insurance through the federal exchange. Despite these efforts, and the uncertainty and increased costs they created in many state exchanges, enrollment in the exchanges fell only 5% in 2018 compared to the previous year (Kaiser Family Foundation 2018a). This suggests that the popularity of the expanded coverage afforded by the ACA endures,

creating challenges as legislators from both parties try to shape the U.S. health care system moving forward.

---

## Assessment

### Overview

The US health system has both considerable strengths and notable weaknesses. These are discussed in the following sections in the context of access, quality and outcomes, and expenditures from the USA and international perspectives.

### Access

In 2013, just prior to the main provisions of the ACA being implemented, it was estimated that 44.6 million Americans under the age of 65 (16.7%) were uninsured (US Department of Health and Human Services 2014, Table 114). This rate had been relatively steady since 2000 except for an uptick during the Great Recession. The distribution of uninsured was skewed toward those who were economically most vulnerable. In 2013, nearly 30% of the non-elderly with incomes below twice the federally designated poverty level were uninsured, compared to just 5% of those whose income exceeded 400% of the poverty level. Coverage varied considerably by race/ethnicity as well. Among those under age 65, about 16% of non-Hispanic whites, 19% of African Americans, and 14% of Asians were uninsured. This compares to 31% of Hispanics/Latinos (US Department of Health and Human Services 2014, Table 114). Poor and near-poor children were the one group that has had increasing insurance coverage over the years. Their uninsurance rate in 2013 was about 7%, less than half that of poor and near-poor parents as well as adults without children. The lower uninsurance rates for poor and near-poor children reflected the success of CHIP.

After nearly 4 years, the 2014 public and private insurance expansions brought about by the ACA have reduced the number of uninsured

considerably. Private health insurance coverage is rising as a result of the employer and individual insurance mandates, coupled with subsidies provided to purchase health insurance. In addition, Medicaid coverage is expanding as program eligibility rules have been loosened in states that accept federal subsidies for expansion. As noted, in those 32 states and D.C., all poor and near-poor persons with incomes up to 138% of the federal poverty level are covered. By the middle of 2016, the uninsurance rate was estimated to have fallen to 9% (28 million) (Kaiser Family Foundation, 2018a).

The ACA also is intended to create more equity between people in like circumstances. This is accomplished in three primary ways. First, where previously about half of poor and near-poor adults (defined here as 138% of the federal poverty level) were ineligible for Medicaid, all such persons are eligible for coverage in the states that have elected to accept federal funding for Medicaid expansion. Second, the great majority of those whose incomes are too high for Medicaid will be insured through subsidized private coverage. Third, individuals with preexisting medical conditions or a history of illness will be eligible to purchase insurance and be able to do so at the same price as others.

In the US, there is a direct relationship between insurance status and having one's usual source of medical care in a physician's office. Generally, those with private health insurance and Medicare have access to physicians' private practices. This is not the case, however, for most of the uninsured and, as mentioned earlier, many persons on Medicaid. Having a usual source of care provides a critical entry into the health-care system through access to primary care, preventive services, and referrals to specialists. In 2013, 76% of women with a usual source of care received mammograms within a 2-year period, and 84% received cervical exams in the past 3 years. For those without a usual source of care, the figures were 30% and 62%, respectively (US Centers for Disease Control and Prevention 2015).

Selected measures of access are discussed next, first for the US and then across countries.

## US Data

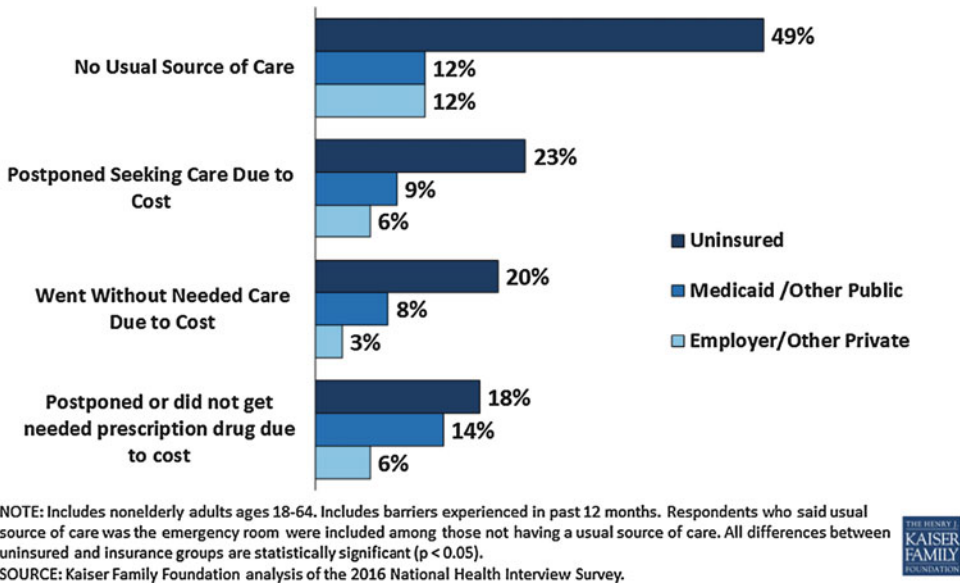
Figure 4 shows the relationship between insurance status and the use of particular services in 2016. The most striking figures relate to having a usual source of care, where 49% of the uninsured report having no usual source of care, versus just about 12% for those with employer coverage or Medicaid (Kaiser Family Foundation 2017b). Among the uninsured, 23% report that they did not obtain needed care due to costs, and 18% say that they could not afford a prescription drug. By comparison, people with Medicaid are roughly half as likely to report these problems, with rates even lower for those with private insurance. These figures demonstrate the critical role that Medicaid plays in facilitating access to care among those with low incomes.

Another impact of being uninsured is the stage at which a person is diagnosed for particular cancers. For melanoma and colorectal, lung, and breast cancers, the uninsured are between two and three times as likely as the insured to be diagnosed at stage III or IV compared to stage I (Kaiser Family Foundation 2012).

## International Comparisons

Comparative international data used in this section are obtained from the Commonwealth Fund, a US-based foundation. Eleven countries were included in the surveys, with samples in each country ranging from approximately 1,000–3,000 (for methodology, see High et al., 2017).

Compared to ten other developed nations included in the survey, access problems due to the cost of medical care are greater in the USA. Table 4 examines sicker adults (those in poor health, having received surgery or hospitalization in the past 2 years, or received care for a chronic illness, injury, or disability in the past year). The table shows five access problems that result from costs, where in each case, Americans had greater problems than those in other countries. To illustrate, the table shows that 33% of Americans had problems accessing medical care due to costs in the past year. The next highest figures were 22% (Switzerland) and 17% (France). In sharp



**Fig. 4** Barriers to health care among non-elderly adults by insurance status, 2016 (Kaiser Family Foundation 2017b)

contrast, the figure was just 7% in the UK and in the Germany (High et al. 2017).

A final set of metrics regarding access regards in how timely of a manner care is received. Table 5 shows several indicators of waiting times in 11 high-income countries. The US performed well internationally with regard to seeing a specialist and getting elective surgery, with Germany and France performing best and Norway and Canada worst. The picture is different for primary care. The US ranked 8 out of the 11 countries for seeing a doctor or nurse on the same or next day. This is not surprising. Access to specialty care and surgery is relatively high because there are ample resources and few restrictions on what and how much medical equipment hospitals, other health facilities, and physicians can purchase and own. In contrast, primary care efforts in the US fall behind many other high-income countries (Starfield and Shi 2002).

**Outcomes and Quality**

The US performs well on some measures of quality and outcomes from an international

perspective, while it does not perform so well on others. Performance on some of these measures is discussed next.

**Mortality**

US life expectancy at birth was 81.2 years in 2015 (Worldbank 2015). It tied for 26th out of the 32 high-income OECD countries, at about 2 years below the median. With respect to infant mortality, US rates have declined substantially over the past two decades but not as fast as other countries. As a result, it ranks the highest among the 31 high-income OECD countries in infant mortality (OECD 2015).

Amenable mortality is defined as “premature deaths from causes that should not occur in the presence of timely and effective health care” (Nolte and McKee 2011). Figure 5, adapted from a 2017 Commonwealth Fund report, illustrates that in the 2014 period, the USA had the highest amenable mortality rate among all countries, nearly double that of Switzerland, the country with the lowest figure (Schneider et al. 2017). Typical explanations for the poor US performance compared to other countries with respect to mortality rates

**Table 4** Cost-related access problems in 11 high-income countries

	Raw scores (%)											
	Source	AUS	CAN	FRA	GER	NETH	NZ	NOR	SWE	SWIZ	UK	US
Overall benchmark ranking	2016	2	9	10	8	3	4	4	6	6	1	11
Had any cost-related access problem to medical care in the past year	2016	14	16	17	7	8	18	10	10	22	7	33
Skipped dental care or check up because of cost in the past year	2016	21	28	23	14	11	22	20	20	21	11	32
Insurance denied payment for medical care or did not pay as much as expected	2016	9	14	24	8	8	2	2	2	12	1	27
Patient had serious problems paying or was unable to pay medical bills	2016	5	6	23	4	7	5	8	5	11	1	20
Doctors report their patients often have difficulty paying for medications or out-of-pocket costs	2015	25	30	17	13	52	30	3	6	9	12	60
Out-of-pocket expenses for medical bills more than \$1,000 in the past year, US\$ equivalent	2016	16	15	7	5	7	7	13	4	46	4	36

Source: (High et al. 2017)

include “a high rate of uninsured and a fragmented delivery system with relatively weak primary care and poor coordination of care between providers and sites” (Schoenbaum et al. 2011).

**Objective Measures of Quality**

There exist voluminous data on outcomes and quality of care in the US. The discussion is divided into three sections: prevention and screening, cancer survival rates, and asthma admissions.

**Table 5** Timeliness of care in 11 high-income countries

	Raw scores (%)											
	Source	AUS	CAN	FRA	GER	NETH	NZ	NOR	SWE	SWIZ	UK	US
Last time needed medical attention was able to see doctor or nurse the same or next day	2016	67	43	56	53	77	76	43	49	57	57	51
Very or somewhat difficult to get medical care in the evening, weekend, or on a holiday without going to the emergency room (base, sought after-hour care)	2016	44	63	64	64	25	44	40	64	58	49	51
Waiting time for emergency care was 2 h or more (base, used emergency room in past 2 years)	2016	523	50	9	18	20	30	34	39	26	32	25
Waiting time to see a specialist was 2 months or more (base, saw or needed to see a specialist in past 2 years)	2016	13	30	4	3	7	20	28	19	9	19	6
Waiting time of 4 months or more for elective/nonemergency surgery (base, those needing elective surgery in the past year)	2016	8	18	2	0	4	15	15	12	7	12	4

Source: (High et al. 2017)



Source: European Observatory on Health Systems and Policies (2017). Trends in amenable mortality for selected countries, 2004 and 2014. Data for 2014 in all countries except Canada (2011), France (2013), the Netherlands (2013), New Zealand (2012), Switzerland (2013), and the U.K. (2013). Amenable mortality causes based on Nolte and McKee (2004). Mortality and population data derived from WHO mortality files (Sept. 2016); population data for Canada and the U.S. derived from the Human Mortality Database. Age-specific rates standardized to the European Standard Population (2013).

**Fig. 5** Mortality amenable to health care (Source: Adapted from Schneider et al. 2017). Data from: European Observatory on Health Systems and Policies (2017). Trends in amenable mortality for selected countries, 2004 and 2014. Data for 2014 in all countries except Canada (2011), France (2013), the Netherlands (2013), New Zealand (2012), Switzerland (2013), and the U.K. (2013). Amenable mortality causes based on Nolte and McKee (2004). Mortality and population data derived from WHO mortality files (Sept. 2016); population data for Canada and the U.S. derived from the Human Mortality Database. Age-specific rates standardized to the European Standard Population (2013).

Switzerland (2013), and the U.K. (2013). Amenable mortality causes based on Nolte and McKee (2004). Mortality and population data derived from WHO mortality files (Sept. 2016); population data for Canada and the U.S. derived from the Human Mortality Database. Age-specific rates standardized to the European Standard Population (2013).

Unless otherwise noted, all data are from OECD (2015).

**Prevention and Screening:** The US immunization rates in 2015 were diphtheria, tetanus, and pertussis, 84.6%; measles, 91.9%; hepatitis B, 92.6%, and influenza, 67%. The US is among the lower half of countries for DTP, measles, and hepatitis B. It is, however, among the countries with the highest rates for influenza vaccination. With regard to screening rates for breast cancer (mammography) and cervical cancer (Pap smears), of the 14 countries OECD compared, the US has the second highest mammography (cancer screening) rate for women age 50–69, at 81% (after the Netherlands) among 12 countries, and (among 11 countries) the highest cervical cancer screening rate for women age 20–69, at 85%.

**Cancer Survival:** Cancer survival is often considered a good measure of the quality of a medical care system because high survival rates are related both to preventive (screening) care and to treatment success. The US has been very successful

with regard to breast cancer treatment, in part due to the high mammography screening rates. The 5-year survival rate, 89%, is highest of 18 OECD countries. The US survival rate for cervical cancer of 62%, in contrast, is the third lowest of the 18 countries. In contrast, for colorectal cancer, with a 5-year survival rate of 64%, the US ranks in the top third of the countries.

**Asthma Admissions:** The hospital admission rate for asthma in the US is among the highest among the 32 high-income OECD countries, at 89.7 per 100,000 population, with only the Slovak Republic and Korea higher. This is likely the result of a high uninsurance rate and poor preventive care.

### Subjective Measures of Quality

The leading source of these data for international comparisons is the Commonwealth Fund, using annual surveys of patients or physicians that have been conducted in up to 11 countries since 2007. The 2011 survey focused on adults with a history of illness, while the 2013 survey examined

nationally representative samples of all adults. The data below are from the 2014 report (Davis et al. 2014).

With regard to care coordination, compared to the other countries, sicker adults in the US had among the highest rates of problems with test results or records not being available when they saw their doctor as well as having duplicate tests ordered. One area in which the US did well was patients receiving a written plan for care after hospital discharge or surgery – at 92%, well higher than the other ten countries.

Five metrics of patient safety are shown in Table 6: that the patient believes there was a medical mistake made in treatment, received the wrong medication or dose, that there were incorrect test results, there were delays in obtaining abnormal test results, and those hospitalized reported an infection from the hospital stay. For the first four measures, the US ranked near the bottom in patient safety among the 11 countries. However, for the last measure (hospital infections), the US figure was the best (Davis et al. 2014).

### Equity of Outcomes

The US suffers from major inequities or disparities in access to health care as well as in health outcomes. A few of the more noteworthy disparities are discussed here (unless noted, all figures are from the US Department of Health and Human Services (2016)). Beginning with infant mortality, the overall rate in 2015 was 5.9 deaths per 1000 live births. The rates for both whites (4.9) and Hispanics/Latinos (5.01) are considerably higher than they are for Asian/Pacific Islanders (3.7). The rate for African Americans, however, is more than double that of whites, at 10.9. The infant mortality rate for American Indians and Alaskan Natives is also considerably high at 7.7, higher than the rate for whites, Hispanics and Asians. Infant mortality also varies considerably by state, with the rate in Massachusetts (4.3) about half that in several states in the South. Given the racial differences just noted, it is not surprising that the states with the highest rates tend to have higher proportions of African American residents. Life expectancy at birth shows similar patterns: In 2015, whites had,

on average, a 3.8-year longer life expectancy than African Americans. This gap had narrowed considerably in the recent years, as in 2006, it was 5.1 years.

This disparity between African Americans and other races also holds for certain diseases. Diabetes rates, for example, are 80% higher among African Americans than whites. For end-stage renal disease, African American incidence and prevalence rates are about three times those of whites. There are disparities by income as well. In the case of diabetes, rates for those below 200% of the FPL are twice those of people above 400% of the FPL. While diet and genetic factors play a strong role in diabetes, disparities in treatment relate to both the medical care system itself and access to it. Similarly, there are different cancer survival rates according to race. Overall 5-year survival rates in the 1999–2006 period were 69% for whites compared to 59% for African Americans. Among ten of the most common types of cancer, whites had higher survival rates for nine of them (all but stomach cancer).

One of the stated objectives of the ACA is to improve quality and outcomes. First, preventive care is encouraged because such services will not be subject to patient co-payments under Medicare and Medicaid. Medicare will also cover one comprehensive risk assessment. Second, ACOs, some believe, can increase quality by encouraging coordination of currently disparate providers and discouraging the provision of unnecessary services. Third, additional comparative effectiveness research will be funded, and fourth, a number of financial incentives based on quality and outcomes are initiated under the legislation. These include reimbursement incentives for hospital performance and value-based payments to providers.

### Expenditures

The US spends far more on health care per person than any other country. There is little agreement on why the US is an outlier in this regard. Those on the left often point to what they see as several contributing factors: lack of consolidated

**Table 6** Measures of patient safety in 11 high-income countries

	Raw scores (%)											
	Source	AUS	CAN	FRA	GER	NETH	NZ	NOR	SWE	SWIZ	UK	US
Overall benchmark ranking												
Patient believed mistake was made in treatment or care in past 2 years	2011	10	11	6	8	11	13	17	11	4	4	11
Patient given wrong medication or wrong dose at a pharmacy or hospital in past 2 years	2011	4	5	6	8	6	7	8	5	2	2	8
Patient given incorrect results for a diagnostic or lab test in past 2 years (base, had a lab test ordered)	2011	4	5	3	2	6	5	4	3	3	2	5
Patient experienced delays in being notified about abnormal test results in past 2 years (base, had a lab test ordered)	2011	7	11	3	5	5	8	10	9	5	4	10
Hospitalized patients reporting infection in hospital or shortly after	2013	9	11	8	10	12	12	10	8	10	12	5

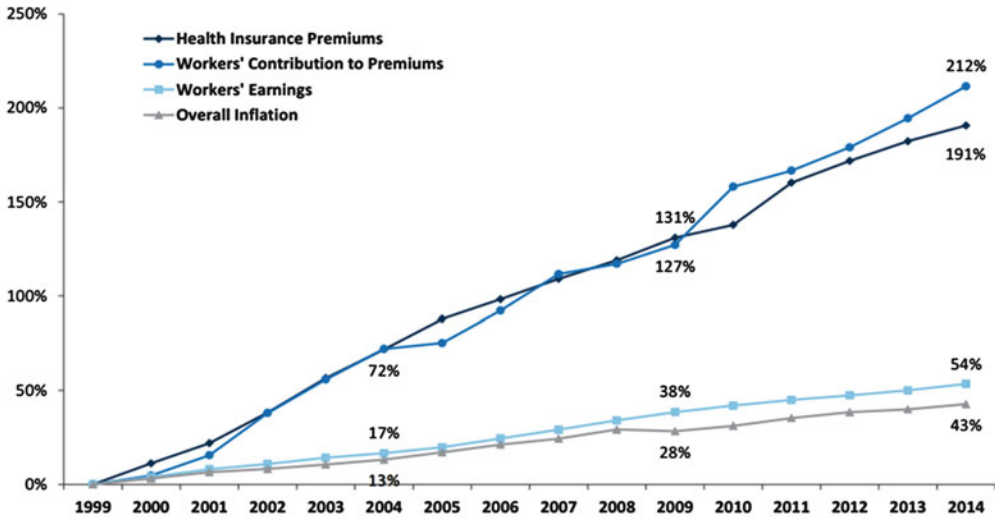
Source: Davis et al. (2014)

purchasing power among buyers of care, the lack of universal insurance coverage, high marketing and administrative costs among private insurers, too many specialists and not enough primary care doctors, and direct-to-consumer advertising of prescription drugs. Those on the right point to a bloated government bureaucracy and a myriad of regulations that stifle competition, along with medical liability laws that encourage over-

provision and overutilization of services. Other factors that observers on both sides point out are high unit prices paid to providers, particularly in the fee-for-service system, proliferation of medical technologies, and unhealthy behaviors.

Per capita spending is more than double the median level for OECD countries, nearly 40% more than the second most expensive country, Switzerland, and health-care expenses constitute





SOURCE: Kaiser/HRET Survey of Employer-Sponsored Health Benefits, 1999-2014. Bureau of Labor Statistics, Consumer Price Index, U.S. City Average of Annual Inflation (April to April), 1999-2014; Bureau of Labor Statistics, Seasonally Adjusted Data from the Current Employment Statistics Survey, 1999-2014 (April to April).



**Fig. 6** Cumulative increases in health insurance premiums, workers’ contributions to premiums, inflation, and workers’ earnings, 1999–2014

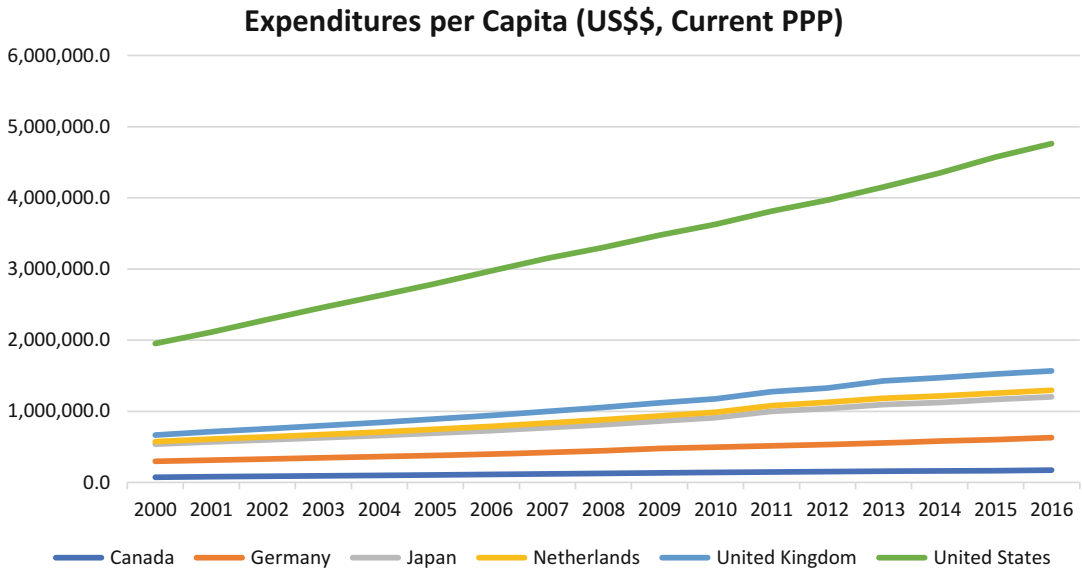
over one-sixth of the US economy (Hartman et al. 2014). The rate of growth in health-care spending exceeded the GDP growth rate every year since at least the 1960s until 2010, which has increasingly squeezed the finances of all levels of government, employers, and individuals.

Employers and employees also have seen large increases in their contributions to the health-care costs of employer-sponsored health insurance. Between 1999 and 2014, total premiums rose by 191% and the workers’ share by 212%. In contrast, wages rose by only 54% over this period (Fig. 6).

Looking now at changes over time, Fig. 7 illustrates growth in national health expenditure per capita expressed in US purchasing power parities for six countries: Canada, Germany, Japan, the Netherlands, the UK, and the US from 2000 to 2016. Growth rates in the Netherlands and Japan exceed those of the other countries. However, in 2016, US spending was more than double that in the UK because the UK started at such a low level of spending. Thus, when one combines both level of spending *and* rate of growth, the US is an international outlier.

There are two overall ways in which the ACA may help contain expenditures. First, it includes a number of initiatives that have the potential to change the financing and delivery system. These include encouraging the development and/or growth of ACOs; bundled payment systems, which provide payment for a set of related services usually related to an episode of illness (as opposed to fee-for-service); medical homes (a physician-directed organization that oversees the provision of access to comprehensive care across health-care facilities and over a patient’s life); electronic medical records; and the linking of reimbursement to performance outcomes (initially, for Medicare hospital stays).

In addition, the ACA includes a number of direct mechanisms that could control expenditures, including large cuts in previously expected payment levels to Medicare Advantage (usually, managed care) plans, which in 2012 were estimated to have been paid 7% more than it would have cost for the same individuals to have been enrolled in the traditional fee-for-service Medicare program (Medicare Payment Advisory



**Fig. 7** National health expenditures, per capita in six countries, 2000–2016 (Source: OECD 2017)

Commission 2012), the tax on “Cadillac” or high-benefit health insurance plans, and the Independent Payment Advisory Board, which is to recommend ways to reduce Medicare costs if they exceed a certain threshold.

The ACA does not include a number of cost-containment methods that have been employed in some other countries. These include global budgets, coordinating provider payment among public and private insurers (i.e., an “all-payers” system), controlling the supply of resources (e.g., through expenditure targets or technology controls), and using cost-effectiveness research to determine which services should be reimbursed and, if so, how much.

## Conclusions

In summary, the US health-care system is among the best in the world in some respects while suffering from significant shortcomings in others. The US is distinguished from its counterparts by its historic distaste for health planning, lack of control over the dissemination of medical technologies, reluctance to take advantage of the potential bargaining power

afforded through large government insurers, the lack of a centralized prices and prospective budgeting, and, most importantly, the absence of guaranteed insurance coverage.

With the adoption of the Affordable Care Act in 2010, and subsequent legal and policy challenges to its core provisions, the US health care system continues to change. Nonetheless, despite many legal and political challenges, the core provisions of the ACA have endured. The ACA addresses major challenging issues such as geographic variation in the use of services and a bias toward subspecialty rather than primary care services but mainly through small programs and pilot studies. The types of changes needed in health-care delivery are unlikely to result from legislation. Rather, they need to be innovated and supported by both the public and private sectors as each grapples with the cost, quality, and access issues they face. They also hinge on changing individual and provider behaviors. Solving the most vexing health-care financing, delivery, and policy issues depends as much on finding a common ground among US policymakers and, more broadly, the American public, as it does on medical, social, behavioral, and organizational sciences.

## References

- AACN. Nursing Faculty Shortage: American Association of Colleges of Nursing Fact Sheet. 2017. <http://www.aacnursing.org/News-Information/Fact-Sheets/Nursing-Faculty-Shortage>
- Adler-Milstein J, DesRoches CM, Furukawa MF, Worzala C, Charles D, Kralovec P, Jha AK. More than half of US hospitals have at least a basic EHR, but stage 2 criteria remain challenging for most. *Health Aff (Proj Hope)*. 2014;33(9):1664–71.
- Aiken L. US nurse labor market dynamics are key to global nurse sufficiency. *Health Serv Res*. 2007;42(3):1299–310.
- Aldridge MD, Canavan M, Cherlin E, Bradley EH. Has hospice use changed? 2000–2010 utilization patterns. *Med Care*. 2015;53(1):95–101.
- Berkowitz SA, Seligman HK, Choudhry NK. Treat or eat: food insecurity, cost-related medication underuse, and unmet needs. *Am J Med*. 2014;127(4):303.e3–10.e3.
- BHPr. The physician workforce: projections and research into current issues affecting supply and demand. BHPr, HRSA, U.S. DHHS. Dec 2008. <http://bhpr.hrsa.gov/healthworkforce/reports/physwffissues.pdf>. Accessed 19 Apr 2013.
- BHPr. The registered nurse population: findings from the 2008 National Sample Survey of Registered Nurses. BHPr, HRSA, U.S. DHHS. 2010. <http://bhpr.hrsa.gov/healthworkforce/rnsurvey2008.html>. Accessed 19 Apr 2013.
- BLS. Current population survey. Bureau of Labor Statistics, Department of Labor. 2011a. <http://www.bls.gov/cps/home.htm>. Accessed 19 Apr 2013.
- BLS. Occupational outlook handbook, 2010–11 ed. Bureau of Labor Statistics, Department of Labor. 2011b. <http://www.bls.gov/oco/>. Accessed 19 Apr 2013.
- Bodenheimer T. Coordinating care: a perilous journey through the health care system. *N Engl J Med*. 2008;358:1064–71.
- Bodenheimer B, Pham HH. Primary care: current problems and proposed solutions. *Health Aff*. 2010;29(5):799–805.
- Brody H, Light DW. Efforts to undermine public health: the inverse benefit law: how drug marketing undermines patient safety and public health. *Am J Public Health*. 2011;101(3):399–404.
- Budetti PP. Market justice and U.S. health care. *JAMA*. 2008;299(1):92–4.
- California HealthCare Foundation. California health care almanac. 2009. <http://www.chcf.org/~media/MEDIA%20LIBRARY%20Files/PDF/E/PDF%20EmployerBenefitsSurvey09.pdf>. Accessed 19 Apr 2013.
- Centers for Medicare and Medicaid Services. Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. 2016. <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/ReportsTrustFunds/Downloads/TR2016.pdf>. Accessed 5 Apr 2018.
- CMS. What's an ACO? Centers for Medicare and Medicaid Services web page. 2012. <https://www.cms.gov/ACO/>. Accessed 19 Apr 2013.
- CMS. National health expenditure data. 2014. <http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/index.html>
- COGME. Physician workforce policy guidelines for the United States, 2000–2020. Washington, DC: Committee on Graduate Medical Education; 2005. [www.cogme.gov/16.pdf](http://www.cogme.gov/16.pdf). Accessed 19 Apr 2013.
- Cohen RA, Martinez ME. Health insurance coverage: early release of estimates from the National Health Interview Survey. Jan–Mar 2015. <http://www.cdc.gov/nchs/data/nhis/earlyrelease/insur201508.pdf>. Accessed 6 Aug 2015.
- Congressional Budget Office. Insurance coverage provisions of the Affordable Care Act – CBO's April 2014 baseline. 2014. <https://www.cbo.gov/sites/default/files/cbofiles/attachments/43900-2014-04-ACATables2.pdf>. Accessed 19 Apr 2013.
- Conti R, Busch A, Cutler D. Overuse of antidepressants in a nationally representative adult patient population in 2005. *Psychiatr Serv*. 2011;62(7):720–6.
- Cram P, et al. Uncompensated care provided by for-profit, not-for-profit, and government-owned hospitals. *BMC Health Serv Res*. 2010;10:90.
- Cunningham PJ. Beyond parity: primary care physicians' perspectives on access to mental health care. *Health Aff*. 2009;28:w490–501.
- Danzon PM, Pauly MV. Insurance and new technology: from hospital to drugstore. *Health Aff*. 2001;20(5):86–100.
- Davis K, Stremikis K, Squires D, Schoen C. Mirror, mirror on the wall, 2014 update: how the U.S. health care system compares internationally. New York: Commonwealth Fund; 2014. <http://www.commonwealthfund.org/publications/fund-reports/2014/jun/mirror-mirror>. Accessed 6 Aug 2015.
- Decker S. Two-thirds of primary care physicians accepted new Medicaid patients in 2011–12: a baseline to measure future acceptance rates. *Health Aff*. 2013;32(7):1183–7.
- Ennis SR, Ríos-Vargas M, Albert NG. The Hispanic population 2010, 2010 Census briefs. U.S. Census Bureau. 2011. <http://www.census.gov/prod/cen2010/briefs/c2010br-04.pdf>. Accessed 19 Apr 2013.
- Field MJ, Cassel CK. Approaching death: improving care at the end of life. Washington, DC: National Academies Press, Institute of Medicine; 1997. <http://www.nap.edu/catalog/5801.html>. Accessed 19 Apr 2013.
- Flynn L, Aiken LH. Does international nurse recruitment influence practice values in U.S. hospitals? *J Nurs Scholarsh*. 2002;34(1):67–73.
- Foley DJ, et al. Highlights of organized mental health services in 2002 and major national and state trends. In: Manderscheid RW, Berry JT, editors. *Mental health, United States 2004*. Rockville: Substance Abuse and Mental Health Services Administration; 2004. p. 203, Table 19.2. <http://store.samhsa.gov/shin/content/SMA06-4195/SMA06-4195.pdf>. Accessed 19 Apr 2013.

- Gallup. U.S. Uninsured Rate Steady at 12.2% in Fourth Quarter of 2017. 2017. <http://news.gallup.com/poll/225383/uninsured-rate-steady-fourth-quarter-2017.aspx>. Accessed 8 Feb 2018.
- GAO. Hospital emergency departments: crowding continues to occur, and some patients wait longer than recommended time frames. Washington, DC: US Government Accountability Office; 2009. <http://www.gao.gov/new.items/d09347.pdf>. Accessed 19 Apr 2013.
- GAO. Drug Industry: Profits, Research and Development Spending, and Merger and Acquisition Deals. 2017. <https://www.gao.gov/assets/690/688472.pdf>. Accessed 5 Apr 2018.
- Harrison TD. Consolidations and closures: an empirical analysis of exits from the hospital industry. *Health Econ*. 2007;16(5):457–74.
- Hartman M, et al. National Health Care Spending In 2016: Spending And Enrollment Growth Slow After Initial Coverage Expansions. 2017. *Health Aff*, p.10.1377/hlthaff. <http://www.healthaffairs.org/doi/10.1377/hlthaff.2017.1299>
- Healthcare.gov. Federal Poverty Level. 2018. Available at: <https://www.healthcare.gov/glossary/federal-poverty-level-FPL/>. Accessed 14 Feb 2018.
- Hersh W. A stimulus to define informatics and health information technology. *BMC Med Inform Decis Mak*. 2009;9:24.
- High E, Schneider C, Sarnak DO. Appendix 1. Eleven-Country Summary Scores on Health System Performance. *Mirror, Mirror 2017: International Comparison Reflects Flaws and Opportunities for Better U.S. Health Care*. 2017. [http://www.commonwealthfund.org/interactives/2017/july/mirror-mirror/assets/Schneider\\_mirror\\_mirror\\_2017\\_Appendices.pdf](http://www.commonwealthfund.org/interactives/2017/july/mirror-mirror/assets/Schneider_mirror_mirror_2017_Appendices.pdf). Accessed 18 Feb 2018.
- Hing E, Hsiao C State Variability in Supply of Office-based Primary Care Providers: United States 2012. 2014. US Department of Health and Human Services.
- Hogan SO, Kissam SM. Measuring meaningful use. *Health Aff*. 2010;29(4):601–6.
- Hsiao C, Hing E. Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001–2013. *NCHS Data Brief*. 2014;143:1–8.
- Kaiser Family Foundation. Kaiser slides. 2012. <http://facts.kff.org/>. Accessed 19 Apr 2013.
- Kaiser Family Foundation. Federal Disproportionate Share (DSH) hospital allotments. 2013. <http://kff.org/medicaid/state-indicator/federal-dsh-allotments>. Accessed 11 Oct 13.
- Kaiser Family Foundation. Employer health benefits: 2014 annual survey. 2014a. <http://files.kff.org/attachment/2014-employer-health-benefits-survey-full-report>. Accessed 9 Aug 2015.
- Kaiser Family Foundation. Health Care Expenditures per Capita by State of Residence. 2014b. <https://www.kff.org/other/state-indicator/health-spending-per-capita/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>. Accessed 9 Aug 2015.
- Kaiser Family Foundation. Uninsured Rates for Non-elderly Adults by Gender. 2016a. <https://www.kff.org/uninsured/state-indicator/rate-by-gender/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>. Accessed 8 Jul 2018.
- Kaiser Family Foundation. Health insurance coverage of total population. 2016b. <https://www.kff.org/other/state-indicator/total-population/?currentTimeframe=0&selectedDistributions=medicaid-medicare-other-public&sortModel=%7B%22colId%22>. Accessed 21 Feb 2018.
- Kaiser Family Foundation. Federal Medicaid Disproportionate Share Hospital (DSH) Allotments. 2016c. <https://www.kff.org/medicaid/state-indicator/federal-dsh-allotments/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>. Accessed 21 Feb 2018.
- Kaiser Family Foundation. Health Insurance Coverage of the Total Population. 2016d. <https://www.kff.org/other/state-indicator/total-population/?dataView=1&timeframe=0&selectedDistributions=employer-non-group-uninsured&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>. Accessed 21 Feb 2018.
- Kaiser Family Foundation. Key facts about the uninsured population. 2017a. <https://www.kff.org/uninsured/fact-sheet/key-facts-about-the-uninsured-population/>. Accessed 8 Feb 2018.
- Kaiser Family Foundation. Medicare advantage. 2017b. Medicare advantage. <http://files.kff.org/attachment/Fact-Sheet-Medicare-Advantage>. Accessed 21 Mar 2018.
- Kaiser Family Foundation. The Medicare Part D Prescription Drug Benefit. 2017c. <http://files.kff.org/attachment/Fact-Sheet-The-Medicare-Part-D-Prescription-Drug-Benefit>. Accessed 21 Feb 2018.
- Kaiser Family Foundation. 2017 Employer Health Benefits Survey. 2017d. <https://www.kff.org/report-section/eHBS-2017-summary-of-findings/>. Accessed 21 Feb 2018.
- Kaiser Family Foundation. Health Insurance Coverage of the Total Population. 2018a. <https://www.kff.org/other/state-indicator/total-population/?dataView=0&timeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>
- Kaiser Family Foundation. Status of State Action on the Medicaid Expansion Decision. 2018b. <https://www.kff.org/health-reform/state-indicator/state-activity-around-expanding-medicaid-under-the-affordable-care-act/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>. Accessed 14 Feb 2018.
- Kaiser Family Foundation. Medicaid waiver tracker: Which states have approved and pending section 115 Medicaid waivers? 2018c. <https://www.kff.org/medicaid/issue-brief/which-states-have-approved-and-pending-section-115-medicaid-waivers/>. Accessed 14 Feb 2018.
- Kaiser Family Foundation. Subsidy calculator. 2018d. <http://kff.org/interactive/subsidy-calculator/>. Accessed 18 Feb 2018.

- Kaiser Family Foundation. Marketplace Enrollment, 2014–2018. 2018. <https://www.kff.org/health-reform/state-indicator/marketplace-enrollment-2014-2017/?current-timeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>. Accessed 21 Mar 2018.
- Kidsdata.org. Child population, by race/ethnicity. 2015. <http://www.kidsdata.org/topic/33/child-population-race/table#fmt=144&loc=2,127,347,1763,331,348,336,171,321,345,357,332,324,369,358,362,360,337,327,364,356,217,353,328,354,323,352,320,339,334,365,343,330,367,344,355,366,368,265,349,361,4,273,59,370,326,333,322,341,338,350,342,329,325,359,351,363,340,335&tf=79&ch=7,11,726,10,72,9,939&sortColumnId=0&sortType=asc>. Accessed 3 Aug 2015.
- Kovner AR, Knickman JR. Health care delivery in the United States. 9th ed. New York: Springer; 2011.
- Ling DC, Berndt ER, Frank RG. Economic incentives and contracts: the use of psychotropic medications. *Contemp Econ Policy*. 2008;26(1):49–72.
- Longtermcare.gov. Costs of Care. 2018. <https://longtermcare.acl.gov/costs-how-to-pay/costs-of-care.html>. Accessed 5 Apr 2018.
- Lorenz K, et al. Charity for the dying: who receives unreimbursed hospice care? *J Palliat Med*. 2003;6(4):585–91.
- Medicare Payment Advisory Commission. Health care spending and the Medicare program. 2012. <http://www.medpac.gov/documents/Jun12DataBookEntireReport.pdf>. Accessed 19 Apr 2013.
- Medicare.gov. Your Medicare Coverage. 2018a. Centers for Medicare and Medicaid Services. <https://www.medicare.gov/coverage/hospital-care-inpatient.html>. Accessed 18 Feb 2018.
- Medicare.gov. Part B Costs. 2018b. Centers for Medicare and Medicaid Services. <https://www.medicare.gov/your-medicare-costs/part-b-costs/part-b-costs.html>. Accessed 21 Mar 2018.
- Mehrotra A, Wang M, Lave J, Adams J, McGlynn E. Retail clinics, primary care physicians, and emergency departments: a comparison of patients' visits. *Health Aff*. 2008;27(5):1272–82.
- Metzl JM, Herzig RM. Medicalisation in the 21st century: introduction. *Lancet*. 2007;369(9562):697–8.
- Milliman Inc. Medicare versus private health insurance: the cost of administration. 2006. [http://www.cahi.org/cahi\\_contents/resources/pdf/CAHIMedicareTechnicalPaper.pdf](http://www.cahi.org/cahi_contents/resources/pdf/CAHIMedicareTechnicalPaper.pdf). Accessed 19 Apr 2013.
- Misurski DA, Lipson DA, Changolkar AK. Inappropriate antibiotic prescribing in managed care subjects with influenza. *Am J Manag Care*. 2011;17(9):601–9.
- NHPCO. NHPCO facts and figures: hospice care in America 2010. National Hospice and Palliative Care Organization. 2010. [http://www.nhpco.org/files/public/Statistics\\_Research/Hospice\\_Facts\\_Figures\\_Oct-2010.pdf](http://www.nhpco.org/files/public/Statistics_Research/Hospice_Facts_Figures_Oct-2010.pdf). Accessed 19 Apr 2013.
- Nolte E, McKee M. Variations in amenable mortality – trends in 16 high-income nations. *Health Policy*. 2011;103:47–52.
- OECD. OECD.Stat. 2015. [http://stats.oecd.org/index.aspx?DataSetCode=HEALTH\\_STAT](http://stats.oecd.org/index.aspx?DataSetCode=HEALTH_STAT)
- OECD. OECD.Stat. 2017. [http://stats.oecd.org/OECDStat\\_Metadata/ShowMetadata.ashx?Dataset=SHA&Coords=%5BLOCATION%5D.%5BDEU%5D&ShowOnWeb=true&Lang=en](http://stats.oecd.org/OECDStat_Metadata/ShowMetadata.ashx?Dataset=SHA&Coords=%5BLOCATION%5D.%5BDEU%5D&ShowOnWeb=true&Lang=en). Accessed 18 Feb 2018.
- Phillips RL, Bazemore AW. Primary care and why it matters for U.S. health system reform. *Health Aff*. 2010;29(5):806–10.
- Pietroburgo J. Charity at the deathbed: impacts of public funding changes on hospice care. *Am J Hosp Palliat Med*. 2006;23(3):217–23.
- Ranasinghe PD. International medical graduates in the US physician workforce. *J Am Osteopath Assoc*. 2015;115(4):236–41.
- RAND. Health care on aisle 7: the growing phenomenon of retail clinics. *RAND Health Research Highlights*. *Clin Sch Rev*. 2010;3(1):10–3.
- Rittenhouse D, et al. Small and medium-size physician practices use few patient-centered medical home processes. *Health Aff (Proj Hope)*. 2011;30(8):1575–84.
- Salinsky E. Governmental public health: an overview of state and local public health agencies, National Health Policy Forum, background paper no. 77. Washington, DC: George Washington University; 2010. [http://www.nhpf.org/library/background-papers/BP77\\_GovPublicHealth\\_08-18-2010.pdf](http://www.nhpf.org/library/background-papers/BP77_GovPublicHealth_08-18-2010.pdf). Accessed 19 Apr 2013.
- Schlesinger M, Mitchell S, Gray B. Measuring community benefits provided by nonprofit and for-profit HMOs. *Inquiry*. 2003;40(2):114–32.
- Schneider EC, et al. Mirror, Mirror 2017. International Comparison Reflects Flaws and Opportunities for Better US Health Care. 2017. Commonwealth Fund. [http://www.commonwealthfund.org/interactives/2017/july/mirror-mirror/assets/Schneider\\_mirror\\_mirror\\_2017.pdf](http://www.commonwealthfund.org/interactives/2017/july/mirror-mirror/assets/Schneider_mirror_mirror_2017.pdf). Accessed 5 Apr 2018.
- Schoenbaum SC, et al. Mortality amenable to health care in the United States: the roles of demographics and health systems performance. *J Public Health Policy*. 2011;32(4):407–29.
- Shen Y, Hsia R. Changes in emergency department access between 2001 and 2005 among general and vulnerable populations. *Am J Public Health*. 2010;100(8):1462–9.
- Shi L, Singh DA. Delivering health care in America: a systems approach. 5th ed. Boston: Jones & Bartlett; 2012.
- Starfield B, Shi L. Policy relevant determinants of health: an international perspective. *Health Policy*. 2002;60(3):201–18.
- Tunis SR, Kang JL. Improvement in Medicare coverage of new technology: how Medicare has responded to the need to improve access to beneficial technologies. *Health Aff*. 2001;20(5):83–5.
- U.S. Census Bureau. NAICS 6211, Offices of physicians. 2010. <http://www.census.gov/econ/census02/data/industry/E62111.HTM#bridge>. Accessed 19 Apr 2013.
- U.S. Census Bureau. 2014. <http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk>. Accessed 4 Jul 2015.

- U.S. Census Bureau. Sumter County, Fla., is Nation's Oldest, Census Bureau Reports. 2016. Press Release: CB16-107. <https://www.census.gov/newsroom/press-releases/2016/cb16-107.html>. Accessed 21 Mar 2018.
- U.S. Census Bureau. Quickfacts US, Population estimates 2017. 2017. Available at: <https://www.census.gov/quickfacts/fact/table/US/PST045217#viewtop>. Accessed 21 Mar 2018.
- U.S. Centers for Disease Control and Prevention. Cancer screening and test use – United States, 2013. *Morb Mortal Wkly Rep*. 2015. [http://origin.glb.cdc.gov/mmwr/preview/mmwrhtml/mm6417a4.htm?s\\_cid=mm6417a4\\_w](http://origin.glb.cdc.gov/mmwr/preview/mmwrhtml/mm6417a4.htm?s_cid=mm6417a4_w). Accessed 6 Aug 2015.
- U.S. Department of Health and Human Services. Health, U.S., 2014. 2014. <http://www.cdc.gov/nchs/data/hus/hus14.pdf>. Accessed 19 Aug 2015.
- U.S. Department of Health and Human Services, Health Resources and Services Administration. The registered nurse population: findings from the 2008 National Sample Survey of Registered Nurses. 2010. Retrieved from <http://bhpr.hrsa.gov/healthworkforce/nrsurveys/msurveyfinal.pdf>
- US Department of Health and Human Services. Health, U.S., 2016. 2016. <https://www.cdc.gov/nchs/data/hus/hus16.pdf>. Accessed 8 Feb 2018.
- U.S. Government Accountability Office. Private health insurance: concentration of enrollees among individual, small group, and large group insurers from 2010 through 2013. 2014. <http://www.gao.gov/assets/670/667245.pdf>. Accessed 2 Aug 2015.
- Van der Hoof C, et al. Inappropriate drug prescribing in older adults: the updated 2002 Beers criteria—a population-based cohort study. *Br J Clin Pharmacol*. 2005;60(2):137–44.
- Weiner J. Expanding the US medical workforce: global perspectives and parallels. *BMJ*. 2007;335(7613):236–8.
- Weiner S, et al. Managing the unmanaged: a case study of intra-institutional determinants of uncompensated care at health care institutions with differing ownership models. *Med Care*. 2008;46(8):821–8.
- Weinick RM, Bristol SJ, DesRoches CM. Urgent care centers in the U.S.: findings from a national survey. *BMC Health Serv Res*. 2009;9:79.
- Weissman J, Gaskin DJ, Reuter J. Hospitals' care of uninsured patients during the 1990s: the relation of teaching status and managed care to changes in market share and market concentration. *Inquiry*. 2003;40(1):84–93.
- Whitmore H, et al. The individual insurance market before reform: low premiums and low benefits. *Med Care Res Rev*. 2011;68(5):594–606.
- WHO. The right to health – fact sheet. 2007. [http://www.who.int/mediacentre/factsheets/fs323\\_en.pdf](http://www.who.int/mediacentre/factsheets/fs323_en.pdf). Accessed 19 Apr 2013.
- Williams SJ, Martin P, Gabe J. The pharmaceuticalisation of society? A framework for analysis. *Sociol Health Illn*. 2011;33(5):710–25.
- World Bank. Life Expectancy at Birth, total (years). 2017. <https://data.worldbank.org/indicator/SP.DYN.LE00.IN>. Accessed 21 Feb 2018.
- Yee T, Lechner AE, Boukus ER. The surge in urgent care centers: emergency department alternative or costly convenience? *Center for Studying Health System Change. Res Brief*. (26). July 2013. [www.hschange.com/CONTENT/1366/1366.pdf](http://www.hschange.com/CONTENT/1366/1366.pdf)



Claus Wendt

## Contents

<b>Introduction</b> .....	927
<b>Typologies</b> .....	929
The Role of Actors and Institutions in Healthcare .....	929
How Do Healthcare Systems Work? .....	933
<b>Discussion</b> .....	935
<b>References</b> .....	936

## Abstract

Since the early 1970s, scholars have been working on typologies for the comparison of healthcare systems. Typologies enable scholars to more easily replicate existing studies and contrast findings from a comparative study with those of other studies that cover different years and countries. Typologies might also help to identify institutional indicators that seem to be of particular promise when comparing healthcare systems and reform processes. This contribution provides an overview of health system typologies and can be roughly divided into two areas of research: (1) classifications that focus on modes of governance, actors, and institutions and (2) classifications that try to capture how healthcare is financed, provided, and regulated. This chapter identifies

prominent examples of both areas of research and also describes and characterizes types of healthcare systems and country classifications.

## Introduction

Healthcare systems are characterized by different levels and modes of financing, service provision, and regulation. Various actors are represented in the healthcare arena. Decision-makers may emphasize the relevance of inpatient and outpatient healthcare, prevention, and rehabilitation. Healthcare systems are, in other words, complex institutions that are difficult to capture. Typologies can be used as a tool to compare healthcare systems based on few preselected indicators. A major goal of comparative typology research is to reduce the massive amount of data and typologies as a concept of comparative research method reduce the level of complexity. The strengths of typologies can be seen in offering a conceptual basis for generalizing across highly diverse

C. Wendt (✉)  
University of Siegen, Siegen, Germany  
e-mail: [wendt@soziologie.uni-siegen.de](mailto:wendt@soziologie.uni-siegen.de)

healthcare systems. In healthcare, typologies have so far mainly been used to contrast different types of healthcare systems, to group countries into types, and to identify similarities and differences among countries. Recently, typologies of welfare states and of healthcare systems have been used for combining macro- and micro-research in healthcare. Comparative scholars have, for instance, studied macro-level effects on patients' access to healthcare, health status, and satisfaction.

With respect to the triangular model of healthcare systems as the backbone of this volume, health system typologies generally refer to one, two, or even all three dimensions: financing agencies, healthcare providers, and patients. Typologies can roughly be divided into two areas of research. The first concentrates on actors and institutions by asking *who* finances, provides, and regulates healthcare services. The second area of research is interested in the *what* and captures levels and structures of financing, provision, and regulation.

Typologies are rooted in Max Weber's methodology of ideal-types. According to Weber (1949: 90; italics in original), an ideal-type "is formed by the one-sided *accentuation* of one or more points of view and by the synthesis of a great many diffuse, discrete, more or less present and occasionally absent *concrete individual* phenomena, which are arranged according to those one-sidedly emphasized viewpoints into a unified *analytical* construct." This method can be used as a tool for grouping countries (real-types) into health system types to identify similarities and differences among healthcare systems, analyze changes over time, and study the effects of healthcare systems characterized by different institutional setups.

Social health insurance and National Health Service have been used as terms for contrasting healthcare systems that have different traditions and institutional designs. The first social health insurance (SHI) was implemented in Germany in 1883 by Bismarck; later, countries such as Austria, Hungary, and France followed the German example, and their systems are often labeled as SHI or Bismarckian healthcare systems. The first

National Health Service (NHS) system was introduced in Britain in 1946 on the basis of the Beveridge-plan and provided an example for countries such as New Zealand, Sweden, and Denmark, which have since been labeled as having NHS or Beveridgian healthcare systems (e.g., Hassenteufel and Palier 2007). In a 1987 study by the OECD, the labels National Health Service, social insurance, and private insurance were used to form a more coherent analytical concept for healthcare system comparison. This concept, however, has been criticized for essentially referring to the real cases of Britain, Germany, and the USA instead of ideal-types. While sharing certain characteristics (such as tax financing vs. social insurance financing vs. private financing), the three types are designed neither for covering all developed healthcare systems nor for capturing changes over time. The social insurance countries of Central and Eastern Europe (CEE), for instance, differ in many respects (e.g., the weak position of corporate actors) from West European social health insurance, and Southern European NHS systems differ (e.g., regarding administrative capacities) from the British and Nordic NHS systems.

Due to the lack of a commonly used health system typology, some scholars of healthcare systems use the typology of welfare states introduced by Esping-Andersen (1990) as a reference. In the original version, social democratic welfare states were separated from conservative-corporatist and from liberal welfare states. In healthcare research, welfare regimes have been used for analyzing public support of healthcare systems (Gelissen 2002), the health status (Conley and Springer 2001), and health inequalities (Eikemo et al. 2008). Arguing that the concept of de-commodification (central to the welfare regime typology) is not designed for capturing characteristics with great importance to health, Bamba (2005) introduced a health de-commodification index consisting of private health expenditure, private hospital beds, and the overall coverage of the healthcare system. The country grouping is slightly different compared with the "classic" welfare state typology. Both concepts, however, were



first and foremost designed to capture social rights and do not reveal the main characteristics of modern healthcare systems.

This chapter summarizes some of the more recent health system typologies related to “organization and governance of health systems,” “health financing,” and “provision of services.” It discusses concepts, country groupings, and findings from studies that use health system typologies. Studies are available that analyze the effects of different health system types on cost containment, access to care, public opinion, health, and health inequality (for an overview, see Beckfield et al. 2013; Burau et al. 2015). Most typologies, however, remain descriptive with the primary task of identifying similarities and differences among today’s healthcare systems. Such studies are critical and have commonly been used for selecting countries for small-n comparative studies in the welfare state discourse. Health system typologies have taken data from data sets such as the OECD Health Data (various years), the WHO Health for All Database (various years), and other international and national sources. Most comparative researchers make use of the OECD data, which is, however, less useful for the detection of institutional boundaries within countries or for the analysis of inequalities (Beckfield et al. 2013). If the effects of different health system types are examined, macro data could be matched with micro data from sources such as the Eurobarometer, the European Social Survey (ESS), the International Social Survey Programme (ISSP), and the Survey of Health, Ageing and Retirement (SHARE).

---

## Typologies

Health system typologies can basically be divided into frameworks that concentrate on the role of actors and the type of governance on the one hand and into frameworks that try to understand how healthcare systems work, what they invest in the people’s health, and what services they provide on the other hand.

## The Role of Actors and Institutions in Healthcare

With respect to the triangular concept of this book, typologies in this area of research ask “who” is responsible for governance and organization as well as the purchasing and provision of healthcare services. The 1987 OECD study was one of the first attempts to classify healthcare systems according to preselected dimensions. The analytical dimensions used in OECD 1987 were “coverage,” “financing,” and “ownership,” and the study investigated “who” was responsible in these areas. The creation of types and the classification of countries, however, did not take place on the basis of available comparative data but on the basis of informed reasoning. The OECD study identified (1) a National Health Service (NHS) model with universal coverage, tax financing, and public ownership of healthcare provision; (2) a social insurance model with universal coverage, social insurance financing, and a combination of public and private ownership; and (3) a private insurance model with selective private insurance coverage, private insurance financing, and private ownership (OECD 1987).

Moran (1999) developed a typology of “healthcare states” by asking “who” governs the “consumption,” “provision,” and “production” of healthcare. Governance of “consumption” refers to patients’ eligibility to access healthcare and to the allocation of financial resources to the healthcare system; governance of “provision” refers to the control of doctors and hospitals; and governance of “production” refers to the regulation of medical innovations. On the basis of these dimensions, Moran (1999) constructed four families of healthcare states: (1) the “entrenched command and control state,” in which the state is distinctive in all three governing areas (e.g., the UK and the Scandinavian countries); (2) the “corporate healthcare states,” in which “consumption” is dominated by public law bodies and the field of outpatient healthcare is dominated by panel doctors’ associations (e.g., Germany); (3) “supply states,” which are dominated by provider interests (e.g., the USA); and (4) “insecure command and

control states,” in which administrative capacities are much lower and private healthcare provision is higher than in the first type (e.g., Italy, Greece, Portugal, Spain). By using Moran’s concept, Bura and Blank (2006) analyzed nine healthcare systems and identified four cases that fully fit one of Moran’s types. Sweden and the UK are perfect examples of the “command and control state”; however, New Zealand and the Netherlands share important characteristics of this type as well. Germany represents the “corporatist healthcare state,” and Australia, Japan, and again the Netherlands match this type in two dimensions. The USA is an example of the “supply healthcare state,” and Singapore also shows major characteristics of this type in addition to corporatist elements.

Wendt et al. (2009) suggest a typology with 27 healthcare system types, 3 of which are ideal-types. These healthcare system types are constructed by combining the dimensions of regulation, financing, and service provision with the involvement of state, nongovernmental (societal), and private actors. In “state healthcare systems,” the state is decisive in all three dimensions; in “societal healthcare systems,” societal and corporate actors are decisive; and in “private healthcare systems,” private actors dominate regulation, financing, and healthcare provision. For each ideal-type, Wendt et al. (2009) identified six combinations in which either the state, societal actors, or private actors are dominant in two dimensions and therefore come close to the respective ideal-type. Six additional combinations do not approach to any of the three ideal-types. Based on this typology, Wendt et al. (2009) suggested that the UK and Denmark form “state healthcare systems,” in which the state is decisive in all three dimensions. Germany is classified as a (societal-based) mixed type due to the great importance of private provision, and the USA is labeled a (private-based) mixed type due to the growing importance of public financing through public programs such as Medicare and Medicaid.

Using Wendt et al.’s model, Böhm et al. (2013) compared and classified 30 OECD countries and found 6 health system types for which real cases could be identified: (1) national health service with

regulation, financing, and service provision by state actors and institutions (e.g., the UK, the Scandinavian countries, and the Southern European countries of Portugal and Spain); (2) national health insurance with regulation and financing by the state and with private healthcare provision (e.g., Australia, Canada, Ireland, New Zealand, and Italy); (3) a societal-based mixed type with regulation and financing by societal actors such as social insurance and public healthcare provision (e.g., Slovenia); (4) social health insurance with regulation and financing by societal actors and private healthcare provision (e.g., Austria, Germany, Luxembourg, and Switzerland); (5) a private health system with private regulation, financing, and service provision (e.g., the USA); and (6) etatist social health insurance with state regulation, social insurance financing, and private provision (e.g., Belgium, Estonia, France, the Czech Republic, Hungary, the Netherlands, Poland, Slovakia, Israel, Japan, and Korea). Böhm et al. therefore identified two of the ideal-types proposed by Wendt et al. (a state healthcare system and a private healthcare system), while according to this study, an ideal-type societal healthcare system does not exist in today’s OECD world. While corporate actors such as social health insurance and doctors’ associations can (and sometimes do) run their own services, most healthcare systems that are financed by social health insurance contributions rely on private provision.

Most typologies that concentrate on the role of the state and other actors in healthcare (i.e., “who” is governing and regulating, financing, and providing healthcare) have identified one type of system in which the state plays a dominant role and includes the UK and the Scandinavian countries. Furthermore, in all typologies, the private US healthcare system forms a type of its own. All other empirical and theoretical observations are far from uniform. Most typologies have identified the German healthcare system as representative of a “societal core”; however, while Bura and Blank (2006) cluster the German case together with Australia, Japan, and the Netherlands, Böhm et al. (2013) place Germany in the same group as Austria, Luxembourg, and Switzerland (see Table 1 below).

**Table 1** Overview of health system typologies

Authors	Dimensions	Data	Types	Country grouping	Main goal
The role of actors and institutions in healthcare					
OECD (1987)	Coverage	No data	(1) National health service	“Paradigmatic cases”: (1) The UK (2) Germany (3) The USA	Construction of types
	Financing		(2) Social insurance		
	Ownership		(3) Private insurance		
Moran (1999)	Consumption	No data	(1) Command and control state	(1) The UK, Scandinavian countries (2) Germany (3) The USA (4) Greece, Italy, Portugal, Spain	Construction of types
	Provision		(2) Corporatist state		
	Production		(3) Supply state (4) Insecure command and control state		
Bureau and Blank (2006) (using Moran’s concept)	Consumption	Partly based on OECD health data	(1) Command and control state	(1) Sweden, the UK (New Zealand, the Netherlands) (2) Germany (Australia, Japan, the Netherlands) (3) The USA (Singapore)	Grouping of countries
	Provision		(2) Corporatist state		
	Production		(3) Supply state		
Wendt et al. (2009)	Regulation	No data	Taxonomy of 27 health systems with three ideal-types:	(1) Denmark, the UK (2) (Germany) (3) (The USA)	Construction of types
	Financing		(1) State healthcare system		
	Provision		(2) Societal healthcare system (3) Private healthcare system		
Böhm et al. (2013) (using Wendt et al.’s concept)	Regulation	OECD health data; HiT reports <sup>a</sup>	(1) National health service (regulation, financing, and provision: state)	(1) Denmark, Finland, Iceland, Norway, Sweden, Portugal, Spain, the UK (2) Australia, Canada, Ireland, New Zealand, Italy (3) Slovenia (4) Austria, Germany, Luxembourg, Switzerland (5) The USA (6) Belgium, Estonia, France, the Czech Republic,	Grouping of countries
	Financing		(2) National health insurance (regulation and financing: state; provision, private)		
	Provision		(3) Societal-based mixed type (regulation and financing: societal; provision, state)		
			(4) Social health insurance (regulation and financing: societal; provision, private)		
			(5) Private health system (regulation, financing, and regulation: private)		
			(6) Etatist social health insurance		

*(continued)*

**Table 1** (continued)

Authors	Dimensions	Data	Types	Country grouping	Main goal
			(regulation: state; financing, societal; provision, private	Hungary, the Netherlands, Poland, Slovakia, Israel, Japan, Korea	
How do healthcare systems work?					
Bambra (2005)	Private health expenditure	OECD health data; WHO data	(1) High public healthcare index (50 or higher)	Grouping suggested by the author and based on Bambra (2005), Table 8:	(Construction of types) and grouping of countries
	Private hospital beds		(2) Middle public healthcare index (around 40)		
	Coverage of the public system		(3) Low public healthcare index (20–30)	(2) Austria, Belgium, France, Ireland, New Zealand, Canada, Denmark, Italy	
			(4) Very low public healthcare index (below 10)	(3) Australia, Germany, the Netherlands, Switzerland, Japan	
Reibling (2010)	Gatekeeping	OECD health data; HiT reports <sup>a</sup> ; MISSOC <sup>b</sup>	(1) Financial incentives states	(1) Austria, Belgium, France, Sweden, Switzerland	Construction of types and grouping of countries
	Cost-sharing		(2) Strong gatekeeping and low supply states	(2) Denmark, the Netherlands, Poland, Spain, the UK	
	Supply		(3) Weakly regulated and high supply states	(3) The Czech Republic, Germany, Greece	
			(4) Mixed regulation states	(4) Finland, Italy, Portugal	
Wendt (2009)	Health expenditure	OECD health data; HiT reports <sup>a</sup>	(1) Health service provision-oriented type	(1) Austria, Belgium, France, Germany, Luxembourg	Constructing of types and grouping of countries
	Public-private mix of financing		(2) Universal coverage – controlled access type	(2) Denmark, Italy, Ireland, Sweden, the UK	
	Privatization of risk		(3) Low budget – restricted access type	(3) Portugal, Spain, Finland	
	Healthcare provision				
	Entitlement to care				
	Payment of doctors				
	Patients' access to providers				
Wendt (2014)	Health expenditure	OECD health data; HiT reports <sup>a</sup>	(1) Health service provision-oriented type	(1) Austria, Belgium, Canada, France, Germany, Japan, Luxembourg, New Zealand	Constructing of types and grouping of countries
	Public-private mix of financing		(2) Universal coverage – controlled access type	(2) Australia, the Czech Republic, Denmark, Estonia, Hungary, Ireland, Italy, the Netherlands, Poland, Slovak Republic, Slovenia, the UK	

(continued)

**Table 1** (continued)

Authors	Dimensions	Data	Types	Country grouping	Main goal
	Privatization of risk		(3) Universal coverage – controlled supply type	(3) Finland, Iceland, Portugal, Spain, Sweden	
	Healthcare provision		(4) Low supply type	(4) Greece (in 2001), Israel, Turkey	
	Payment of doctors				
	Patients' access to providers				

<sup>a</sup>HiT reports: European Observatory of Health Care Systems, Healthcare Systems in Transitions series, see <http://www.euro.who.int/en/about-us/partners/observatory/health-systems-in-transition-hit-series>

<sup>b</sup>MISSOC: The EU's Mutual Information System on Social Protection, see <http://ec.europa.eu/social/main.jsp?catId=815&langId=en>

## How Do Healthcare Systems Work?

While the first category of typologies focuses on the role of the state, the second category is mainly interested in how healthcare systems work, in the level of resources invested in healthcare, in the actual process of service provision, and in patients' access to healthcare. In this respect, these typologies of health systems are closer to welfare regime types since whether or not citizens have a right to access certain healthcare services is a key factor. Since a strong focus is placed on what healthcare systems actually do, there is a higher potential for analyzing institutional effects on health outcomes. These types of healthcare systems can (but do not need to) be related to typologies that concentrate on actors and institutions (Marmor and Wendt 2012).

So far, only a few typologies have included selected information on levels of expenditure, financing, healthcare provision, or institutional indicators of the healthcare system. Frenk and Donabedian (1987), for instance, focused on the basis for eligibility (citizenship, contributions, poverty) to access the healthcare system (not shown in Table 1), and the OECD 1987 study used the related question of coverage. Both concepts, however, are placed within the more general concept of governance and regulation. By drawing on the extent of private financing, the level of private provision, and the general access provided by the public healthcare system,

Bambra (2005) developed a healthcare de-commodification index. The main theoretical argument is that patients have easier access to healthcare provision if public coverage is higher and private financing and service provision are lower. A possible country grouping (not provided by Bambra, who was primarily interested in combining cash and service indicators to construct a welfare typology that takes health and social services into account) could juxtapose countries with a high public healthcare index (Finland, Sweden, Norway, the UK), a middle public healthcare index (Austria, Belgium, France, Ireland, New Zealand, Canada, Denmark, Italy), a low public healthcare index (Australia, Germany, the Netherlands, Switzerland, Japan), and a very low public healthcare index (the USA).

Reibling (2010) also used the concept of de-commodification as her starting point but focused more directly on access to welfare programs, whereby access is defined by benefit levels and the conditions by which benefits can be accessed. This focus strengthens the patients' perspective and draws a closer link between healthcare services and individual health. Dimensions for the comparative analysis of access are gatekeeping, cost-sharing, and supply. Gatekeeping is defined as legal regulations that structure patients' entry and passage through the healthcare system (Reibling 2010). Access, furthermore, is influenced by cost-sharing that may create financial incentives not to use healthcare services,

particularly for minor diseases. Supply, as a major precondition for access, is assessed by provider density and medical technology. By using gatekeeping, cost-sharing, provider density (GPs, specialists, and nurses), and medical technology (magnetic resonance imaging units/MRI, computed tomography scanners/CT), four types of European healthcare systems were constructed: (1) “financial incentive states” that regulate patients’ access to medical care first and foremost by cost-sharing (Austria, Belgium, France, Sweden, Switzerland); (2) “strong gatekeeping and low supply states” that are characterized by low cost-sharing (but where access is controlled by extensive gatekeeping), low numbers of healthcare providers, and medical technology (Denmark, the Netherlands, Poland, Spain, the UK); (3) “weakly regulated and high supply states” with low legal access regulation and a high supply of healthcare providers (the Czech Republic, Germany, Greece); and (4) “mixed regulation states” that use both gatekeeping and cost-sharing.

In two publications, Wendt (2009, 2014) additionally focused on gatekeeping, cost-sharing, and supply and combined these dimensions with information on entitlement to healthcare, the level of healthcare expenditure, the public-private mix of healthcare financing, and doctors’ remuneration. Healthcare provision is captured by service provider numbers in inpatient and outpatient healthcare, gatekeeping by a healthcare regulation index, and doctors’ remuneration by the payment of general practitioners in the outpatient sector (fee-for-service, capitation, salary). The 2009 article compares European countries, whereas the 2014 article covers both European and non-European countries. By applying cluster analyses in the 2009 typology, Wendt arrived at three types of healthcare systems:

1. The “health service provision-oriented type,” which captures Austria, Belgium, France, Germany, and Luxembourg. This type is characterized by a high level and unquestioned importance of service provision. Patients often have direct access and a choice of both general practitioners and specialists. Cost-sharing is

comparatively low, and self-employed doctors are generally paid fee-for-service.

2. The “universal coverage – controlled access type,” represented by Denmark, Italy, Ireland, Sweden, and the UK. While these healthcare systems provide universal coverage, access to care is strictly regulated. Patients typically have to sign up on a general practitioner’s list for a longer period of time, and a referral is required if specialist care is needed. Access to care is further restricted by a comparatively low level of healthcare provision in the outpatient sector. General practitioners are mainly paid on a capitation basis.
3. The “low budget – restricted access type,” which includes Finland, Portugal, and Spain. This type of system is characterized by a low level of healthcare expenditure. Patients’ access is controlled not only by strict access regulation but also by high private co-payments. Most general practitioners receive a salary, and the degree of doctors’ autonomy can therefore be considered to be even lower than in the “universal coverage – controlled access type.”

In Wendt (2014), the number of countries was extended, and the research now covers both European and non-European healthcare systems. When using the same dimensions (except entitlement to care) and newer data, the “health service provision oriented type” can be confirmed and now also covers Canada, Japan, and New Zealand. The “universal coverage – controlled access type” has also been confirmed and now additionally includes Australia and countries from Central and Eastern Europe. A third type identified in Wendt (2014) is the “universal coverage – controlled supply state,” represented by Finland, Iceland, Portugal, Spain, and Sweden. In this type, the control of doctors’ remuneration is even stricter, and cost-sharing is even higher than in the “universal coverage – controlled access type.” In the publication from 2014, the “low supply type” has been identified as a fourth type of healthcare system, represented by Israel, Turkey, and (in 2001) Greece. This type is characterized by both very low levels of total health

expenditure and low public financing. Levels of healthcare provision in both inpatient and outpatient healthcare are quite low. Patients' access to medical doctors, however, is hardly controlled by instruments of regulation.

---

## Discussion

The typologies summarized in Table 1 cover two different areas of research. The first group is more focused on types of governance and on the role of the state and other actors in healthcare. The dimensions used are "coverage," "financing," "consumption," and "provision," and the focus is on "who" is responsible in these areas of the healthcare arena. In almost all typologies, Germany (and to some extent Australia, Japan, and the Netherlands), the UK (often together with the Scandinavian countries and to some extent with New Zealand and the Netherlands), and the USA (with no other countries representing this type) are contrasted. Böhm et al. (2013) put forward one of the first empirical classifications of healthcare systems that covers a larger number of countries. Like earlier "role of actors and institutions" typologies, the UK and the Scandinavian countries are grouped into the same type; however, this time they are together with the Southern European countries. Furthermore, Germany is grouped together with Austria, Luxembourg, and Switzerland. This grouping is much in line with arguments laid down in the OECD 1987 study and in Moran's comparative work but has so far not been demonstrated empirically. Two other types that have not been suggested in earlier studies are the "social health insurance type," represented by Australia, Canada, Ireland, New Zealand, and Italy, and the "etatist social health insurance type," represented by countries from Central and Eastern Europe as well as by Belgium, France, the Netherlands, Israel, Japan, and Korea.

The second group of typologies is more focused in "how" healthcare systems work, what services they provide, and how patients access necessary healthcare services. Both areas of research are necessarily interrelated, for the way healthcare systems are governed influences the

way they function. "Command and control states" should be characterized by lower healthcare spending and stronger access regulation. "Supply states," in which doctors' associations and other corporate actors are involved in the governance of healthcare, should be characterized by higher levels of healthcare provision, greater doctors' autonomy, and lower access regulation. However, strong state actors could also use their power and financial capacities to invest more in healthcare. If we want to know how healthcare systems actually work (e.g., for analyzing healthcare systems' effects on health, health inequalities, and healthcare utilization), dimensions with a stronger focus on healthcare provision and patients' access to healthcare providers are required.

The different focus of the two concepts becomes clear when comparing two typologies that include the largest number of countries (see Table 2). We almost always find the Scandinavian countries in the same type of healthcare system, irrespective of whether the focus is on governance (Böhm et al. 2013) or on how healthcare systems work (Wendt 2014). Since the mid-2000s, Portugal and Spain have appeared to be close to the Scandinavian group. Almost all CEE countries can be found in a common type of healthcare system. However, while the form of governance seems to be close to that of some Western social health insurance systems (Belgium, France, the Netherlands) and of the Japanese social health insurance, levels of financing and healthcare provision as well as patients' access to medical care are more similar to the situation in NHS countries such as Denmark, Ireland, Italy, and the UK (see Table 2). The Western social health insurance countries of Austria, Germany, and Luxembourg are similar in both their governance and the way they work. When focusing on levels of financing, healthcare provision, and patients' access, Germany, Austria, and Luxembourg are close to Belgium, Canada, France, Japan, and New Zealand, which, according to Böhm et al., represent different governance types. The USA seems to be distinct from any other type of healthcare system, both in the way it is regulated and in its level of financing, provision, and patients' access to care.

**Table 2** Comparing two typologies

Böhm et al. (2013)	Wendt (2014)
(1) Denmark, Finland, Iceland, Norway, Sweden, Portugal, Spain, the UK	(3) Finland, Iceland, Portugal, Spain, Sweden
(6) Belgium, Estonia, France, the Czech Republic, Hungary, the Netherlands, Poland, Slovakia, Israel, Japan, Korea	(2) Australia, the Czech Republic, Denmark, Estonia, Hungary, Ireland, Italy, the Netherlands, Poland, Slovak Republic, Slovenia, the UK
(2) Australia, Canada, Ireland, New Zealand, Italy	
(4) Austria, Germany, Luxembourg, Switzerland	(1) Austria, Belgium, Canada, France, Germany, Japan, Luxembourg, New Zealand
(3) Slovenia	
	(4) Greece (in 2001), Israel, Turkey
(5) The USA	(5) The USA

This overview of health system typologies suggests that the way healthcare systems are governed does not directly dictate the way they function. Even if very similar actors are involved in the regulation, financing, and provision of healthcare, the results can be very different levels of financing, healthcare provision, and access regulation among individual countries. It is therefore essential to construct health system typologies for both areas of research. It depends on the specific research question at hand what the more useful typological category is. Classifications capturing the role of actors and modes of governance are better suited to analyze reform options, cost containment, and physical and human resource strategies in different health system types, whereas classifications capturing how healthcare systems actually work are better suited for assessing health systems and their influence on health, inequalities in health, and utilization of healthcare services. The triangular model of health systems is of importance for health system typologies not only with respect to the main players and their interactions in the three health markets (the health insurance market, the healthcare purchasing market, and the healthcare provision market) but also with respect to the way patients can use the healthcare system, which is related to factors such as the resources spent on healthcare, cost-sharing arrangements, the level of healthcare services actually provided, and how patients can use these healthcare services.

Health system typologies also have limitations that are in part related to their strength of simplification. The identification of health system types always depends on the indicators chosen, and therefore the selection of indicators and their theoretical justification is key to healthcare system typologies. Furthermore, the correct definition of indicators is not always an easy task. For instance, does health insurance offered by private organizations in the Netherlands, that are highly regulated, count as private or as social health insurance? Also, so far typologies have used national averages that conceal regional differences. Due to the trend of decentralization, future typologies may have to take geographic inequalities into account (Reibling 2010). More generally, according to Freeman and Frisina (2010) and Burau et al. (2015), a trade-off between simplification and accuracy is inherent to typologies.

## References

- Bambra C. Cash versus services: 'worlds of welfare' and the decommmodification of cash benefits and health care services. *J Soc Policy*. 2005;34(2):195–213.
- Beckfield J, Olafsdottir S, Sosnaud V. Healthcare systems in comparative perspective: classification, convergence, institutions, inequalities, and five missed turns. *Annu Rev Sociol*. 2013;39:127–46.
- Böhm K, Schmid A, Götzte R, Landwehr C, Rothgang H. Five types of OECD healthcare systems: empirical results of a deductive classification. *Health Policy*. 2013;113(3):258–69.
- Burau V, Blank RH. Comparing health policy: an assessment of typologies of health systems. *J Comp Policy Anal*. 2006;8(1):63–76.



- Burau V, Blank RH, Pavolini E. Typologies of healthcare systems and policies. In: Kuhlmann E, Blank RH, Bourgeault IL, Wendt C, editors. *The Palgrave international handbook of healthcare policy and governance*. Basingstoke: Palgrave Macmillan; 2015. p. 101–15.
- Conley D, Springer KW. Welfare state and infant mortality. *Am J Sociol*. 2001;107(3):768–807.
- Eikemo TA, Bambra C, Judge K, Ringdal K. Welfare state regimes and differences in self-perceived health in Europe: a multilevel analysis. *Soc Sci Med*. 2008;66:2281–95.
- Esping-Andersen G. *The three worlds of welfare capitalism*. Cambridge: Polity Press; 1990.
- Freeman R, Frisina L. Health care systems and the problem of classification. *J Comp Policy Anal Res Pract*. 2010;12(1):163–78.
- Frenk J, Donabedian A. State intervention in medical care: types, trends and variables. *Health Policy Plan*. 1987;2(1):17–31.
- Glissen J. *Worlds of welfare, worlds of consent? Public opinion on the welfare state*. Leiden: Brill; 2002.
- Hassenteufel P, Palier B. Towards neo-Bismarckian health care states? Comparing health insurance reforms in Bismarckian welfare systems. *Soc Policy Adm*. 2007;41(6):574–96.
- Marmor T, Wendt C. Conceptual frameworks for comparing healthcare politics and policy. *Health Policy*. 2012;107(1):11–20.
- Moran M. *Governing the health care state. A comparative study of the United Kingdom, the United States and Germany*. Manchester: Manchester University Press; 1999.
- OECD. *Financing and delivery of health care. A comparative analysis of OECD countries*. Paris: OECD; 1987.
- Reibling N. Healthcare systems in Europe: towards an incorporation of patient access. *J Eur Soc Policy*. 2010;20(1):5–18.
- Weber M. *The methodology of the social sciences*. New York: The Free Press; 1949.
- Wendt C. Mapping European healthcare systems. A comparative analysis of financing, service provision, and access to healthcare. *J Eur Soc Policy*. 2009;19(5):432–45.
- Wendt C. Changing healthcare system types. *Soc Policy Adm*. 2014;48(7):864–88.
- Wendt C, Frisina L, Rothgang H. Health care system types. A conceptual framework for comparison. *Soc Policy Adm*. 2009;43(1):70–90.

## Further Reading

- Freeman R. *The politics of health in Europe*. Manchester: Manchester University Press; 2000.
- Freeman R, Moran M. Reforming health care in Europe. *West Eur Polit*. 2000;23(2):35–59.
- Gauld R. *The new health policy*. Maidenhead: Open University Press; 2009.
- Gaiimo S, Manow P. Adapting the welfare state – the case of health care reform in Britain, Germany, and the United States. *Comp Pol Stud*. 1999;32(8):967–1000.
- Immergut EM. *Health politics: interests and institutions in Western Europe*. Cambridge: Cambridge University Press; 1992.
- Marmor T, Wendt C, editors. *Reforming healthcare systems. Two Volumes*. Cheltenham/Northampton: Edward Elgar Publishing; 2011.
- Montanari I, Nelson K. Social service decline and convergence: how does healthcare fare? *J Eur Soc Policy*. 2012;23(1):102–16.
- Moran M. Understanding the welfare state: the case of health care. *Br J Polit Int Relat*. 2000;2(2):135–60.
- Rothgang H, Cacace M, Frisina L, Grimmeisen S, Schmid A, Wendt C. *The state and healthcare. Comparing OECD countries*. Basingstoke: Palgrave Macmillan; 2010.
- Smith P, Anell A, Busse R, Crivelli L, Healy J, Lindahl AK, et al. Leadership and governance in seven developed health systems. *Health Policy*. 2012;106:37–49.
- Tuohy C. *Accidental logics: the dynamics of change in the health care arena in the United States, Britain, and Canada*. New York: Oxford University Press; 1999.



# Organization and Governance: Stewardship and Governance in Health Systems

# 42

Scott L. Greer

## Contents

<b>Introduction</b> .....	939
<b>Definitions: Into the Mire</b> .....	940
<b>Comparing and Measuring Governance</b> .....	941
<b>Good Enough, or Better, Governance</b> .....	942
<b>Attributes of Governance</b> .....	943
Transparency .....	943
Accountability .....	944
Participation .....	944
Integrity .....	945
Policy Capacity .....	945
A Diagnostic Approach .....	946
<b>Conclusion</b> .....	946
<b>References</b> .....	946

### Abstract

Governance, how decisions are made and implemented, is an important part of health care and health policy. It is also the subject of a large and often confusing literature. This chapter presents the results of a review of the governance literature for health. First, it notes that not all problems are of governance. Second, it introduces five domains of governance in which governance problems, challenges, and policies are located: Transparency,

Accountability, Participation, Integrity and Capacity. Together they make the TAPIC framework and can be used to identify governance dimensions of policy problems. Third, better governance through the TAPIC model can also reduce the likelihood of other problems.

### Introduction

Stewardship and governance, like “resilience” or “strategic,” are “power words” (Frederickson 2005). They sound desirable, are difficult to argue with, and give an automatic advantage in most arguments to the people who invoke them.

S. L. Greer (✉)  
 Department of Health Management and Policy, University  
 of Michigan, Ann Arbor, MI, USA  
 e-mail: [slgreer@umich.edu](mailto:slgreer@umich.edu)

As a result, both have been stretched by academics, governments, international organizations, consultants, and other ideological entrepreneurs who want the power that comes with its invocation.

This chapter will first separate out stewardship and governance, providing key definitions and making the point that while they might be in the hands of political rivals, they are not intellectually rivalrous concepts. It then presents the results of our review of concepts, presenting the five attributes of governance (which are also among many desirable objectives of stewardship) that emerged as mutually exclusive and able to cover the many activities and ideas classified as “governance.”

---

### Definitions: Into the Mire

*Governance* has several kinds of meaning. On one hand, it has spread across multiple fields that use it in different ways to discuss topics as different as the proper constitution of a company board and the nature of public management in the Internet age. On the other hand, it is used for a variety of normative, empirical, and mixed projects.

While there have been sporadic uses of the word for many years, it became a common modern concept first in the discussion of management, specifically corporate governance, the organization of power within commercial firms. In the 1980s, it started to pick up a second usage; it was used in political economy research to discuss arrangements in which organizations such as unions, professions, and government collectively coordinated activity (e.g., Campbell and Lindberg 1991). In the aftermath of the Cold War, more academics became interested in it as a descriptive term for systems that produced collective decisions without having clear centers of hierarchical power (as distinct, in some once-fashionable formulations, from “government”). In this capacity, the term drew on and partially displaced perfectly good older terms such as “networks.” In the hands of these scholars, governance came to mean almost anything that generated order without hierarchy; its meanings in transaction cost economics (Williamson 1996), European studies (Marks et al. 1996), international relations (Rosenau and

Czempiel 1992), and public management (Rhodes 1997), for example, differed greatly.

International organizations became particularly interested as part of the backlash against structural adjustment lending and, in particular, their role in the Asian financial crisis and its aftermath. Fifteen years of increasingly invasive policy conditionality in the service of structural adjustment failed to produce the desired effects in the structurally adjusted countries (Greer 2013; Woods 2006). They turned to good governance as a solution (e.g., World Bank 1992, 1994). The essential logic was simple enough: reforms, especially those imposed through conditional loans, frequently had serious noncompliance problems, faced serious implementation problems, and had the wrong effects. The response was to blame these problems on the governance – the organization, probity, competence, and coordination – of the countries involved and try to improve that as a part of development or financial rescue (Nunnenkamp 1995).

In 2013, all three preoccupations are alive and well: we have governance as a field of management, including corporate governance and clinical governance in health (Walshe and Smith 2011), governance as a sprawling and contested term applied in endless different ways by social scientists in analyzing the world (Kjaer 2004; Bevir 2013), and governance as a normative concept used when policymakers speak about improving, essentially, international public management (Fukuyama 2013).

In each of these incarnations, governance-speak has two essential uses. One is empirical: the description and analysis of what *is*. One is normative: calls for how it *ought* to be. Empirically, governance in almost any account is some form of authoritative coordination, which means decisionmaking and implementation. Such analyses tend to try to capture the mechanisms by which authoritative decisions are made, analyzing the powers, responsibilities, and coordination of professions, insurers, providers, governments at different levels, and the other actors who make and implement decisions in health systems.

Normatively, governance can be termed good, or better or worse, and the parallel normative,

policy-oriented literature seeks to improve it by promoting, essentially, various forms of “good governance.” In general, this normative literature is focused on policy interventions and institutional changes. The real solution to corruption, social science makes quite clear, is reducing inequality in society by expanding social rights and economic redistribution (Uslaner 2008; Rothstein 2011). That seems to be beyond the scope of most governance advice, which focuses on the level of individuals (hiring the right people) and organizations, and perhaps legal frameworks (Sabet 2012, 21 for the distinction). Many accounts, of course, mix normative and empirical in more or less coherent, articulated, and useful ways.

*Stewardship*, by contrast to governance, is a word with a more limited history in health policy. While the word is as old as the concept of a steward – a person entrusted with looking after something – its grand entrance into the global health policy vocabulary came in the 2000 World Health Report (World Health Organization 2000) (*WHR*), which defined it as one of four key functions of health systems alongside resource generation, financing, and service delivery. The *WHR* defines stewardship as “the careful and responsible management of the well-being of the population,” and “... the very essence of good government” (Travis et al. 2003 for a lucid discussion in the WHO context).

Separating governance and stewardship is conceptually easier than it might look. Firstly, governance is a structure or pattern, whereas stewardship is an activity. As a result, pursuing an item such as capacity or development or transparency from a long list of policies can be seen as good stewardship or establishment of better governance. A person occupying a position in a system of governance can be a better or worse steward. Secondly, stewardship is almost always normative in health policy discussions. Governance in the sense of authoritative coordination exists in almost any functional society (by definition), even if it is not good. Stewardship, by incorporating care, responsibility, good government, and the well-being of the population, makes itself a normative rather than empirical

concept. Thirdly, stewardship was a concept largely confined to global health policy discussions, while governance is, for better or for worse, discussed in many fields of human activity.

---

## Comparing and Measuring Governance

Measuring the quality of governance has been a preoccupation of scholars and international organizations for some years now, and the result has been a variety of initiatives that attempt to define governance in quantitative, comparable terms. Given that the latest initiatives are the most ambitious yet, the next years should be fertile ones for the quantitative, comparative study of governance.

The largest project is based at the University of Gothenburg. The “Quality of Government” project, as it is known, aggregates a wide variety of databases (its key findings are in Rothstein 2011). A variety of other projects, including the Varieties of Government project based at the University of Notre Dame (Coppedge et al. 2012), try to enhance our comparative understanding and measurement of a wide spectrum of governance indicators (Fukuyama 2013 for a review). These databases, which face the data and coding problems of all large-scale international quantitative comparative research efforts, are mostly focused on general regime types and put less focus on the actual management of health systems.

The measurement of health systems *governance* is somewhat less developed, since it is not unintelligent to focus instead on actual health outcomes (an imperfect enough set of outcomes) (Smith et al. 2008). The comparison of health systems and their governance is, by contrast, rather more developed. The European Observatory on Health Systems and Policies has significantly advanced comparative health systems research by producing books, written to templates, on the health systems of every country in the WHO European region and a variety of others. Its Health Systems and Policy Monitor is a regularly updated source of information on health policies, from which much can be learned about governance.

## Good Enough, or Better, Governance

Two words, three broad traditions of their use, a plethora of international comparative enterprises, and both normative and empirical applications: this is a dispiriting starting point for a discussion of how the vocabulary of governance and stewardship may be used to understand or improve health systems.

The first problem to address is the confusion created by political analysts of many stripes, ranging from entrepreneurial consultants to entrepreneurial academics, who sought to distinguish governance as a type of organization from government. This approach defined governance in terms of self-organization, networks, and a blend of public, nongovernmental, and private actors, rather than “government,” which connoted hierarchy, legalism, and inflexibility. The essential distinction was spurious and misleading; networks were hardly new forms of political organization, in the West or anywhere else, and the hierarchical authority of states and other big organizations such as corporations remained very powerful and effective (Bevir 2013). Here, following on current usage and the international institutions, governance is a description of overall decisionmaking and implementation rather than an ideal type rendering of a particular form of public administration.

The next problem is with the concept of “good governance.” If governance can be better or worse, then it seems reasonable to seek to identify and generalize practices of good governance, whether it is corporate governance activists trying to generalize good recruitment practices for boards or international financial institutions trying to generalize good governance for the recipients of their funds. Two difficulties arise. The first is revealed by the syllogism: if governance is how decisions are made and implemented, then good governance is good decisionmaking and implementation throughout a whole society. The likelihood that the same things, defined with any level of specificity, will constitute good governance in every society on earth seems limited (Andrews 2013). Excessive concreteness is a besetting problem in advice about good governance.

The third problem is that governance, being a power word (Frederickson 2005) whose invocation strengthens all sorts of arguments and claims, therefore has had a wide range of attributes added to it. These are often self-contradictory or hard to derive from either data or first principles. For example, some international organizations view “conflict prevention” as an important aspect of good governance, and others do not (Barbazza and Tello 2014). Does this mean that the WHO regards conflict as part of good governance? Obviously not. Rather, what it shows is that lists of attributes of good governance have a tendency to be arbitrary and utopian. Defining the aspects of good governance is tantamount to defining the good society, and that is questionable on matters of taste and practicality.

Notably, few if any systems show all the attributes that have been assigned to “good governance,” and many highly functional systems have aspects of poor governance – opacity, corruption, nepotism, clientelism, and other problems occur in many places. Few if any countries vaulted into high-income brackets while enjoying good governance as many define it today (Greer and Jarman 2011; Brewer et al. 1999), and a few practices we associated with bad governance have proved flexible and effective, for example, clientelism can mean disruption and bad administration by political jobbers but also allows reformers to put technically skilled people into important posts (Grindle 2012).

The problem, therefore, is the one noted by Tolstoy: all happy families are the same, but all unhappy families are different. So many things have to go right to produce a happy family that the variation within the category of happy families is limited. Unhappy families have many more degrees of freedom. And it is in the realm of unhappy families that policy scholars and policymakers must operate. The solution lies in the simple concept of “Good enough governance.” Good enough governance is a concept formulated by Merilee Grindle, who pointed out that many lists of governance attributes have an arbitrary and utopian character (Grindle 2004, 2007; Thomas 2015).

Drawing from this, a more intellectually and practically satisfying approach to governance is to view governance not as a desirable end state but rather as an activity that can be carried out in different ways with different effects. This diagnostic approach views governance as a phenomenon that exists in essentially all societies and sometimes causes a problem for something else. Governance problems can be diagnosed as a reason for policy failures, and strengthening one aspect or another of governance can remedy policy failures. Likewise, some policies are just not sustainable in some systems; governance that is good enough for maintaining basic public health functions might not be good enough to operate sophisticated quasi-markets for health care.

In other words, rather than insistently defining good governance it makes more sense to identify aspects of governance that improve the ability of health systems to achieve a sustainable balance of equity, access, and cost containment. So, then, what are aspects of governance that influence the ability of health systems to achieve their goals, and which can in some cases be improved? Or, on the other side of the coin, what is a governance problem (as distinct from some other kind of problem), and what is a detailed taxonomy of governance problems that might need understanding or remedy?

---

## Attributes of Governance

The first question in using governance analysis to improve policies and systems: is the challenge, or problem, or opportunity one of governance? There are other reasons programs fail. They can be fundamentally bad ideas (though high-capacity, participative, transparent governance might reduce the odds of bad ideas being adopted). They can be underfunded. They can also lack political support.

By a process of elimination, a workable, funded, and supported policy that fails suggests a governance issue. More positively, do problems appear to lie in the decisionmaking and implementation systems of society? If so, that means the problems lie in governance.

More specifically, our review found five key aspects of governance that matter and in many cases can be strengthened. They are not a list of attributes to which every society should aspire; they are, rather, five aspects of health systems that influence the success or failures of policies. One of the remarkable aspects of the governance literature is that, beneath a level of apparent conceptual confusion, the same words and concepts constantly recur. In other words, despite many different terms and many different lists with different inclusions and exclusions, and many different conceptual hierarchies, the same five issues recur. We sorted them into groups with minimal overlap that scholars or policymakers interested in governance should consider (Greer et al. 2016; Greer et al. 2017). The result is the TAPIC framework, for its domains of transparency, accountability, participation, integrity and capacity any of the five might be the first or most important issue, and all can exist relatively independently of each other (accountability without transparency, for example, is the norm in both medical care and automobile repair). The literature review and analysis is presented in (Greer et al. 2016). Case studies exploring and showing the uses of the TAPIC framework can be found in that book, and in (Jarman 2017, Wolfe et al. 2017, Exworthy et al. 2017, Trump 2017, Vasev 2017, Willison 2017 and Greer et al. 2017)

## Transparency

Transparency involves two things: making decisions clear and making clear grounds on which decisions were made (Woods 1999). At a minimum, this means the kind of basic publicity long familiar in functional governments – official notifications, open meetings, and latterly informative websites that make policies and policy processes understandable.

There are a variety of problems with such a simple form of transparency, however; for a start, as every consumer knows, “fine print” can look transparent and effectively hide companies’ actions. Transparency can be taken too far; decisionmaking necessarily involves both deals

and ambiguity, and problems arise if transparency displaces real decisionmaking into shadows or becomes a weapon for those who want to replace argument and prioritization with some more mechanistic (Best 2005). It also has the problem that policy information can be intricate, and efforts to simplify it can also distort it (as frequently happens with both politics and website redesigns). The result is that simple notification should probably be flanked by devices that permit informed access to the policy process so that informed journalists, NGOs, citizens, and experts can contest decisions and their grounds. These mechanisms can include inspectorates, ombuds procedures, public data releases, and freedom of information laws.

Effective transparency should improve policy by enhancing accountability and participation, deterring or quickly identifying corruption and incompetence, and making policies more predictable. The result, in theory, will be trust that an organization will not be erratic and in constant pressure to be competent.

## Accountability

Accountability is a relationship in which an actor (such as a government agency) must account for its actions to a forum (such as a legislature) which can sanction it. In other words, it has three key attributes: actions, reporting, and sanction. A good accountability relationship means that the interests of the forum (legislature, population) is always in the mind of the actor, but the actor has autonomy to formulate superior solutions. It can also allow productive innovation; holding somebody accountable for outcomes within limits rather than process can produce learning and better policy outcomes in general (Sabel 2001; Behn 2001).

Mechanisms that policymakers use to achieve accountability are diverse, including contracts; reporting requirements; financial mechanisms such as pay for performance; laws that specify objectives, reporting, and mechanism; competitive bidding; organizational separation such as purchaser/provider splits; conflict of interest

policies; ombuds processes; legislative oversight and committees of oversight, and regulation including the establishment of dedicated regulatory agencies. Each of these focuses on increasing the extent of reporting and the ability of the forum to sanction the actor.

Accountability is not the same thing as a principal-agent relationship, which favored form of economic modeling. In a public sector principal-agent relationship, a principal chooses an agent to carry out its wishes (Smith et al. 1997; Besley and Coate 2003). Governance, in this analysis, is better insofar as it shortens and clarifies principal-agent relationships. There are two key problems with this style of analysis. The first is that frequently the relationship is hard to characterize in that way – it might actually be a fiduciary model rather than an agency relationship. The second is that it is essentially normative rather than political; it assumes that there should be a clear principal, agent, and instructions. A quick reflection on, for example, the many missions of a hospital shows the empirical limits (Marmor 2001).

## Participation

Participation means that affected parties have access to decisionmaking and power so that they acquire a meaningful stake in the work of an institutions (Woods 1999). Participation has many normatively desirable aspects – it is the basis of democracy, after all – but there is also a pragmatic case for participation of affected parties in decisions that spans political regimes. That is simple: participation helps to reduce or avoid the problems that emerge when key affected groups resist a policy or when a policy is made without knowing what they know. For example, complex medical payment incentive systems do not work as intended if they are made without understanding how doctors work and are paid (a common problem in “pay for performance” schemes). In the worst case, it makes it clear what depth of opposition a policy will face once enacted.

There are a variety of well-established participation mechanisms, as well as a very large and notably confused literature on public participation

in health that rarely explains the point of participation (for a critical discussion Stewart 2013) and some experiments in novel forms of public participation, such as participatory budgeting, whose popularity outside their places of origin is clearer than their effectiveness (Seekings 2013). Established mechanisms of participation include stakeholder forums, public consultations, elections, appointed community representatives on boards, and legal remedies (e.g., legislation that allows aggrieved outsiders to litigate processes). They can also include research, e.g., surveys of local opinion about a given option. When affected bodies are other governments or organizations, advisory committees, partnerships, joint budgets, and special forums for consultation are effective mechanisms for ensuring that different governments will be aware of decisions and make their views clear.

The benefit of participation is the potential creation of “ownership,” i.e., a sense among affected parties that they have a stake in the success of an initiative. Without ownership, there is a real risk of sabotage, lassitude, or simple ignorance, all of which amount to implementation failure. There is also the potential benefit of increased legitimacy – the sense that decisions are taken in ways that reflected the relevant interests.

## Integrity

Integrity is one of many words for the key attributes of a well-run modern bureaucracy: processes of representation, decisionmaking, and enforcement should be clearly specified; all members should be able to understand and predict the processes by which an institution will take decisions and apply them; and individuals should have clear roles and responsibilities. In other words, an organization with a high level of integrity is meritocratic, separates the person and the office, and is not corrupt. These are the bases for well-functioning, long-lasting trustworthy organizations.

Mechanisms policymakers can use to promote or entrench organizational integrity include internal audit (so that money moves as intended and can be

traced), clear personnel policies (regular hiring, job descriptions, and procedures to weed out flawed people), a clear mandate for each organization, a clear and reliable budgeting process, administrative procedures such as document management and minuted meetings, external audit (to put a check on people within the organization), and a clear sense of organizational roles and purposes. Many of these policies, if added together, are bureaucracy – for better or for worse. The challenge of public management is to gain the benefits of bureaucracy in terms of merit, impartiality, and efficiency without risking too much wasted effort or incompetence.

## Policy Capacity

Finally, most accounts of effective health governance include a discussion of policy capacity: the ability to develop policy that is aligned with resources in pursuit of societal goals. Policy capacity is a property of what Edward Page calls the “policy bureaucracy,” that part of an organization, especially a government, whose purpose is to produce policy (Page and Jenkins 2005). Just as a health policy initiative can run into trouble for a lack of medical staff, it can run into trouble for a lack of policy staff who are capable of identifying, synthesizing, and analyzing a wide variety of information in order to spot problems, make the case against ill-considered policies, and work through the procedural and practical challenges of implementation. It can look good to reduce policy capacity – civil servants at the heart of the state do not always have public sympathy – but it can have negative consequences in the form of poorly thought-out policies.

The development and improvement of policy capacity is a central preoccupation of public management scholarship, and the list of tools for doing it is long. It includes mechanisms to produce intelligence on developments in the system and its performance, so that policymakers can identify and react to problems and intelligence on process such as budgetary and legal issues (all too often neglected in health policy analysis), research and analysis capacity (trained staff who can conduct or commission research and deal with literature and



outside experts), staff training (e.g., so that a doctor hired into a health ministry can learn about budgeting and law), strong hiring procedures that balance merit and responsiveness in the central policy bureaucracy, procedures to incorporate experts with their different career structures and incentives, and, all too often forgotten, extensive capacity for purchasing and managing relationship with outsiders such as regulated industries or government contractors. This long list suggests something important: while policy bureaucracies are routinely dwarfed by the systems they manage and they go beyond the minister's immediate office. Civil servants further from the minister, and from the glamor of politics, fulfill an important role and can respond to investment and organizational development.

### A Diagnostic Approach

Reading scholarly and grey literature, almost everything framed as a component of good governance or as an attribute of governance in general, can be fitted into these five categories. If we use them as a diagnostic tool (before or after a problem arises), then we can first see if a policy failure, or risk, depends on decisionmaking and implementation and then work out what kind of governance issue exists and might be remedied – if, for example, the problem is of sabotage and poor implementation by excluded interested parties, then greater transparency and participation might be called for. It is less productive to elevate them, or any other framework, into good governance, for the simple reasons that there are tensions between them, all of them can be taken to extremes (e.g., transparency can make productive dealmaking impossible), and not all of them will mean the same thing or have the same salience in every system (e.g., integrity is much less of an issue in Northern Europe than in most of the rest of the world). We can, however, try to use the TAPIC framework for diagnoses not just of specific policy problems but of policymaking problems. This should in turn reduce the likelihood of unworkable policies being adopted, or workable policies adopted without adequate finance.

### Conclusion

Governance and stewardship might seem like hopelessly fuzzy concepts, but the exercise of grouping the many things said about them reveals five relatively coherent attributes of a health system that are the object of policies for improvement and that can have an effect on the ultimate cost, quality, and access of health.

### References

- Andrews M. *The limits of institutional reform in development*. Cambridge: Cambridge University Press; 2013.
- Barbazza E, Tello JE. A review of health governance: definitions, dimensions and tools to govern. *Health Policy*. 2014;116(1):1–11.
- Behn RD. *Rethinking democratic accountability*. Washington, DC: Brookings; 2001.
- Besley T, Coate S. Centralized versus decentralized provision of local public goods: a political economy approach. *J Public Econ*. 2003;87(12):2611–37.
- Best J. *The limits of transparency: ambiguity and the history of international finance*. Ithaca: Cornell University Press; 2005.
- Bevir M. A theory of governance. In: *A theory of governance*. Berkeley: University of California Press; 2013.
- Brewer J, Hellmuth E, Brewer J. *Rethinking Leviathan: the eighteenth-century state in Britain and Germany*. Oxford: Oxford University Press; 1999.
- Campbell JL, Lindberg LN. The evolution of governance regimes. In: Campbell JL, Rogers Hollingsworth J, Lindberg LN, editors. *Governance of the American economy*. Cambridge: Cambridge University Press; 1991.
- Coppedge M, Gerring J, Lindberg S. V-Dem: varieties of democracy project description. In: *Varieties of democracy project description*. South Bend: University of Notre Dame Kellogg Institute; 2012.
- Exworthy M, Powell M, Glasby J. The governance of integrated health and social care in England since 2010: great expectations not met once again? *Health Policy*. 2017;121(11):1124–1130.
- Frederickson HG. Whatever happened to public administration? Governance, governance everywhere. In: Ferlie E, Lynn Jr LE, Pollitt C, editors. *Oxford handbook of public management*. New York: Oxford University Press; 2005.
- Fukuyama F. What is governance? *Governance*. 2013;26(3):347–368.
- Greer SL. Structural adjustment comes to Europe: lessons for the eurozone from the conditionality debates. *Global Soc Policy*. 2013;14:51.
- Greer SL, Jarman H. The British civil service system. In: van der Meer FM, editor. *Civil service systems in Western Europe*. Cheltenham: Edward Elgar; 2011.

- Greer SL, Wismar M, Figueras J, editors. Strengthening health system governance: better policies, stronger performance. Brussels/Philadelphia: European Observatory on Health Systems and Policies/ Open University Press; 2016.
- Greer SL, Vasev N, Wismar M. Fences and ambulances: Intersectoral governance for health. *Health Policy*. 2017;121(11):1101–1104.
- Grindle MS. Good enough governance: poverty reduction and reform in developing countries. *Governance*. 2004;17(4):525–48.
- Grindle MS. Good enough governance revisited. *Dev Policy Rev*. 2007;25(5):533–74.
- Grindle MS. Jobs for the boys: patronage and the state in comparative perspective. Cambridge, MA: Harvard University Press; 2012.
- Jarman H. Trade Policy Governance: What Health Policymakers and Advocates Need to Know. *Health Policy*. 2017;121(11):1105–1112.
- Kjaer AM. Governance. New York; 2004.
- Marks G, Hooghe L, Blank K. European integration from the 1980s: state-centric v. multi-level governance. *J Common Mark Stud*. 1996;34(3):341–78.
- Marmor T. Fads in medical care policy and politics: the rhetoric and reality of managerialism. London: The Nuffield Trust; 2001. [Rock Carling Fellowship Lecture 2001].
- Nunnenkamp P. What donors mean by good governance: heroic ends, limited means, and traditional dilemmas of development cooperation. *IDS Bull*. 1995;26(2):9–16.
- Page EC, Jenkins B. Policy bureaucracy: government with a cast of thousands. Oxford: Oxford University Press; 2005.
- Rhodes RAW. Understanding governance: policy networks, governance, reflexivity and accountability. Philadelphia: Open University Press; 1997.
- Rosenau JN, Czempel E-O. Governance without government: order and change in world politics. Cambridge: Cambridge University Press; 1992.
- Rothstein B. The quality of government: corruption, social trust and inequality in international perspective. Chicago: University of Chicago Press; 2011.
- Sabel C. A quiet revolution of democratic governance: towards democratic experimentalism. In: OECD, editor. Governance in the 21st century. Paris: OECD; 2001.
- Sabet DM. Police reform in Mexico: informal politics and the challenge of institutional change. Stanford: Stanford University Press; 2012.
- Seekings J. Is the south Brazilian? The public realm in urban Brazil through a comparative lens. *Policy Polit*. 2013;41(3):351–70.
- Smith PC, Stepan A, Valdmanis V, Verheyen P. Principal-agent problems in health care systems: an international perspective. *Health Policy*. 1997;41(1):37–60.
- Smith PC, Mossialos E, Papanicolas I. Performance measurement for health systems improvement: experiences, challenges and prospects. Copenhagen: WHO Regional Office for Europe; 2008.
- Stewart E. What is the point of citizen participation in health care? *J Health Serv Res Policy*. 2013;18(2): 124–6.
- Thomas MA. Govern Like Us: U.S. Expectations of Poor Countries. Columbia University Press; 2015.
- Travis P, Egger D, Davies P, Mechbal A. Towards better stewardship: concepts and critical issues. In: Global programme on evidence for health policy discussion papers. 2002. [www.who.int/healthinfo/paper48.pdf](http://www.who.int/healthinfo/paper48.pdf)
- Travis P, Egger D, Davies P, Mechbal A. Towards better stewardship: concepts and critical issues. In: Murray CJ, Evans DB, editors. Health systems performance assessment: methods, debate and empiricism. Geneva: World Health Organization; 2003.
- Trump BD. Synthetic biology regulation and governance: Lessons from TAPIC for the United States, European Union, and Singapore. *Health Policy*. 2017;121(11):1139–1146.
- Uslaner EM. Corruption, inequality, and the rule of law. Cambridge: Cambridge University Press; 2008.
- Vasev N. Governing energy while neglecting health - The case of Poland. *Health Policy*. 2017;121(11):1147–1153.
- Walshe K, Smith J. Leadership and governance. In: Healthcare management. 2nd ed. Maidenhead: Open University Press; 2011.
- Williamson OE. The mechanisms of governance. New York: Oxford University Press; 1996.
- Willison C. Shelter from the Storm: Roles, responsibilities, and challenges in United States housing policy governance. *Health Policy*. 2017;121(11):1113–1123.
- Wolfe I, Mandeville K, Harrison K, Lingam R. Child survival in England: strengthening governance for health. *Health Policy*. 2017;121(11):1131–1138.
- Woods N. Good governance in international organizations. *Glob Gov*. 1999;5:39–61.
- Woods N. The globalizers: the IMF, the World Bank, and their borrowers. Ithaca: Cornell University Press; 2006.
- World Bank. Governance and development. Washington, DC: World Bank; 1992.
- World Bank. Governance: the World Bank's experience. Washington, DC: World Bank; 1994.
- World Health Organization. The world health report 2000: health systems: improving performance. Geneva: WHO; 2000.



# Provision of Health Services: Long-Term Care

# 43

Vincent Mor and Anna Maresso

## Contents

<b>Introduction</b> .....	950
Who Uses Long-Term Care Services and Supports? .....	951
Background to Long-Term Service and Support “Systems” .....	951
Structure of Chapter .....	953
<b>Financing of Long-Term Care</b> .....	953
Expenditure on Long-Term Care .....	953
Coverage .....	954
Paying for Long-Term Care Services .....	959
<b>Structure of the Delivery System</b> .....	961
The Long-Term Care Services and Supports Continuum .....	962
Long-Term Care Bed Capacity .....	963
Community-Based Service Capacity .....	965
Informal Care Provision and Cash Payments for Dependent Care Allowances .....	966
<b>Regulating Quality</b> .....	968
Different Regulatory Approaches to Quality Assurance .....	968
The Regulatory Reach of Quality Monitoring .....	969
Challenges Facing Quality Monitoring .....	974
<b>Summary and Conclusions</b> .....	974
<b>References</b> .....	976

V. Mor (✉)

Department of Health Services, Policy and Practice, Brown University School of Public Health, Providence, RI, USA

Providence Veterans Administration Medical Center, Center on Innovation, Providence, RI, USA  
e-mail: [vincent\\_mor@brown.edu](mailto:vincent_mor@brown.edu)

A. Maresso

European Observatory on Health Systems and Policies, London School of Economics and Political Science, London, UK  
e-mail: [a.maresso@lse.ac.uk](mailto:a.maresso@lse.ac.uk)

## Abstract

This chapter examines the financing, organization and regulation of long-term care in OECD countries. Historically, long-term care services and supports constitute a blending of social welfare benefits and health care provision. Depending on the complexity and severity of care recipients’ needs, delivery is characterized by both specialized nursing and medical care and personal and home-help services such as assistance with meals, grooming and

household chores. The delivery of long-term care is accomplished via institutional (residential) care, formal home care services, as well as through informal care provided by family members or hired care givers. In line with the preferences of older people to remain in their own homes, the past decade has seen a substantial shift in most OECD countries towards more home and community based care. This trend has regulatory and cost implications for monitoring the quality of care, which in the past has focused predominantly on institutions. Moreover, increased demand for formal services, in both residential and home care settings, due to ageing population pressures, also has implications for the long-term care workforce, with shortages anticipated over the next 20–40 years.

While funding of long-term care services comes mainly from public sources, there are very large variations between OECD countries in the resources dedicated to this sector. Eligibility for coverage also varies between countries, ranging from universal systems - based solely on need and not on income - to long-term care systems that apply means testing and safety-net principles to determine who qualifies for publicly-provided long-term care services and benefits. However, irrespective of financing model, all countries use some form of needs assessment to judge an applicant's level of functional impairment and care needs. Financial support is provided via in-kind services or through cash benefits to recipients to purchase the services they need (with varying degrees of restrictions). Cost-sharing, in the form of user charges, play a role in all countries, to different degrees, with service users, unless they are destitute, having to meet a proportion of the cost of their care from their own private resources.

The chapter also looks at the regulatory mechanisms used across a selection of countries to monitor the quality of long-term care, particularly in residential facilities, identifying three broad quality assurance approaches. The chapter ends with a

discussion of key challenges in quality monitoring and its role in enhancing user choice and stimulating improvements in providers' performance.

---

## Introduction

The network of long-term care services and supports that provide assistance, both financial and personal, to frail and disabled individuals in society is not really a system. While many who study long-term care document and compare the policies and practices that characterize country's service structure talk about the "long-term care system," in most countries it is best to view long-term services and supports as an amalgam of laws, policies, rules, practices, and service providers that emerged over the decades as a response to social and demographic changes in developed and developing societies. Unlike medical care or even public health structures, historically long-term care services and supports constitute a blending of social welfare supports and health care provision. The supports required by frail older, or seriously disabled, individuals include enhanced finances made necessary to buy the help needed to sustain daily life or the services from appropriate agencies to provide that help. Since the principal cause of frailty and/or disability is compromised health, it is almost always the case that more and more complex and comprehensive medical care is needed in conjunction with support services.

Another factor that differentiates long-term care services and supports from the provision of health care services is that most often long-term care services and supports is a family affair made possible by a person's spouse and children and less often extended family. Indeed, evidence suggests that, depending upon the country, between 10 and 40 times more care is provided by informal carers than by formal agency staff, whether institutionally based or providing home-based services (Columbo et al. 2011). Unlike demands for primary care medicine, as demand for long-term care services and supports increases it can be met

by policies favoring provision of care by families or by formal sources, the former more reminiscent of how long-term care has been historically provided.

The history of formal long-term care in western societies is closely bound up with the emergence of state sponsored social welfare efforts ranging from “almshouses” to outdoor relief efforts designed to support paupers and others unable to care for themselves (Kellog 1883; Katz 1996). Almshouses in Britain, the Netherlands, France, and Belgium housed indigent elderly and disabled persons unable to care for themselves without family members to whom they or local authorities or charities could appeal for support. In European and Anglo-Saxon countries these facilities emerged from a tradition of sectarian or local charitable organizations but were not infrequently conflated with support for the poor, the destitute, and the alcoholic. Local authorities, not just in England where the poor laws prevailed, established almshouses or hospices to care for the unfortunate and dying as a civic responsibility.

There are two overlapping dualities that characterize the scope and delivery of long-term services and supports. First, services and supports represent both financial support for basic food and shelter and the provision of physical support and care for those unable to do even the simplest daily tasks without help. Second, because care recipients are in need largely due to the complexity and severity of their medical conditions, services generally involve both unskilled homemaker services as well as specialized nursing and medical care. Third, different countries have adopted varying mechanisms to meet the long term care needs of population ranging from cash payments to eligibility determination processes which fundamentally define each country’s long term services and supports structure. As will be observed in the paragraphs below, many countries make all these different dimensions of services and supports available under public funding with or without means testing the client and/or her family.

## **Who Uses Long-Term Care Services and Supports?**

Most users of formal long-term care services (institutional or community-based) are women aged 80 and over. However, according to OECD health data, there is considerable country to country variation in the proportion of women aged 80 and over who use long-term care services from a low of 2% in Poland to a high of over 45% in Norway (Columbo et al. 2011). Interestingly, depending upon the country, a sizeable minority of long-term care service recipients are under age 65, with Poland leading the way with almost half (48%) of formal care recipients being under 65. However, the substantial variation in the availability of home versus institutionally based long-term care services and how cash allowances provided to frail elderly persons and their families are counted in long-term care user statistics makes it difficult to be too precise in comparing rates of use across OECD countries. This is a theme which will be revisited throughout this chapter since without reliable data to characterize the nature of the services provided and the characteristics of the recipient population, it is difficult to have a great deal of confidence in many of the statistics used to compare the long-term care systems of one country with another.

## **Background to Long-Term Service and Support “Systems”**

As noted, long-term care “systems” are in almost all cases a misnomer because it is the exceptional country that actually has an integrated system. Regardless of its “system-ness,” in general, long-term care can be conceptualized as three interlocking sets of policies and forces which apply, regardless of the country. These three features include: (1) financing and reimbursement, that is, who pays and how the services rendered are reimbursed; (2) the organization of the delivery system, that is, how the providers of long-term care services and supports are organized and coordinated, since clients often receive a multiplicity of services from different providers and sources;

and (3) the regulatory or quality assurance system, that is, the regulations, rules, and procedures governing licensure and quality standards for agencies serving the long-term care population.

These three components are interdependent; changes made in each affect the implementation and impact the others have, either directly or indirectly. For example, over the past decade most OECD countries have increased their emphasis on home and community-based services to meet the preferences of a new generation of seniors who are less willing to be relegated to an institutional setting (Grabowski et al. 2010; Damiani et al. 2011). Indeed, the movement toward “consumer directed care” represents the epitome of the shift toward home and community-based care since the underlying assumption is that older persons will use the new discretion to remain in the community and outside of institutions (Alakeson 2010).

Shifting payments to home care providers, away from the past dominance of institutions, has immediate implications for the structure of the delivery system as well as how it is regulated. To implement policies stimulating the development of home care services, regulatory structures that have historically been oriented toward monitoring quality in institutions must be realigned to manage a much more diverse and complex oversight process. To assure that agencies charged with meeting the needs of the elderly in their homes actually are providing the care for which they are paid requires visiting clients and their families in their homes and/or demanding extensive care management auditable documentation. This means that the costs of realigning long-term care services from the institution to the community will require a very different, and costly regulatory and oversight structure. Furthermore, financing home and community services represents a substantial departure from the institutional approach, where purchasing a day of care in a nursing home is well understood. In the case of home care services, whether to pay by the hour, the skill level of the staff person, or even to bundle payment with other post-acute care services or via capitation are all decisions that have different implications for how payments are made. To the

extent that financing changes are also designed to give the eligible service recipient some choice as to how their needs are to be met in the form of “consumer direction” (Doty et al. 2010), additional operational complications arise related to personal care workers’ compensation, indemnification, and even whether family members can be paid to provide the care. As is obvious, these wrinkles in the financing rules and allowances introduce a further complication in the regulatory control structures since it is difficult for government to regulate the quality of familial relationships.

Changes in financing also have implications for the organization of the long-term care delivery system. For example, countries that instituted universal long-term care insurance policies that include cash transfers must determine whether those funds can be used to purchase home care services, regardless of the licensure status of the agency or worker employed to provide the service. However, policies which only reimburse recipients and their families for services rendered by licensed or professionally supervised staff necessarily means costs will be higher. Without such requirements, institutional care providers, who are required to adhere to professional licensure requirements and labor laws including tax withholdings, would have a legitimate complaint about there not being a “level playing field,” since cash transfer payments that result in families hiring illegal immigrants can be seen as undermining the formal health and social care services labor market.

Understanding how changes in one component of the system affect the others is further complicated by the fact that it spans the health care sector as well as the formal and informal labor market. The emphasis on home care places increased pressure on family caregivers who, in rich countries, often supplement direct family care time with undocumented workers’ time, thereby violating labor laws and possibly endangering the frail older person. Recent efforts within the OECD to better characterize the variation in long-term care

systems across member countries has identified clusters of countries based upon variation along two dimensions pertinent to long-term care. The two dimensions first include, the “generosity” of the formal entitlements to the long-term care services and supports that recipients require and second, the ease with which individuals can access needed community services and the organizational complexity associated with monitoring or regulating the array of available services. A recent OECD report which characterized most EU countries on these dimensions found that countries with more generous long-term care financing systems also offered a broader array of services and more choice for service users (Mot and Willemé 2012)

### Structure of Chapter

In this chapter we characterize the major issues facing industrialized countries with respect to the financing, organization, and regulation of long-term care services and supports. Each of these aspects of countries’ long-term care “systems” is discussed separately, although, as noted, these are integrally intertwined. To the extent that the available literature allows, we compare selected countries with respect to their approach to financing long-term services and supports and their organization (private vs. public, institutional vs. home based) with special reference to the level of informal care support provided by family and friends. Finally, the regulatory structure for licensing and certifying institutional and home care agencies meeting the needs of frail elders are described across selected countries. Since we rely heavily upon research studies funded by the EU and/or OECD for our comparative data, not all countries are consistently represented. We close the chapter by identifying several salient issues that could benefit from additional rigorous cross-national review along with the challenges facing industrialized nations as they strive to meet the long-term care needs of their growing population of frail elderly persons.

## Financing of Long-Term Care

### Expenditure on Long-Term Care

Expenditures on long-term care vary significantly among countries, and spending reflects differences in care needs, utilization rates, the comprehensiveness of formal long-term care services, and the role of families in providing informal care. Another factor that affects expenditures is whether services are defined as health or social services. With this proviso in mind, a recent report (Columbo et al. 2011) calculates that OECD countries spent an average of 1.5% of GDP on long-term care in 2008, (Long-term care spending is calculated on the basis of health-related long-term care services (including palliative care, nursing care, personal care services, and health services supporting family care) and social services related to long-term care (including home help and care assistance and residential care services).) but with levels for individual countries ranging from less than 0.5% (e.g., Hungary, Slovenia, South Korea, and Poland) to more than 3% (The Netherlands and Sweden). These differences are put into even starker relief when per capita expenditure on long-term care is considered. On this metric, the highest spenders devote almost three times as many resources as mid-range countries (such as Australia, Germany, Japan, and the United States) and between 20 and 30 times more than the lowest spending countries (see Table 1). Table 1 also indicates the source of LTCF funding in each country, revealing that long-term care is mainly funded through public sources. It should be noted, however, that data systems in most countries are not sufficiently robust to capture all aspects of private spending on long-term care, and that there is a great deal of underreporting of direct out-of-pocket spending. Nevertheless, with the exception of Switzerland, government sources account for the lion’s share of public financing for long-term care. Interestingly, the data also indicate that in Portugal, Germany, and Spain private expenditures on long-term care are considerable.

**Table 1** Funding for publicly provided long-term care, selected OECD countries

Country <sup>c</sup>	Per capita spending on LTC (US\$ PPP) <sup>a</sup>	LTC funding as % of GDP <sup>a</sup>	Total government /state component (%) of public LTC expenditure (incl. taxes and social insurance) <sup>b, d</sup>	Private share/out-of-pocket component (%) of public LTC expenditure <sup>b, d</sup>	Private insurance component (%) of public LTC expenditure <sup>b, d, e</sup>
Slovak Republic	42	0.2	–	–	–
Czech Republic	59	1.4	100	0	0
Poland	68	0.4	92.3	0.3	0
Korea	73	0.3	76.9	17.8	0
Hungary	108	0.3	90.3	2.4	0.9
Spain	271	0.6	71.9	28.1	0
Slovenia	302	0.8	75.4	24	0.5
Australia	367	0.8	88.9	8.5	0.3
New Zealand	383	1.3	92	4.4	1.3
United States	455	0.6	–	–	–
Germany	470	0.9	67.2	30.4	1.7
Austria	497	1.1	81.8	17.1	0
Japan	527	1.4	88.9	7.1	4
France	564	1.7	99.2	0.4	0.4
Canada	574	1.2	82	16.8	0.4
Iceland	638	1.7	100	0	0
Belgium	707	1.7	90	0.2	9.8
Denmark	724	1.8	89.6	10.4	0
Finland	790	1.8	84.4	14.2	0
Luxembourg	822	1.4	–	–	–
Norway	1276	2	89.3	10.7	–
Sweden	1332	3.6	99.2	0.8	0
The Netherlands	1431	3.5	99.9	0	0
Portugal	–	0.1	53.4	45.4	1.1
Switzerland	–	0.8	38.8	58.4	0.4

Source: Adapted from Columbo et al. (2011)

Notes

<sup>a</sup>Data from 2008.

<sup>b</sup>Data from 2007.

<sup>c</sup>Countries are listed from lowest to highest per capita expenditure on long-term care

<sup>d</sup>Funding from government sources, private out-of-pocket expenditures, and private insurance do not always add up to 100% as the following other minor funding sources are excluded from this table: nonprofit institutions serving households, corporations (other than health insurance), and “other”.

<sup>e</sup>Data on out-of-pocket spending for some of the countries are underestimated. For example, in the Netherlands, cost-sharing on long-term care services is estimated to account for 8% of the total long-term care expenditure. The share of out-of-pocket spending for Switzerland is overestimated as cash benefits granted for care in care facilities are not considered.

## Coverage

One way of looking at the financing mechanisms behind public long-term care is to consider three aspects of coverage: the scope of entitlement (i.e., on what basis are citizens entitled

to publicly provided services and benefits?), the range of benefits covered, and the proportion of the benefit cost that is covered, including those services that are excluded from public funding (cost-sharing and user charges).



### Types of Public Long-Term Care Systems

The scope of entitlement provides a useful way to classify countries' public long-term care systems since this approach captures whether entitlement is universal or whether access to services is means-tested and thus reserved for the poorest individuals who are protected through a public safety-net. In addition, such coverage may be financed through a single program (such as general taxation or a mandatory long-term care insurance scheme) or through multiple programs and benefits. Using these criteria, Columbo et al. (2011) identify three long-term care models: (1) universal coverage systems with a single program, (2) mixed systems, and (3) means-tested safety-net systems. Table 2 classifies a number of OECD countries according to this typology. The main feature of single-program universal systems, as found in Germany, Japan, Luxembourg, the Netherlands, and South Korea, which have mandatory long-term care insurance schemes, and the Nordic countries, which have tax-based financing programs, is that they provide public long-term care services to everyone who is assessed as needing care, based on their dependency level and regardless of income. That is, access to services is not dependent on the income level or assets of beneficiaries. Mixed systems, as seen in Australia, Austria, France, and Spain, typically have a number of different programs and benefit schemes operating side by side, which can be either universal or means-tested, with the amount of the benefit adjusted downwards as the recipient's income level increases. The countries in this group may also have medical-related or nursing benefits covered universally (free) through the health system. Finally, means-tested safety-net systems (found in the United Kingdom and USA) use income or asset tests to set a threshold for entitlement to publicly provided long-term care services and benefits. Income and asset-testing is used to target those with the highest care needs and to protect those who otherwise would not have the means to purchase care privately. However, if means-testing thresholds are set quite low, a large proportion of elderly people in need of long-term care may be excluded from receiving publically provided

services, until they become impoverished paying for such services privately. The only other way that these individuals' long-term care needs can be met is if services are provided as part of the health care system (as in the case of nursing care in the United Kingdom) which sets up the dynamic of cross-subsidy between the health and social services sector, the latter generally being a more costly means of meeting the same need.

It should be noted, however, that whatever long-term care coverage model a country has, needs assessments to judge an applicant's level of functional impairment and care needs are a central component of determining eligibility. Nor is it the case that countries' approach to coverage necessarily corresponds with their level of spending. For example, while the Netherlands, Sweden, Norway, and Luxembourg dedicate the highest per capita spending to long-term care services, other universal-system countries such as Germany, Japan, South Korea, as well as Denmark and Finland fall within the mid to upper-mid expenditure range. Similarly, all mixed system countries have mid-range long-term care spending, as does the USA, which belongs to the means-tested safety-net group.

### What Long-Term Care Services Are Covered?

Most public long-term care systems cover both institutional and home-based services, although the range of services covered varies, as does the proportion of the cost (see Table 2, as well as the subsection on cost-sharing below). Universal long-term care systems tend to provide comprehensive long-term care packages encompassing institutional/residential services, home care nursing, domestic assistance as well as sheltered housing schemes, assistive devices, home modification, and transport to community services. However, in some universal systems, such as those in Germany and South Korea, a notable omission from the long-term care package is accommodation (all rooms in Germany and private rooms in South Korea) and meal costs in nursing homes; these must be paid for out-of-pocket. In Japan, lodging and meals in nursing homes are only partially covered. In mixed

**Table 2** Public coverage of long-term care, selected OECD countries, 2010

Country	Type of system based on eligibility (universal; mixed; means-tested/low income)	Financing source: tax, social security contribution	Government levels contributing to financing	Program characteristics	Who is covered	Income/means-testing to determine eligibility?	Needs assessment	Types of benefits provided
Australia	Mixed	Tax-based	Federal, state, and local	Multiple programs	Older people	Yes	Yes	In-kind only: home and institutional care
Austria	Mixed	Tax-based	Federal and regional (lander)	Cash benefit programs: universal cash benefit (Pflegegeld) and 24-hour care benefit	All disabled people	Universal cash benefit: no 24-hour care benefit: yes	Yes	Cash: home and institutional care Some in-kind benefits provided by regional governments
Canada	Mixed	Tax-based	Federal and provinces	Various programs by province	All people in need	Universal (home care) and means-tested (institutional care)	Yes	In-kind: home and institutional care
Denmark	Universal	Tax-based	National and local	Single program of assistance	All people in need	No	Yes	Cash and in-kind home and institutional care
Finland	Universal	Tax-based	National and municipalities	Single program of assistance	All people in need	No	Yes	Cash and in-kind home and institutional care
France	Mixed	Taxes, social contributions	Central and local	Various income-related benefits	All people in need; Handicap allowance is for those aged 60+	Income	Yes	Cash and in-kind home and institutional care
Germany	Universal	Social insurance	Payroll contributions	LTC insurance system with multiple insurers	All people in need	No	Yes	Cash and in-kind home and institutional care

Italy	Mixed	Tax	National and regional	Institutional care benefits part of the health system; cash care allowance covers home care	All people in need	No	Yes	Cash and in-kind home and institutional care
Japan	Universal	Social insurance, plus personal contributions	National	LTC insurance system Insured individuals aged 40–65 pay 30% of total LTC costs	Over 65, or 40–65 with age-related disease	No	Yes	In kind only: home and institutional care
Korea	Universal	Social insurance and taxes	National	LTC insurance system	Over 65 s, or under 65 suffering from geriatric diseases	No	Yes	Cash and in-kind home and institutional care
Luxembourg	Universal	Social insurance, tax and a special tax	National	Single LTC insurance system part of health insurance system	All people in need	No	Yes	Cash and in-kind home and institutional care
The Netherlands	Universal	Social insurance	National	LTC insurance system with multiple insurers	All people in need	No	Yes	Cash and in-kind home and institutional care
New Zealand	Mixed	Tax-based	National	Health funding authority responsible for LTC provision; Residential Care Subsidy	All people in need	Yes	Yes	In kind only: home and institutional care
Norway	Universal	Tax-based	National and local	Single program	All people in need	No	Yes	Cash and in-kind home and institutional care
Spain	Mixed	Tax-based	Central and regional	National long-term care system administered by regions	All people in need	Yes	Yes	Cash and in-kind home and institutional care
Sweden	Universal	Tax-based	Local and national (11–12%)	Single program	All people in need	No	Yes	Cash, in-kind and vouchers: home and in-kind care varies across municipalities

(continued)

**Table 2** (continued)

Country	Type of system based on eligibility (universal; mixed; means-tested/low income)	Financing source: tax, social security contribution	Government levels contributing to financing	Program characteristics	Who is covered	Income/means-testing to determine eligibility?	Needs assessment	Types of benefits provided
Switzerland	Mixed	Social (health insurance), state budget	National and cantons	Mandatory health insurance program plus complementary cash benefits under Disability Insurance	All people in need	Asset tested (for some benefits)	Yes	Cash and in-kind institutional care; home care mainly provided by private organizations
United Kingdom	Means-tested safety-net	Tax-based	National and local	Various programs and allowances	Social care benefits to all adults in need; specific allowances for the disabled and elderly disabled	Asset tested (for some benefits)	Yes	Cash and in-kind home and institutional care
United States	Means-tested safety-net	Tax-based	National and state	Medicaid and Medicare programs	People of low income (Medicaid) Seniors (Medicare)	Medicaid is means-tested; Medicare is universal for seniors	Yes	Mainly in-kind: institutional benefits. Optional state home care benefits

Sources: Adapted from Fernandez et al. (2009); Columbo et al. (2011); Swartz (2013)

Note: LTC – long-term care.

systems, typically nursing care, either in home or institutional settings, is financed on a universal basis by the parallel health system while personal (social care) is covered under separate benefit schemes. For example, in Italy, special nursing homes for elderly people are covered via the health system budget while home care services are mainly financed by a non-means-tested cash care allowance whose modest level means it is most often used to pay for informal care. In Canada, most provinces cover nursing and personal care (such as help with bathing and grooming) in home settings, but other assistance such as domestic help and meal preparation may require the user to pay a fee. In the group of means-tested safety-net countries, the United States sets a basic mandatory basket of long-term care services (such as nursing facility services and home health-related services) through its Medicaid program for people on low incomes, but individual states determine what other services may be covered. In most states while benefit structures cover support for daily living activities in home-care settings as well as accommodation and meals in nursing homes, the latter services are only available to those who meet strict means-testing and who have exhausted their own resources before becoming eligible for public support (Columbo et al. 2011).

### **Paying for Long-Term Care Services**

Comparing countries' approach to paying for services is complicated by the lack of comparable international data on the different reimbursement mechanisms used to pay providers for different types of care, whether it be fee-for-service payments, capitation, or day-rates for nursing costs. The importance of having data on the impact of different reimbursement vehicles may be illustrated by the case of how institutional services (i.e., in nursing homes) are paid for. If a country's reimbursement mechanism does not recognize, and adjust payment levels proportionally for clients/patients who have more complex needs and require more care, there will be a disincentive for providers to admit such individuals as their greater

impairment will cost the facility more in terms of the time, labor, and skills required to care for them. In this example, a form of case-mix reimbursement (such as the Resource Utilization Groups case-mix system used in many US states and in Ontario, Canada) that provides an incentive to care for sicker patients would be more appropriate than a flat-rate reimbursement model that pays the same amount per nursing home resident, regardless of the intensity of their care needs. While most often applied to the institutional setting, it is possible to devise case-mix reimbursement models for home care services. Such considerations impact on both the efficiency of the long-term care system as well as on its capacity to meet the growing care needs of the population requiring long-term care services.

### **Cash Benefit Schemes**

There is some cross-national information on cash-benefit schemes, which offer recipients the choice to purchase care services that they feel best meet their needs from the provider they prefer. While most countries offer a combination of in-kind services and cash benefits, a few, like Austria, France, and the Czech Republic, use cash benefits as the main type of long-term care purchasing mechanism. These schemes differ among countries as to whether they are available alongside in-kind benefits or whether recipients must choose either one or the other, whether the level of the cash benefit is determined through means/income testing, and whether any restrictions are placed on how the benefit may be used. For example, some countries require that only accredited formal services be hired while others have very few restrictions and allow the benefit to be used to pay family members or other informal carers for services rendered in the home. There is some evidence that the use of unregulated cash payments seems to incentivize the hiring of migrant care workers in countries such as Austria and Italy, who either substitute or compliment personal care and domestic assistance traditionally provided by the family (van Hooren 2008; Columbo et al. 2011; Phillips and Schneider 2007; see also section "[Structure of The Delivery System](#)" on Provision). Table 3 provides an overview of the cash

**Table 3** Cash for care schemes for long-term care services, selected OECD countries

Country	Benefits available	Cash benefit programs	Income/ asset tested	Use restrictions
Austria	Both in-kind and cash	1) Cash Allowance for Care ( <i>Pflegegeld</i> ) 2) 24-hour care benefit 3) Dementia care benefit	1) No 2) Income 3) No	No. Can be used to pay for care by relatives or other carer
Czech Republic	Only cash benefits	Care allowance	No	No. For services or care by relatives
Denmark	In-kind, cash, and vouchers	BPA (Citizen Controlled Personal Assistance)	No	Yes. Not for nursing care
France	In-kind and cash benefits are separate	<i>Allocation personnalisée d'autonomie</i> (APA)	Income	Yes. Use of APAs is strictly controlled. Can be used to pay for care by relatives but not a spouse
Germany	Users must choose between either in-kind or (lower value) cash benefits	Cash benefits part of LTC insurance scheme: 52% of users opt for cash benefits	No	Yes. Cannot be used to pay for care by relatives or for some services (such as GP services)
Italy	In-kind and cash benefits are separate	<i>Indennità di accompagnamento</i> (Carer/ Companion allowance)	No	No. Can be used to pay relative or other carer
Korea	Users must choose between either in-kind or (lower value) cash benefits	Cash benefits part of LTC insurance scheme	No	Only available to users who live in remote areas with few facilities, are unable to use LTC facilities due to national disasters, or are unsuitable for institutional LTC due to physical or mental condition. Cannot be used to pay for care by relatives
Luxembourg	Users must choose between either in-kind or (lower value) cash benefits	Cash benefits part of LTC insurance scheme: Cash Allowance for Care	No	Cash for the first 10.5 hours of care per week
The Netherlands	Users must choose between either in-kind or (lower value) cash benefits	Cash benefits (Personal Care Budgets) are part of LTC insurance scheme: 12% of users opt for Personal Care Budgets	No	98.5% of expenses must be justified and unspent funds returned. Personal Care Budgets can be used to pay for care by relatives but they must have a contract
Spain	Users must choose between either in-kind or cash benefits (the latter vary according to program)	1) Allowance for user to hire services 2) Allowance for user receiving informal care 3) Allowance for Personal Assistance	1) Income 2) Income 3) Income	1) Hire through accredited centers 2) To compensate informal carers who must be a relative or in rural areas; a neighbor can qualify 3) Expenses must be justified; carer must have professional qualifications
Sweden	In-kind and cash benefits are complementary; also vouchers	1) Attendance Allowance 2) Assistance Allowance	1) No 2) No	Yes. Cannot be used to cover medical expenses or to pay for care by relatives

(continued)

**Table 3** (continued)

Country	Benefits available	Cash benefit programs	Income/ asset tested	Use restrictions
United Kingdom	In-kind and cash benefits are complementary	1) Attendance Allowance 2) Direct Payments 3) Individual (social care) Budgets	1) Income and asset tested 2) Income and asset tested 3) Income and asset tested	1) No 2) Yes. Spending record required 3) Yes. Cannot be used to pay for care by relatives

Sources: Adapted from van Hooren (2008); Columbo et al. (2011); Swartz (2013); Wirmann Gadsby (2013)

benefit schemes available in a selection of countries, highlighting these major differences.

### Cost Sharing

Cost-sharing, in the form of copayments, deductibles, or user-charges apply to all long-term care systems, whether they are universal, mixed, or means-tested, safety-net systems. Commentators (Swartz 2013) have noted that rising long-term care costs, aging populations, and pressure on public sector spending due to structural deficits and the recent financial crisis in Europe since 2008 have seen a shift to greater cost-sharing among users of long-care services or their relatives. In most cases, cost-sharing is subject to income thresholds, with exemptions available for those meeting set criteria, such as low-income status (Columbo et al. 2011; Swartz 2013). For example, in the Nordic countries with universal systems, cost-sharing mechanisms account for relatively low shares of publically financed formal long-term care services and in Sweden and Norway, such contributions are capped. In contrast, beneficiaries in South Korea are required to pay a coinsurance rate of 20% for residential care and 15% for home care (Jung et al. 2014). Similarly, in Australia, those eligible for public long-term care services still need to contribute to the cost of their personal care in both residential and home settings, with the amount determined through means-testing (Columbo et al. 2011). Table 4 summarizes a number of cost-sharing approaches to long-term care and provides some country examples. Like user charges and cost-sharing approaches in health care, it is clear that the different approaches found in long-term care systems reflect not only the incentive structures that

modulate access to care but also countries' different emphasis on social protection for vulnerable or low-income groups.

### Structure of the Delivery System

As can be seen in Fig. 1, based upon OECD data on long-term care, there is substantial variation in the percentage of the population over 65 using long-term care services (OECD 2013b). Consistent with its origins in the medieval alms house and hospice, traditionally long-term care was synonymous with residential arrangements provided in an institution. Indeed, in spite of the concerted effort that most OECD governments have made in "rebalancing" long-term services and supports from institutions to community-based services, spending on long-term care in institutions was higher than spending at home in virtually all OECD countries in 2008. On the other hand, there are many more using long-term care services residing at home. While in the average OECD country 12.9% of those 65 and over receive formal long-term care services, less than half (under 5%) receive care in a residential or institutional setting. Indeed, most countries report that about twice as many in the population of long-term care users are receiving those services at home (Columbo 2011). Figure 2, reflecting OECD Health Statistics data reveals that in many countries with data on the distribution of home-based and institutional care over time, it is evident that the share of long-term care users receiving home care has increased in most countries and as an OECD average.

**Table 4** Cost-sharing approaches for long-term care services in selected countries

Cost-sharing approach	Country examples
Users have to first exhaust their own means (means-tested systems)	<i>United Kingdom</i> Eligibility for residential care is means-tested and individuals with savings over a threshold are not eligible for public support. Cost-sharing is applicable according to income/savings under the threshold but some support from local government is available. Individuals with less than GBP 14,250 in savings qualify to have their residential costs fully covered.
Residual cost-sharing – after defined public benefits	<i>France</i> The <i>Allocation personnalisée d'autonomie</i> (APA) cash benefit is subject to a national ceiling, and the level of benefit decreases as a proportion of income. <i>Germany</i> Cost-sharing applies when the costs of long-term care services go beyond the fixed public benefits. Families are required to help cover costs that exceed statutory benefits. For residential care, recipients must cover accommodation and meals; means-tested social assistance benefits may be available to those who cannot meet these costs.
Flat rate (as a percentage) cost-sharing	<i>Japan</i> The long-term care insurance scheme sets a user-charge rate of 10% on all public long-term care services (excluding preventive services). <i>South Korea</i> Under the national long-term care insurance scheme beneficiaries pay 20% of total institutional care costs and 15% of home care services costs, with reductions or exemptions for low-income individuals.
User charges are linked to income and/or assets-based benefits	<i>Finland</i> In home care, private contributions are set according to the amount of care needed and income of the recipient and other household members, covering about 15% of total costs. In institutional care, personal contributions are set at 85% of the recipient's net income. <i>Norway</i> Municipalities have the flexibility to set personal contributions within given frameworks. Personal contributions are typically income-related, except for short-term stays in nursing homes, where contributions are set independently from income. For long-term nursing home stays, personal contributions cannot exceed 80% of a resident's income in excess of a given amount. For home care, user charges are set so as to leave the recipient with a minimum income for extra expenses. <i>Spain</i> Private contributions are determined by each autonomous region and differ according to care setting and type of service. The level of cost-sharing depends on an assessment of financial capacity, typically based on available capital, the beneficiary's estate, and household income. Private out-of-pocket payments range from 70–90% for residential care and 10–65% for home care.

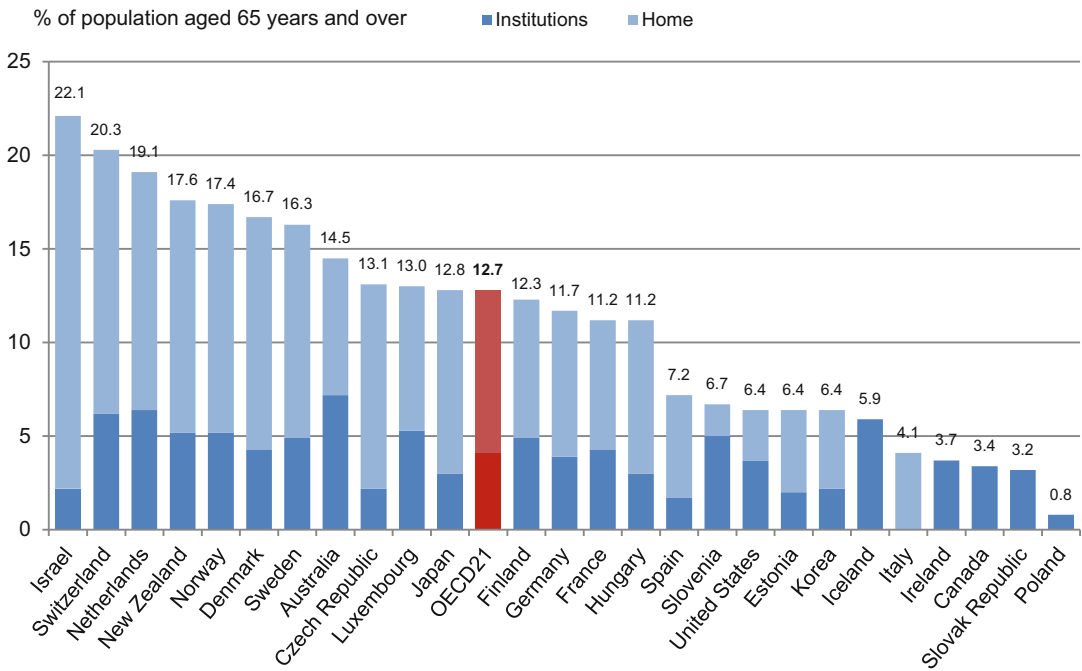
Source: Columbo et al. (2011)

## The Long-Term Care Services and Supports Continuum

As noted, presently most OECD countries have a higher percentage of long-term care users receiving care at home than institutionally

based care. While many countries have had a range of different types of long-term care residential arrangements for frail older persons for many decades, the full array of community-based services has been a relatively recent development. This required the development





**Fig. 1** Percentage of population aged 65 or over receiving long-term care services, by country, 2011 (Source: OECD 2013a)

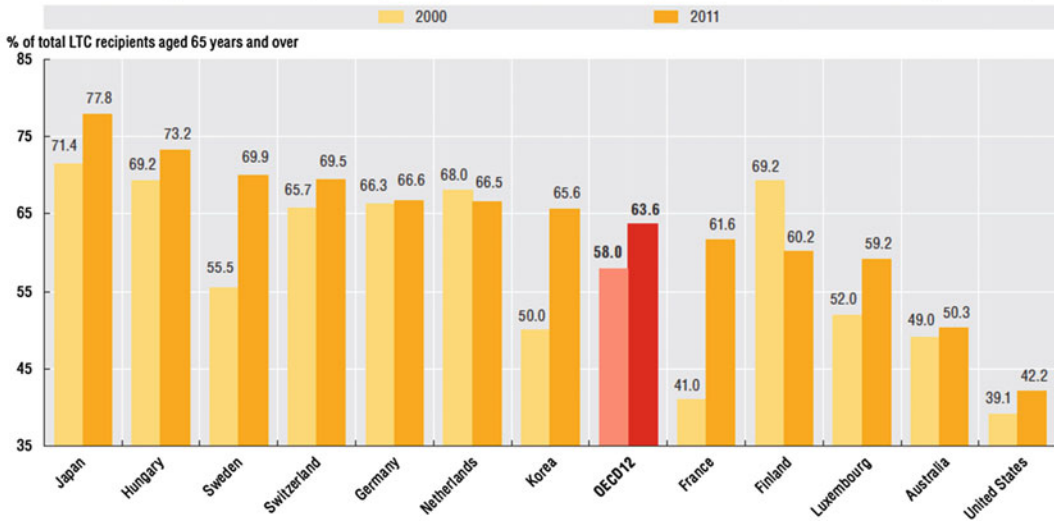
of a continuum of long-term services and supports, ranging from household chores to intensive, medically oriented nursing home care to serve individuals as their needs increase. The different levels of intensity of nursing home care offered range from facilities that manage chronically bed-bound patients requiring oxygen, artificial feeding, and intravenous care to facilities specializing in short term rehabilitation to independent small apartments offering congregate meals. Nonresidential long-term care services can range from intensive round the clock “respite” services to weekly chore and cleaning services, and all the range of nursing to meals services offered in the homes of dependent elders or in community settings. Day care programs, with and without medical and nursing support, increasingly serve frail older individuals who otherwise live with caregiver children. Finally, some have argued that even the differentiation between residential and home care services can be false since, regardless of where one lives, needed services can be provided to meet their needs (Kane et al. 1998).

Indeed, some have argued that financing rules and contradictory regulatory controls are the major drawbacks to having more comprehensive and responsive long-term care delivery systems.

### Long-Term Care Bed Capacity

Even though home care is more prevalent than institutional care, the best data regarding long-term care services across the OECD refers to the availability of residential long-term care beds per elderly person. Figure 3 reveals the substantial intercountry variation in the number of residential long-term care beds per 1000 elderly, ranging from under 20 in Italy, Poland, and Korea to over 60 in the Nordic and northern European countries. (In the OECD data, Japan has very few nursing home beds but many long stay hospital beds which are not counted as nursing homes although they serve a very similar population (Ikegami et al. 2014.) Not included in these figures are OECD countries that do not report data on long-term care use such

### 8.5.2. Share of long-term care recipients aged 65 years and over receiving care at home, 2000 and 2011 (or nearest year)

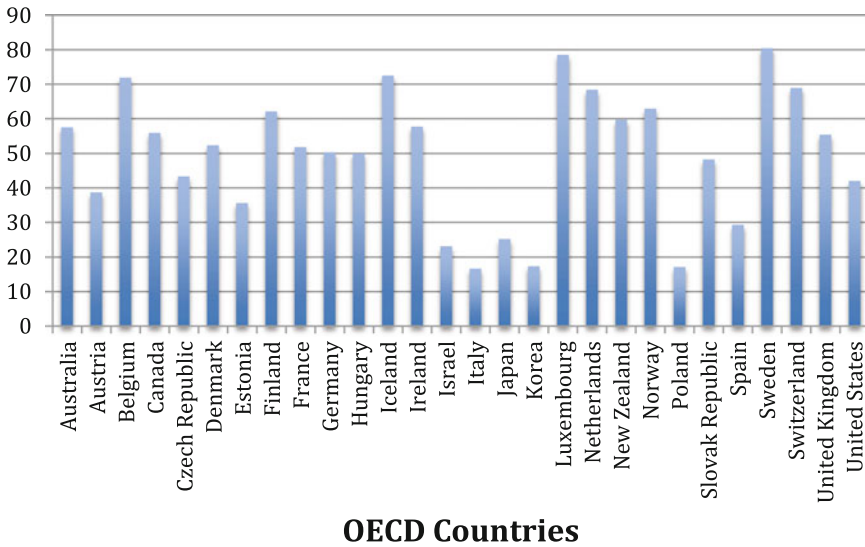


**Fig. 2** Share of long term care recipients aged 65 and over receiving care at home, 2000 and 2011 (Source: Columbo et al. 2011)

as Mexico and Chile which must be presumed to have even less well-developed resources than countries like Italy, Korea, and Poland. In spite of the higher prevalence of home care recipients, spending on long-term care in institutions is higher than spending on home care in all OECD countries reporting with the exception of Denmark (Columbo et al. 2011). This reflects two phenomena: first, institutional care is more costly than home care and second because a higher proportion of those receiving institutional care are very impaired, particularly in the absence of very involved family caregivers (Carpenter and Hirdes 2013).

In the absence of standardized definitions for what constitutes a “long-term care bed,” OECD data on the rates of such beds per 1000 elderly are necessarily vague. Many countries have different definitions for what constitutes a long-term care bed. For example, Japan has many small, long stay hospitals licensed quite differently from other Japanese acute hospitals but serving populations that are not all that different from licensed nursing facilities (Ikegami et al. 1994; Ikegami et al. 1997). On the other hand, over the last several decades a new class of residential long-term care home, Assisted Living

Facilities, in the USA, that are not nursing homes per se, has made it appear as if the number of nursing home beds per capita elderly has fallen dramatically, even though the average impairment level of Assisted Living residents now is as great as it used to be among nursing homes two decades ago (Sloane et al. 2005; Smith et al. 2007; Stevenson and Grabowski 2010). Thus, the recorded number of long-term care beds per 1000 elderly is very sensitive to the different definitions of what constitutes a long-term care bed, suggesting that in both Japan as well as in the USA, OECD data undercounts the number of long-term care beds, although for different reasons. Indeed, according to the first national survey of long-term care providers done by the US National Center for Health Statistics, in 2012 there were some 15,700 nursing homes but 22,000 Assisted Living Facilities with 39 nursing home beds per 1000 elderly and 20 Assisted Living beds per 1000 elderly. Adding these together places the USA at the same level of total long-term care beds per thousand elderly as exists in the Netherlands and Belgium, (but still below Sweden) and substantially higher than Germany or France. However, it is not known whether certain classes



### OECD Countries

**Fig. 3** Residential Care Home Beds per 1000 Elderly aged 65+, selected OECD Countries, 2009 (Source: OECD 2013a)

of retirement housing in those countries actually serve some of the same functions as do Assisted Living Facilities in the USA.

Notwithstanding the difficulty of measuring current bed capacity in a comparable way, OECD planners have recently projected the future need for residential long-term care in multiple countries. Between 2010 and 2060, a recent report estimates needing a 100% increase in the number of residential long-term care placements in Germany, whereas they project nearly twice that in the Netherlands. In Spain and Poland, which begin from a much lower base, the rate of increase estimated to be needed is around 150%. Such projections are notoriously unreliable, at least based upon the experience in the USA where estimates of future demand for nursing home care were anticipated to more than double by now. However, changes in personal preferences, expectations, and the poor public image nursing homes have in the mind of the public, all worked to undermine those projections (Miller et al. 2010, 2012). The continually dropping nursing home occupancy rates, in spite of declining bed supply, is a clear indication of reductions in demand for nursing home care in spite of the ongoing aging of the population (Larson Allen 2008; Feng et al. 2011).

### Community-Based Service Capacity

In response to a growing preference for care at home, over the past decade many OECD countries have implemented programs and benefits to support home-based care. Where trend data are available, the share of people aged 65 and over receiving long-term care at home, as a share of the total number of long-term care recipients, has increased substantially over the past 10 years (Columbo et al. 2011). Other than this general statement, however, there are few studies that have documented the changing supply or the detailed levels of use of the variety of different home care services offered by agencies in Europe. The OECD data on long-term care services provides specific information only on the number of nursing home beds in the country and not information regarding the number of adult day care centers or home care agencies, even though, the literature and OECD data itself clearly reveal that for the last decade or more the majority of recipients of formal long-term services and supports have them provided in their homes.

One obvious reason for this lack of systematic data on service supply relates to the difficulty of arriving at common definitions of services across the boundaries of different countries, languages,

and cultural histories. Additionally, in many OECD countries the organization and provision of long-term care services is a local matter for municipalities operating under national guidelines that still allow considerable local discretion in how long-term care policies are implemented (Tarricone and Touros 2008). This means that national governments may not have the kind of detailed data that would make it possible to characterize the supply of services across the whole country. For example, a 2008 report on home care included a section on the supply of home care without offering any data regarding the number of agencies, workers, or services available to the elderly population (Tarricone and Touros 2008).

A report compiled as part of a conference hosted at the University of Amsterdam by Professor Dyvendak and his colleagues included a series of detailed case studies regarding the structure of the long-term care service delivery systems in Greece, Germany, Italy, Poland, the Netherlands, England, Sweden, and Norway. The project was designed to assess the adequacy of political commitment to providing support services to the frail elderly and others with long-term care needs (Duyvendak et al. 2009). No consistent information about the supply of home care services was available across all the countries, other than statements that formal home nursing and care aide services provided by established and authorized entities were largely unavailable in countries like Greece, Italy, and Poland whereas in countries like Sweden and Norway, all municipalities have these kinds of service agencies. This is consistent with OECD data indicating that the proportion of the elderly population receiving any formally provided long-term services and supports was very low in the southern European countries but much higher in northern European countries.

One of the recent issues facing all OECD countries are projections of the number of needed long-term care workers relative to the growing number of frail, aged individuals in the population (Columbo et al. 2011). Projections for the numbers of workers relative to the size of the population in need conducted by the European Network of Economic Policy Institutes strongly point to the fact that most countries will face a significant

labor shortage (Mot and Willeme 2012). Indeed, demand for formal institutional care, as well as formal home care support services, are anticipated to increase from 100% to over 200% between 2010 and 2060 in the face of a relative flattening of the available number of informal caregivers and a projected decline in the number of long-term care workers employed in the formal sector. These projections for the four countries examined as part of the ANCIEN project are consistent with other European countries, and the authors suggest that policy makers face major challenges in the coming years as demand outstrips the supply of caregivers, both informal family members and formal agency employees. The only way to increase the supply of formal care workers is to substantially alter their compensation and improve their working conditions, both of which will dramatically increase the costs of services that are already projected to bankrupt countries based only upon changing demography.

### **Informal Care Provision and Cash Payments for Dependent Care Allowances**

While policies offering cash payments to frail elders and their family members are discussed under the financing section above, these policies have a direct bearing on the structure of the market for long-term care services for several reasons (Ungerson and Yeandle 2007). First, many countries which have some form of cash payments provided to eligible frail elders and their families allow those funds to be used by family members, ostensibly as compensation for foregone labor force activity (Wiener 2007). Second, unless there are explicit limitations on the use of such cash transfers, according to the limited empirical research that has been done on the issue, recipients and their families appear to be more likely to purchase unskilled household and personal care help from the unregulated labor market. That is, there are many reports documenting that this kind of work is frequently done by undocumented workers (Bettio and Solinas 2009). Third, the interrelationship between cash transfer programs

from long-term care insurance and the role of informal care and the undocumented “grey” labor force for domestic help has begun to receive considerable attention in EU and OECD countries in the context of the raging debate regarding illegal immigration and the cost of employment. While it is not the place of this chapter to address these issues thoroughly, the implications for the organization of long-term services and supports are considerable since, many would argue that the growth of the illegal labor market for domestic help with the aging of the population undermines the development of a robust home care services system. It is this issue we address in the final paragraphs of this segment of the chapter (van Hooren 2008).

Informal care is the dominant source of support for most community dwelling elderly throughout the developed world. While many have argued that the availability of formal agency support undermines and substitutes for endogenous informal care from families, the evidence both from microeconomic and macroeconomic studies do not support this contention (Tarricone and Touros 2008; Rothgang 2003; Foster et al. 2007). While the proportion of the frail elderly in northern Europe who receive some form of formal care services from municipalities is much higher than in southern Europe, the proportion receiving assistance from families and friends is similar, although the relative share of support may be more highly weighted toward formal service providers (OECD 2013a).

As noted, numerous OECD countries have some form of cash payment system for frail elders and their families who require long-term care services and supports. While there are many details that differentiate the manner in which the funds are provided and the conditions under which they can be used, it is clear that these are extremely popular programs (Columbo et al. 2011). Indeed, the popularity of the cash option is best seen in the fact that the vast majority of German households eligible for support under the long-term care insurance law elect to receive cash rather than services even though the value of the cash is far less than the replacement cost of the service. In the USA the success of the “cash and

counseling” demonstration in three states was the stimulus for major expansions of this option for Medicaid programs across the country since both family members serving as informal caregivers and the workers hired by the family and patients reported improved satisfaction with their circumstances when compared to the control group (Foster et al. 2007; RWJF 2013).

One of the key provisions of these cash allowance programs is the extent to which the use of the funds by recipients and their families are regulated, that is, how the money is spent is predetermined and/or whether there are restrictions on the kinds of workers that can be hired (van Hooren 2008). In a comparative policy analysis, van Hooren appears to suggest that countries with little restriction on how such cash allowances are used are associated with higher proportions of eligible households using illegal domestic workers. Although the data are necessarily limited, comparisons of Italian families’ use of undocumented domestic workers to care for the elderly and the virtually nonexistent use of such workers in the Netherlands suggests that families in countries with better developed formal agency community-based services will rely less upon the unregulated labor market to care for frail elders (van Hooren 2008; Simmonazzi 2009).

In the USA, recent policies have expanded the applicability of “cash and counseling” programs which encourage income and functionally eligible elders and their families to use their cash allowance to arrange for personal care attendants and home care assistants directly, thereby making their money stretch. Like many of the European cash transfer programs, the autonomy inherent in such self-directed care seems to promote an informal economy with workers, whether legally documented or undocumented, being paid without standard employer benefits such as holidays, vacation, and sick time which would be expected by a worker employed by an agency. For agencies to compete for labor when they have to withhold taxes and other payments from employees (not necessarily required for cash payments) as well as for customers, who do not necessarily want to pay for the higher cost of agency-supervised

workers, places them in a highly disadvantaged position. Many US states committed to implement such “consumer directed care” programs are exploring ways of creating a labor market that elderly and disabled clients and their families and advocates can rely upon which also have provisions for paying workers’ benefits while indemnifying the care recipient. (See the “Cash and Counseling” Resource Center Web site for presentations and discussions of care worker training, benefits payment, and liability insurance. [http://www.bc.edu/schools/gssw/nrcpds/cash\\_and\\_counseling.html](http://www.bc.edu/schools/gssw/nrcpds/cash_and_counseling.html).)

Over the next decades, OECD countries will have to devise more cohesive policies to support the population’s desire to receive care at home without inadvertently stimulating demand for undocumented workers and illegal immigration which, in turn, undermines the ability of a well-functioning formal market in long-term care services. Northern European countries, like wealthier US states, have well-developed home care agency structures, whether publicly or privately operated, precisely because they have invested in this sector of the long-term care system, whereas southern European countries, like their US southern state counterparts, have relied more extensively on family caregiving which, when unable to meet elders’ needs, seeks to purchase assistance from the informal labor market. How to devise financing and reimbursement policies as well as regulatory structures to address these issues presents a major challenge to developed countries.

---

## Regulating Quality

Ensuring the quality of long-term care involves more than just putting into place regulatory rules and procedures to govern the licensure (registration) and certification of providers of long-term care services. It also involves having systems to monitor the safety, effectiveness, and success of those services in terms of maintaining the well-being, health outcomes, and dignity of long-term care recipients. In practice, the latter goal is more difficult to achieve and

requires increased investment in time, skilled staff, and resources to ensure appropriate reporting dimensions, standardized data gathering procedures, and consistent assessment protocols that can then feed into quality improvement measures. However, just as there is a great deal of variation in how countries organize and finance long-term care, there are also differences in the regulatory approaches to assuring quality.

## Different Regulatory Approaches to Quality Assurance

In a recent comparative study that we conducted on the regulation of long-term care quality in 14 countries, we identified three main approaches that underpin the quality assurance frameworks in the countries with relatively well-developed long-term care systems (Mor et al. 2014). The first approach, as seen in countries such as Austria, Germany, Japan, and Switzerland, delegates the main responsibility for upholding standards, training, and staff certification requirements for the long-term care workforce, as well as for monitoring quality, to professional organizations. In this approach, government is a partner in quality assurance rather than assuming a primary “policing” role. While government is still involved in setting standards for long-term care via legislation, the “professionalism-based” approach to quality regulation places considerable trust in “self-regulation.” This position is predicated on the assumption that associations of professionals involved in long-term care have distinctive expertise that the state can rely upon to ensure their commitment to training and ethical good practices in caring for the elderly.

In contrast, a second approach (followed in countries such as Australia, England, the Netherlands, and Spain) is much more empirical and inspection-based, where government authorities assume the primary role in rule-making and monitoring providers’ compliance with statutorily defined regulations. This “inspection-based” approach stresses the need for close oversight

by central authorities as there is generally less societal confidence that professionals, or providers, will always act in the interests of frail elderly people using long-term care services. A third approach, in place in Canada, Finland, New Zealand, and the United States, builds upon the existing inspection-oriented approach to licensing, inspection, and complaints investigation by adding quality measurement and public reporting protocols based on intensive data gathering and analysis. This “data management and public reporting” approach emphasizes standardization and reporting of data so that long-term care users, ideally, can act as consumers and choose the best services suited to their needs, with quality being boosted by market competition among providers. The best example of this approach can be seen in the United States where the RAI Minimum Data Set (MDS) (The MDS is a Resident Assessment Instrument (RAI) which is required in all US nursing homes in order to ensure that a resident’s care plan is based upon a comprehensive assessment of their needs.) is used and the government’s *Nursing Home Compare* Web site reports information on a range of MDS-based assessment measures for short- and long-stay nursing home residents [See <http://www.medicare.gov/nursinghomecompare/search.html>].

### The Regulatory Reach of Quality Monitoring

Monitoring the quality of long-term care can address structures, processes, and, less often, outcomes. It is also useful to divide quality regulation functions in terms of three broad domains: (1) standard setting and initial inspection and licensure, (2) ongoing surveillance and enforcement, and (3) reporting and/or rewarding performance. Table 5 summarizes a wide selection of these regulatory functions in a selection of OECD countries, with the check marks in the columns indicating that a particular function is an integral part of the quality regulation regime in the particular country. It is important to note that the table rows includes quality assurance functions that are

mainly applicable to the residential care sector as quality regulation of home care agencies is very underdeveloped. Moreover, a check mark refers only to the fact that the regulatory function takes place and does not purport to indicate the effectiveness of the regulations or the overall quality of care.

As can be seen, the first four rows of Table 5 look at structural standards that are relevant to the licensing of long-term care providers. All countries in this sample require providers to register with designated authorities and in the case of residential facilities must demonstrate that requirements for the physical plant (such as fire and safety arrangements and quality of life considerations such as room size) are met. Moreover, the OECD reports that in two-thirds of its member countries accreditation or certification of care facilities is compulsory, a condition for reimbursement and contracting or common practice (OECD 2010). In addition, formal regulations govern the level of education and training that groups of long-term care workers (e.g., registered nurses, personal care workers) must attain in order to be employed by a long-term care provider. However, it is noteworthy that the levels of required training as well as experience vary markedly among countries. For example certified care workers need 75 h of training and experience in the United States, 430 h in Australia, 75 weeks in Denmark, and 3 years in Japan (OECD 2010; OECD/European Commission 2013; Table 4). In addition, market conditions, such as local unemployment rates or the availability of excess labor (such as illegal immigrants), has a big influence on the strictness with which providers apply these professional standards. Another consideration is the cost implications for mandating minimum training of staff in formal care settings. Along with minimum wages, social security contributions, and other labor-related overhead, training requirements add to higher wage costs in the formal care sector, making formal long-term care services too expensive for many users, particularly in countries where public coverage is limited. In such cases, informal care from relatives or hiring cheaper care workers from an available pool of migrant workers in the “grey labor market” becomes the





13. Complaint collection and monitoring system	X	X	X	X	X	X	X	X	X <sup>b</sup>	X	X	
14. Telephone or Web-based action-line complaint process	X	X	X	X	X	X	X	X	X			
15. Public reporting	X	X	X	X	X	X	X <sup>n</sup>	X	X	X	X	
16. Consumer choice data	X	X	X	X	X					X		
17. Pay-for-Performance quality assurance	X	X										

Source: Mor et al. (2014)

Notes:

<sup>a</sup>Functions can vary slightly across nursing homes and community-based options.

<sup>b</sup>Varies across regions.

<sup>c</sup>Regulations pertaining to structural aspects are very broadly specified for the most part.

<sup>d</sup>Standards for all groups exist but enforceability depends on the staff group (e.g., nurse, social worker, care worker).

<sup>e</sup>Industry associations of aged care providers and nursing professional bodies.

<sup>f</sup>Some provider regulations refer to aspects of the care process but these are broadly specified. Within the regulations, cross-references are made to best practice guidelines, but inspectors seem to have a good deal of leeway on how to interpret this standard. NICE is developing quality standards which set out care process minimum standards, but it is not yet clear to what extent these are enforceable or merely guidelines.

<sup>g</sup>Legislation sets out a Schedule of Specified Care and Services and requires providers to deliver care of “an appropriate standard.” However, the legislation does not set out minimum staff-resident ratios or hours of care.

<sup>h</sup>Poorly or variably enforced.

<sup>i</sup>There are no fines for poor care per se. However, poor providers can be sanctioned by having government funding withheld for new residents until they meet care standards. This is a form of financial sanction.

<sup>j</sup>These exist but are poorly enforced and with very low level of fines.

<sup>k</sup>Contracts can be terminated by the Long-term Care Fund.

<sup>l</sup>Appeals can be brought against registration or outcome of inspections.

<sup>m</sup>Providers have appeal rights against regulatory decisions. Care recipients have appeal rights against a decision by the government to not to approve them for subsidized care.

<sup>n</sup>Limited to inspections.

only viable alternative, especially where unregulated cash benefits are available to further incentivize this solution (van Hooren 2008) (See also section “[Structure of The Delivery System](#)” above). Obviously, the use of unskilled informal, hired, carers may have consequences for the quality of care provided – although many positive benefits of informal or hired care, such as fostering empowerment in the care user and building strong relations of trust, also have been reported (Columbo et al. 2011; Dale et al. 2005).

One final aspect in this group of functions is the role of professional organizations and/or independent, nongovernment organizations, in helping to set standards for long-term care providers (Row 4). Again, there is wide variability among countries in both the participation rates of such organizations and the rigor with which they pursue their roles in standard setting, i.e., whether they actively participate in developing benchmarks for best practice (e.g., in Austria and Japan) or whether they tend to limit their role to advocating in favor of minimum standards (USA).

Rows 5 to 12 consider different functions associated with ongoing monitoring and enforcement as captured by inspection regimes. Such monitoring focuses in particular on process standards that are applied to encourage positive aspects of care (such as weight monitoring, wound monitoring, fall prevention, and infection control) or to prohibit practices that often have a negative impact (e.g., the use of physical restraints or the use of antipsychotic medications) as well as the sanctions that can be imposed in cases of poor performance and/or noncompliance. Such standards exist in several countries to minimize the use of these behavior control schemes by requiring extensive documentation to justify their use, and making them the subject of inspection. In contrast, the monitoring of resident’s health or well-being outcomes as an aspect of quality control is a relatively advanced and complex objective, involving the definition of such “outcomes” for frail elderly individuals and then establishing standardized data systems that carers and inspectors can use to determine whether such outcomes have been achieved.

The best example of a such an assessment tool is the Resident Assessment Instrument Minimum Data Set version (RAI-MDS) applied to nursing home contexts and the versions developed for the assessment of individuals receiving home care (interRAI-HC) and care in community settings (inter-RAI CHA). First developed in the United States in the late 1980s and subsequently extended through an international consortium of experts, the current comprehensive suite of assessment instruments are standardized tools to detect long-term care users’ strengths, needs, and potential risks to enable individualized monitoring and care planning. In addition, collected data are aggregated to produce quality indicators on processes and outcomes both at the individual and facility/organizational level (Mor et al. 2010; Hutchinson et al. 2010). The mandated use or testing of RAI assessment instruments internationally has been growing over the last decade, with a presence in several countries in North America, Europe, South-East Asia, and Australasia (see <http://www.interrai.org/worldwide.html>).

Focusing directly on approaches to inspection, we can see from Table 5 that after a provider has been certified or licensed, routine inspections are almost universally carried out, albeit according to different time-frames and conditions that may trigger an inspection (Rows 7–9). For example, inspections may take place every few years, or they may be less frequent based on a provider’s good performance in previous inspections; the latter kind of “risk-based” regulation relies on historical inspection data and using it to shape the regularity and intensity of subsequent inspections. Alternatively, a desk audit of data submitted in advance by a provider may take place in some circumstances, either in lieu of or prior to an on-site inspection. In addition, ad hoc inspections may be triggered by a complaint by a resident or family member, and the assessor will often investigate both the source of the particular complaint as well as seek to document other problems in the same care domain as the complaint. On-site inspections may also be carried out according to regular schedules or be random (unannounced) with the aim of observing providers in carrying

out their day-to-day care duties with no prior notification. To date, however, there is no empirical data on the efficacy of one approach over the other in terms of stimulating better quality of care (Mor et al. 2014).

Once inspections uncover a problem with an aspect of care, different regulatory frameworks employ various means to rectify the problem (Rows 10–12). Some countries prefer to view inspections as collaborative, compliance-based exercises in which inspectors first work in tandem with providers to find solutions to the identified deficiencies. This may take the form of informal negotiations and persuasion. In other cases, more formalized “deterrence-based” procedures are used, such as issuing warnings to return to compliance within a given timeframe. Most systems, however, do have some form of official sanctions if providers fail to respond adequately or if repeated cases of noncompliance are found. By linking quality to financial penalties, the ability of regulators to levy fines or other penalties against poorly performing providers represents one way of incentivizing improvements. This can be done by fining the provider directly, restricting further admissions (and therefore potential revenue), and/or in countries with public long-term care financing, withholding reimbursement until the specific problem is fixed.

A last-resort and very rarely applied sanction is decertification, or revocation, of a provider’s license to operate (Angelelli et al. 2003). Regulators are often reluctant to impose this ultimate penalty as relocating frail elderly residents from a facility that has been sanctioned with closure would mean finding other suitable and available places in the same area and may also potentially impose “relocation” stress on residents. It is also worth noting that all of the compliance enforcement methods mentioned above may be subject to lengthy procedural requirements that regulators have to adhere to and/or legal appeal processes open to sanctioned providers. These processes often involve considerable periods of time before a noncompliance issue is resolved.

Finally, the quality of long-term care can be monitored via various means to report on providers’ performance, whether through established

complaints channels or making available systematic data on provider quality performance. In some countries, financial incentives (such as pay-for-performance tariffs) may be in place through public funders of long-term care to encourage providers to participate in quality assurance programs (Rows 13–17). Although complaints monitoring data is scarce, systems for submitting complaints about the treatment of long-term care recipients exist in most countries, again with substantial variation in the means available (e.g., written complaints or telephone/internet action-lines) and the requirements for responding to such complaints.

Of equal saliency is making information about providers’ quality performance available to consumers. This could take the form of presenting the results of recent inspections or supplying specific performance data on providers through easily accessible media such as the Internet. For example, in the USA various measures of quality, ranging from staffing levels, to inspection results, to indicators of process and outcome quality are computerized and posted on government Web sites. In Finland, data on residents’ outcomes are voluntarily fed back to providers with the intention that ultimately this information (particularly if it is positive) might be used by the providers themselves to inform potential long-term care consumers in their areas. A similar structure is in place in New Zealand for their home care agencies, with plans to extend the practice to residential care facilities. This transparency can have several advantages. Firstly, if directly available to consumers, performance data based on various indicators can inform choices in selecting a long-term care provider (Werner et al. 2009). Secondly, public reporting of providers’ quality can exert pressure on them to address problems and maintain standards, particularly if they are performing poorly and do not wish to compromise their reputation in the local long-term care market. Thirdly, having access to performance data, particularly on comparable quality indicators across a number of providers (either local or national), can feed into individual providers’ voluntary quality improvement strategies (Werner and Konetzka 2009).

## Challenges Facing Quality Monitoring

While comparative international information is still quite scarce, available sources (Mor et al. 2014) highlight that despite how developed or underdeveloped a long-term care system is, regulatory frameworks and some form of monitoring activities always exist for residential facilities (nursing homes). In contrast, due to the existence of informal care arrangements as well as less developed quality assurance programs applicable to formal home care agencies, there is much less regulation and knowledge about the regulation of home care, although some countries (such as the Netherlands, New Zealand, Canada, the USA, and Switzerland) have made inroads into monitoring the quality of home care. Given the relative growth of the latter sector and the general preference by long-term care users to remain in their own home for as long as feasible (European Commission 2008) the relative scarcity of quality assurance frameworks for home care settings will be a major but necessary challenge for governments in the future.

A second key challenge is the availability of data that can be used for quality assurance purposes – not only in the residential care sector, where data gathering can often be sporadic and not standardized across facilities, but also in the home care sector, where data are even more limited. As far as European countries go, it is still the case that standardized information derived from the inspection process is not routinely collected nor archived for subsequent use. One reason for this is the difficulty of standardizing inspections, or assessments, across different regions, particularly if this function is decentralized to lower levels of government administration (such as municipalities or local agencies) or to providers themselves. Another hurdle is variability in the interpretations and evaluations of individual assessors. Thus, the lack of consistency hinders meaningful comparisons across providers. Indeed, even in the USA where standardized inspection protocols are in place and computerized, there is substantial interstate variation in the conduct and results of inspection. (Mukamel et al. 2012).

Related to this is the issue of whether any collected data is made publicly available. One stumbling block is the opposition of providers to sharing information not only with the public but specifically with competitors, particularly if there is a reputational risk involved in releasing information about poor performance. While some countries have started to collect inspection-based data, very few follow the example of the United States where the availability of such data on residential services helps would-be residents to make choices about what nursing home facilities would best suit their needs or to vote with their feet if their existing facility falls short. The exercise of choice based on quality data also extends to health and long-term care insurers who in the future will increasingly have to make purchasing decisions about competitive service suppliers based on both cost-effectiveness and the quality of care. Indeed, this process is already taking place in the USA, Canada, Finland, and New Zealand.

A final consideration is affecting a sea change in the attitude to quality monitoring and its role in incentivizing improvements in the quality of long-term care. While the primary purpose of measuring long-term care structures, processes, and outcomes is to ensure the safety and dignity of service users, the information harnessed by this process can be an invaluable tool for providers to assess the relative quality of their performance against relevant benchmarks, be it officially set standards or industry averages. Armed with such meaningful data, and skilled staff to interpret it, long-term care providers would then be in a much better position to shape their improvement strategies, not only to enhance their marketability but more importantly, for the benefit of the elderly clients in their care.

---

## Summary and Conclusions

Governments of the rapidly aging industrialized countries are just beginning to be aware of the enormous challenges they will face in meeting the care needs of the frail elderly. Over the last decade or so, most countries have begun the difficult process of rebalancing the provision of long-

term care from a system that was almost entirely supporting residential or institutional care to one in which the majority of service recipients were cared for in their homes. This shift was accelerated, or made possible, in many countries by the introduction of direct cash payments to eligible individuals and their families, allowing them to direct their long-term care mix of services using entitlement funds for which they are eligible because of their need for functionally based support.

There are several consequences that these shifts in care orientation have brought about. First, giving older consumers and their family members control over who they hire has substantially altered the labor market for long-term care services, particularly in those areas where well-developed agency-based long-term home care services do not exist. Indeed, the availability of some financial support may allow late middle-aged children of frail elderly persons to remain out of the formal workforce or provide another reason for these individuals to retire early, becoming full time, partially paid caregivers to their aged and frail parent. Second, as we have seen, monitoring the quality of care and services rendered to frail older persons is difficult enough when only provided in large residential care settings. Administrative procedures for reporting staffing levels and quality as well as documenting services rendered are sufficiently burdensome that many countries do not require this form of reporting. Furthermore, hiring independent inspectors to monitor the performance of these institutional providers constitutes a large expense even if facilities are only inspected annually. However, to truly monitor quality issues requires more frequent inspections, unannounced inspections, and inspections instituted in response to residents' and families' complaints. While this constitutes a very difficult task in the case of residential care, monitoring quality in the home care setting, much less, policing family members' own provision of care to the frail older person in their own homes is considerably more complex and costly, requiring close collaboration with what in the USA is known as "adult protective service" given the real potential for abuse.

Extending quality measurement to the home care setting, particularly if including frail older persons receiving cash payments which they apply to paying family or undocumented and unlicensed workers, presents numerous challenges. Home care providers in the USA, many Canadian provinces, and New Zealand have implemented individualized quality metrics as part of a routine client assessment process, and these data have been used to report on provider quality (Mor et al. 2014). These experiences suggest that the use of this kind of "microlevel" information is certainly a feasible approach to quality performance measurement of home care services, but they necessarily depend upon professionals periodically assessing the client and using those data to calculate indicators of quality performance. In the case of cash payments to family and informal labor market participants, this approach is not viable without introducing a mandatory assessment in recipients' homes, a process that may be perceived as excessively onerous, entailing an invasion of privacy. Furthermore, since most OECD countries have not even established a solid data reporting system in reference to institutional care provision and quality, it would be hard to imagine that most would be willing to institute an even more complicated and costly data-based approach to quality oversight of home care services.

Newer challenges which policy makers in the health care delivery space are increasingly worrying about is the linkage between reimbursement and quality measurement. Strategies to assure quality by applying "value-based purchasing" have been tried with limited success in the USA but are likely to emerge as the next emerging policy debate. To even consider this approach, however, there is a critical need for consistent data about patients' outcomes in selected areas and providers' characteristics and services. To date, these types of data exist in only a few countries, but the complexities of introducing such systems are substantial even after the data collection and assembly challenges have been met. It is likely, however, that as demand for various types of long-term care services increase due to

population aging and inadequate private savings among the elderly, public support for long-term care services may well be contingent upon those services being viewed as value for money.

## References

- Alakeson V. International development in self-directed care. *Issue Brief (Commonw Fund)*. 2010;78:1–11.
- Angelelli J, Mor V, et al. Oversight of nursing homes: pruning the tree or just spotting bad apples? *Gerontologist*. 2003;43(2):67–75.
- Bettio F, Solinas G. Which European model for elderly care? Equity and cost-effectiveness in home based care in three European countries. *Econ Lavoro*. 2009;43(1):53–71.
- Carpenter I, Hirdes J. A good life in old age: monitoring and improving quality in long term care. *OECD Health Policy Studies*, OECD Publishing; 2013. <https://doi.org/10.1787/9789264194564-en>
- Columbo F, Llena-Nozal A, Mercier J, Tjadens F. *Help wanted? Providing and paying for long-term care*. Paris: OECD Publishing; 2011.
- Dale S, Brown R, Phillips B, Carlson BL. How do hired workers fare under consumer-directed personal care? *Gerontologist*. 2005;45(5):583–92.
- Damiani G, Farelli V, Anselmi A, Sicuro L, Solipaca A, Burgio A, Iezzi DF, Ricciardi W. Patterns of long term care in 29 European countries: evidence from an exploratory study. *BMC Health Serv Res*. 2011;11:316.
- Doty P, Mahoney KJ, Sciagaj M. New state strategies to meet long-term care needs. *Health Aff*. 2010;29(1):49–56.
- Duyvendak JW, Grootegoed E, Savernije MT, Tonkens E. Day 1: long-term care in Europe, the state of the art. Presentation given at does Europe care? European Conference on Long-Term Care and Diversity, Amsterdam; 2009. <http://www.careconference.eu/site/sites/default/files/Part201.pdf>.
- European Commission. *Long-term care in the European Union*. Brussels: Commission of the European Communities, DG Employment, Social Affairs and Equal Opportunities; 2008.
- Feng Z, Lepore M, Clark MA, Tyler D, Smith DB, Mor V, Fennell ML. Geographic concentration and correlates of nursing home closures: 1999–2008. *Arch Intern Med*. 2011;171(9):806–13.
- Fernandez JL, Forder J, Trukeschitz B, Rokosová M, McDauid D. How can European States design efficient, equitable and sustainable funding systems for long-term care for older people? Copenhagen: World Health Organization and World Health Organization on behalf of the European Observatory on Health Systems and Policies; 2009.
- Foster L, Dale S, Brown R. How caregivers and workers fared in Cash and Counseling. *Health Serv Res*. 2007;42(1 Pt 2):510–32.
- Grabowski DC, Cadigan RO, Miller EA, Stevenson DG, Clark M, Mor V. Supporting home- and community-based care: views of long-term care specialists. *Med Care Res Rev*. 2010;67(Suppl 4):82S–101S.
- Hutchinson AM, Milk DL, Maisey S, Johnson C, Squires JE, Teare G, Estabrooks CA. The resident assessment instrument-minimum data set 2.0 quality indicators: a systematic review. *BMC Health Serv Res*. 2010;10:166. <https://doi.org/10.1186/1472-6963-10-166>.
- Ikegami N, Fries BE, Takagi Y, Ikeda S, Ibe T. Applying RUG-III in Japanese long-term care facilities. *Gerontologist*. 1994;34(5):628–39.
- Ikegami N, Morris JN, Fries BE. Low-care cases in long-term care settings: variation among nations. *Age Ageing*. 1997;26(Suppl 2):67–71.
- Ikegami N, Ishibashi T, Amano T. Japan's long-term care regulations focused on structure – rationale and future prospects. In: Mor V, Leone T, Maresso A, editors. *Regulating long-term care quality: an international comparison*. Cambridge: Cambridge University Press; 2014.
- Jung H-Y, Jang S-N, Seok J-E, Kwon S. Quality monitoring of long-term care in the Republic of Korea. In: Mor V, Leone T, Maresso A, editors. *Regulating long-term care quality: an international comparison*. Cambridge: Cambridge University Press; 2014.
- Kane RA, Kane RL, Ladd RC. *The heart of long-term care*. New York: Oxford University Press; 1998.
- Katz MB. *In the shadow of the Poorhouse: a social history of welfare in America*. Tenth anniversary edition. New York: Basic Books; 1996.
- Kellogg DO. The pauper question. *Atl Mon*. 1883;51(307):638–652.
- Larson Allen L. *Mapping the future: estimating Florida aging service needs 2008–2030*. Tallahassee: Agency for Health Care Administration; 2008.
- Miller EA, Mor V, Clark M. Reforming long-term care in the United States: findings from a national survey of specialists. *Gerontologist*. 2010;50(2):238–52.
- Miller EA, Tyler DA, Rozanova J, Mor V. National newspaper portrayal of U.S. nursing homes: periodic treatment of topic and tone. *Milbank Q*. 2012;90(4):725–61.
- Mor V, Miller EA, Clark M. The taste for regulation in long-term care. *Med Care Res Rev*. 2010;67(Suppl 4):38S–64S.
- Mor V, Leone T, Maresso A, editors. *Regulating long-term care quality: an international comparison*. Cambridge: Cambridge University Press; 2014.
- Mot E, Willemé P, editors. *Assessing needs of care in European nations*, ENEPRI policy brief no. 14, vol. 2012. Centre for European Policy Studies: Brussels; 2012.
- Mukamel DB, Weimer DL, Harrington C, Spector WD, Ladd H, Li Y. The effect of state regulatory stringency on nursing home quality. *Health Serv Res*. 2012;47(5):1791–813.
- OECD. *Ensuring quality long-term care for older people*. Paris: OECD Publishing; 2010. Policy Brief.

- OECD. Recipients of long-term care. In: *Health at a glance 2013: OECD indicators*. Paris: OECD Publishing; 2013a. [https://doi.org/10.1787/health\\_glance-2013-75-en](https://doi.org/10.1787/health_glance-2013-75-en)
- OECD. OECD health data: long-term care resources and utilisation. Paris: OECD; 2013b.
- OECD/European Commission. *A good life in old age? Monitoring and improving quality in long-term care, OECD health policy studies*. Paris: OECD Publishing; 2013. <https://doi.org/10.1787/9789264194564-en>
- Phillips B, Schneider B. Commonalities and variations in the Cash and Counseling programs across the three demonstration States. *Health Serv Res.* 2007;42 (1 Pt 2):397–413.
- Rothgang H. Long-term care for older people in Germany. In: Comas-Herrera A, Wittenberg R, editors. *European study of long-term care expenditure. Investigating the sensitivity of projections of future long-term care expenditure in Germany, Spain, Italy and the United Kingdom to changes in assumptions about demography, dependency, informal care, formal care and unit costs. Report to the European Commission, Employment and Social Affairs DG: 24–42*. 2003. [http://ec.europa.eu/employment\\_social/soc-prot/healthcare/lc\\_study\\_en.pdf](http://ec.europa.eu/employment_social/soc-prot/healthcare/lc_study_en.pdf)
- RWJF – Robert Wood Johnson Foundation. *Executive summary: cash and counseling program*. Princeton: Robert Wood Johnson Foundation; 2013. Available at: [http://www.rwjf.org/content/dam/farm/reports/program\\_results\\_reports/2013/rwjf406468/subassets/rwjf406468\\_1](http://www.rwjf.org/content/dam/farm/reports/program_results_reports/2013/rwjf406468/subassets/rwjf406468_1)
- Simmonazzi A. Home care and cash transfers. Effects on the elderly care-female employment trade-off. *Cost Conference*. Rome; 2009.
- Sloane PD, Zimmerman S, Gruber-Baldini AL, Hebel JR, Magaziner J, Konrad TR. Health and functional outcomes and health care utilization of persons with dementia in residential care and assisted living facilities: comparison with nursing homes. *Gerontologist*. 2005;45 Spec No 1(1):124–32.
- Smith DB, Feng Z, Fennell ML, Zinn JS, Mor V. Separate and unequal: racial segregation and disparities in quality across US nursing homes. *Health Aff.* 2007;26 (5):1448–58.
- Stevenson DG, Grabowski DC. Sizing up the market for assisted living. *Health Aff.* 2010;29(1):35–43.
- Swartz K. Searching for a balance of responsibilities: OECD countries' changing elderly assistance policies. *Annu Rev Public Health.* 2013;34:397–412.
- Tarricone R, Touros AD, editors. *The solid facts: home care in Europe*. Copenhagen: World Health Organization Regional Office for Europe and Università' Commerciale Luigi Bocconi; 2008.
- Ungerson C, Yeandle S. *Cash for care in developed welfare states*. Houndmills: Palgrave Macmillan; 2007.
- van Hooren F. Bringing policies back in: How social and migration policies affect the employment of immigrants in domestic care for the elderly in the EU-15. Paper presented at Transforming elderly care at local, national and transnational level, International Conference at the Danish National Centre for Social Research (SFI), Copenhagen; 2008.
- Werner RM, Konetzka RT. What drives nursing home quality improvement under public reporting? An examination of post-acute care. *Chicago: AcademyHealth*; 2009.
- Werner RM, Konetzka RT, Stuart EA, Norton EC, Polsky D, Park J. Impact of public reporting on quality of postacute care. *Health Serv Res.* 2009;44 (4):1169–87.
- Wiener JM. *Commentary: cash and counseling in an international context*. *Health Serv Res.* 2007;42(1 Pt 2):567–76.
- Wirmann Gadsby E. *Personal budgets and health: a review of the evidence*. London: PruComm. Policy Research Unit in Commissioning and the Health Care System, Department of Health; 2013.



# Provision of Health Services: Mental Health Care

# 44

Jon Cylus, Marya Saidi, and Martin Knapp

## Contents

<b>Introduction: Why Is Mental Health Important?</b> .....	980
<b>Definitions and Spectrum of Mental Health Disorders</b> .....	980
<b>Direct and Indirect Costs</b> .....	981
<b>Stigma</b> .....	982
<b>Comorbidity</b> .....	983
<b>Provision of Mental Health Care: How Is Care Delivered?</b> .....	983
Who Delivers Care: Medical Professionals, Unpaid Caregivers .....	983
<b>Financing Mental Health Services: How Is Care Financed?</b> .....	986
<b>Key Policy Dimensions/Recent Policies and Trends</b> .....	986
Personalization and Empowerment .....	987
Carer and Family Impact .....	987
Prevention, Promotion, Public Mental Health (e.g., Campaigning) .....	988
Aging and Dementia .....	989
Employment .....	990
New Advancements in Treatments and Technologies .....	992
<b>Discussion</b> .....	992
<b>References</b> .....	994

## Abstract

Despite estimates that one in four people experience a significant episode of mental illness during their lifetime (Kohn et al., *Bull World Health Organ* 82:858–866, 2004), mental health

remains a much neglected issue. Throughout the world, the level of resources dedicated to mental health is incommensurate with its prevalence or with its burden on society. Stigma and social exclusion can make it harder for people with mental health problems to obtain and maintain work, access appropriate health services, participate in their communities, or enjoy family life. Public attitudes toward mental illness, although showing signs of improving in recent years, are often negative and sometimes

J. Cylus (✉) · M. Saidi · M. Knapp  
The London School of Economics and Political Science,  
London, UK  
e-mail: [j.d.cylus@lse.ac.uk](mailto:j.d.cylus@lse.ac.uk);  
[m.saidi1@lse.ac.uk](mailto:m.saidi1@lse.ac.uk); [m.knapp@lse.ac.uk](mailto:m.knapp@lse.ac.uk)



discriminatory. A recent Eurobarometer (2010) survey found that two thirds of EU citizens reported feeling uncomfortable talking to someone with a “significant mental health problem,” while one in five found it “difficult.”

---

## Introduction: Why Is Mental Health Important?

Despite significant gaps, awareness of mental health problems continues to improve. Many countries, particularly in Europe, have taken steps to develop or modernize their mental health policy frameworks highlighting the need to try to prevent mental health problems, to raise awareness of them when they arise, and to improve the volume and quality of resources available for treatment and care. However, frameworks in many other countries remain outdated with their mental health-care capacity correspondingly deficient. Recently the World Health Organization launched its Mental Health Action Plan 2013–2020 (World Health Organization 2013), which introduced six crosscutting principles that they suggested should be at the heart of the policy discussion for global mental health. Universal health coverage for all was emphasized, as well as the promotion human rights, in a way that all mental health strategies should be compliant with international and regional human rights instruments. Evidence-based practice should be promoted for mental health strategies as well as interventions for treatment, prevention, and promotion. Policies should adopt a life course approach, taking into account health and social needs at all stages and ages of life. Partnership between multiple sectors (health, education, housing, social, judicial, and other relevant sectors as well as the private sector) should be encouraged, taking into account local (country and regional) contextual factors. Finally, and perhaps for the first time, the WHO put empowerment at the forefront of their mental health agenda, stating that people with mental health problems should be involved in several aspects of policy.

Saxena et al. (2007) have highlighted the global scarcity of resources for mental health,

inequity in their access, as well as inefficiencies in their use; and these have serious consequences. Treatment gaps, defined as the nonreceipt of care when it is needed, still occur, and as many as a third of individuals with schizophrenia and other non-affective psychoses do not receive any treatment (Kohn et al. 2004).

In this chapter, we look at the state of mental health, mental health care, and policy across the world. We discuss how people with mental health problems are treated and how services designed to care for these individuals are financed. We end by focusing on key policy trends which will help to shape future priorities and actions.

---

## Definitions and Spectrum of Mental Health Disorders

The most widely used classifications of mental disorders are the *International Classification of Diseases-10* (ICD-10) (World Health Organization 1992) and the *Diagnostic and Statistical Manual-V*; the former is currently being revised and updated. The current form of the DSM represents its first update in more than two decades; however it has not gone without controversy. The controversy mainly surrounds the perhaps oversimplification of the description of autistic spectrum disorder (Wing et al. 2011).

Otherwise, the ICD-10 defines mental disorders as “the existence of a clinically recognisable set of symptoms or behaviour, associated in most cases with distress and with interference with personal functions.”

For the most part, the etiology of mental health problems is not fully known, but its determinants and risk factors can be grouped into three distinct categories: biological factors (e.g., heredity or physical diseases), psychological factors (e.g., traumatic experiences or early separation), and social factors (e.g., lack of social support and deprivation) (Lehtinen et al. 2007, p. 127).

In terms of definitions, common mental disorders (CMDs) are mental conditions that cause marked emotional distress and interfere with people’s daily functioning; however, they do not usually affect insight or functioning.

These comprise different types of depression and anxiety, and their symptoms include low mood and a loss of interest and enjoyment in usual activities. Anxiety disorders include generalized anxiety disorder, panic disorder, phobias, and obsessive and compulsive disorder (OCD). OCD, the most severe form of anxiety disorder, is characterized by a combination of obsessive thoughts and compulsive behaviors, where obsessions are defined as recurrent and persistent thoughts and impulses or images that are intrusive and inappropriate and cause anxiety or distress, while compulsions are repetitive, purposeful, and ritualistic behaviors or mental acts that are performed in response to obsessive intrusion and to a set of rigidly prescribed rules (National Centre for Social Research 2007). Psychoses are disorders of the mind that can produce disturbances in thinking as well as perceptions severe enough to produce distortions in perceptions of reality. Psychoses may also impair motivation and may be associated with affective dysregulation (depression, mania), as well as alterations in information processing (cognitive impairment) (Van Os et al. 2010). Van Os et al. conclude that overall, psychotic outcomes are associated with living in an urban area, being part of a minority group, cannabis use, and developmental trauma – hence is linked to the three risk factors described above. An important systematic review of the literature on the epidemiology of schizophrenia (McGrath et al. 2004) found a lifetime prevalence rate of 0.5–1%. Rates varied across the dimensions Van Os et al. suggested, as well as gender: schizophrenia was more common in males compared to females.

Kessler et al. (2009) studied the prevalence rates reported in the first 17 World Mental Health Surveys and found lifetime prevalence estimates of any DSM-V disorder to be 18.1–36.1%. These were the highest in Columbia, France, New Zealand, Mexico, the Netherlands, and South Africa and the lowest in China and Nigeria. Kessler et al. (2009) commented that the low prevalence rates in the last two countries may be downwardly biased. Anxiety disorders were consistently found to be the most prevalent class of mental disorder in the general population

(16%), again higher in Western developed countries. Mood disorders were also very common (12%) and were also mainly reported in Western countries. Kessler et al. suggested that a reason for their possible underestimation of prevalence rates in some countries may be because the DSM categories are less relevant to symptom expression in some countries than others.

---

## Direct and Indirect Costs

The latest data from the Global Burden of Disease study (2013) estimate that mental and behavioral disorders accounted for 198.3 million disability-adjusted life years (DALYs), with unipolar depressive disorders accounting for 37.8% of them. Anxiety disorders were the second biggest contributor, at 13.6% of DALYs of mental health. The WHO (2013) ranked mental and behavioral disorders the sixth leading cause of DALYs worldwide for 2011, surpassing respiratory diseases, neurological and sense organ conditions, musculoskeletal diseases, and endocrine, blood, immune disorders and diabetes.

Importantly, the burden of mental and substance abuse disorders had increased significantly since 1990. However, almost a third of countries – surprisingly perhaps – still do not have a designated budget for mental health; and 21% of the countries that do have a specific mental health budget spend less than 1% of their total health budgets on mental health (World Health Organization 2008).

People with mental health problems experience high rates of unemployment. For example, in OECD countries and depending on level of severity, people with mental health problems are between two to three times and six to seven times more likely to be unemployed compared to people without such conditions (OECD 2012). One reason for this difference is that illness can make it difficult to perform a job, but perhaps bigger problems are stigma and discrimination.

People with a history of mental health problems still face problems in the open employment market, including stigma, and a reluctance from employers to give them a job (McDaid 2008). The

fact that some people with mental health problems receive social security benefits also may also hinder their chances of seeking and obtaining employment (OECD 2011).

## Stigma

Stigma can be a “*mark of disgrace associated with a particular circumstance, quality or person*” (Oxford Dictionaries 2010), yet it is no longer physical or bodily in nature (Goffman 1963; Wahl 1999); it is now viewed as personal, psychological, and social. People are no longer physically branded but labeled by society as poor, homosexual, criminal, or, in this case, mentally ill. These labels have influenced perceptions and behaviors and lead to the devaluation and denigration of those who are so labeled (Thornicroft 2007).

Research on stigma in mental health has largely relied on attitude surveys and has been descriptive; very few studies have investigated this aspect from the standpoint of a person with mental illness. Discrimination against people with mental health problems has still been consistent in different parts of the world (Thornicroft et al. 2009). For example, in Ethiopia, key informants were asked about their perceptions of several different health and mental health conditions – they judged schizophrenia to be the most severe, and mental illness was frequently associated with talkativeness, aggression, and strange behavior (Alem et al. 1999). In the Arab world, there is much stigma associated with mental health services (Al-Krenawi and Graham 2000; Savaya 1995). Thornicroft et al. (2009) have reported high and consistent rates of experienced discrimination among people with schizophrenia across countries of various income levels. A cross-sectional survey conducted in 27 countries using face to face interviews with 732 participants with schizophrenia found that across all countries, the most common areas of negative experienced discrimination were seeking or maintaining friendships, discrimination by relatives, keeping and finding a job, and intimate or sexual relationships. Examining self-stigma and discrimination across several European countries, Brohan et al. (2010)

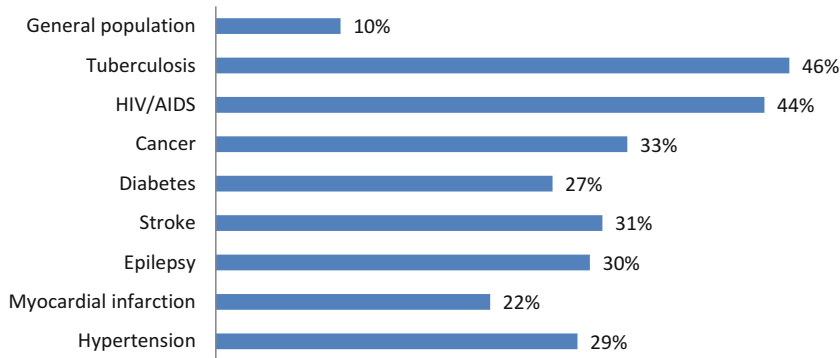
found that people with schizophrenia reported the highest self-stigma scores and perceived discrimination in Greece and the lowest empowerment scores in Ukraine. In this case, empowerment and increased social contact were significantly associated with reduced self-stigma scores.

The greatest barriers to social inclusion for people with mental health problems are said to be stigma and discrimination (Baldwin and Marcus 2011; Social Exclusion Unit 2004). Indeed, misconceptions about mental health can also lead to the belief that these diseases are untreatable and people who have them are not valued members of their communities, subsequently leading to appropriate support and resources not being delivered (Funk et al. 2012).

Lack of access to proper judicial mechanisms that would protect their rights (World Health Organization 2005b) means that people with mental health problems may often experience human rights violations in the community (Drew et al. 2005; Funk et al. 2005), and sometimes major life-changing decisions are made on their behalf with regard to housing or treatment, for example (World Health Organization 2005b). Where institutions still exist, living conditions are paltry and present risks to peoples’ physical health (Drew et al. 2011). Sharma (1999) reported on the state of mental hospitals found that many had undergone no structural transformations after previously having been jails. Other hospitals were at risk of serious overcrowding as single-person cells were used to house several patients. Others lacked any sanitary facilities and received inappropriate treatment.

Public education campaigns have produced mixed results (Thornicroft 2007, p. 244) and have perhaps only been reserved to certain countries. A concerted program in Australia called “beyondblue” aimed at conveying accurate information about depression, and its initial evaluations showed a series of benefits, including better community recognition of people with depression, reforms in life insurance and income protection, as well as intervention programs in schools (Ellis et al. 2002).

Combating mental health stigma has been at the forefront of mental health policy in England.



**Fig. 1** Prevalence of major depression in patients with physical illnesses (World Health Organization 2003a)

The Time to Change campaign led by two mental health charities promoted public mental health awareness. A study measuring its efficacy (Evans-Lacko et al. 2013c) suggested that their marketing tools – promoting social contact between members of the public and people with mental health problems – had positive outcomes on social stigma, and perhaps more so on behavior and attitudes, rather than knowledge (Evans-Lacko et al. 2013a). Smith (2013) commented that although the economic analysis seems to indicate benefits from the program, the assumptions used in the model seem to lead to uncertain conclusions: from a net cost to a benefit of £223 million.

## Comorbidity

Comorbidity is common within the mental health population, as 30% of all people with a long-term health condition also have a mental health problem (Cimpean and Drake 2011). Other estimates have shown that people with long-term conditions were up to three times more likely to experience a mental health problem compared to the general population (Naylor et al. 2012). Although much of the evidence relates specifically to affective disorders such as depression and anxiety (Naylor et al. 2012), studies have shown higher rates of conditions such as asthma, arthritis, cancer, and HIV/AIDS (Chapman et al. 2005; Sederer et al. 2006), among people with mental health problems, compared to people without. There are

also associated risk factors with the development of mild cognitive impairment or Alzheimer’s disease (Velayudhan et al. 2010); studies in Japan (Ohara et al. 2011) and Sweden (Xu et al. 2009) have also shown associations between diabetes and dementia.

In England, recent policy has focused on the association between poor health and mental health (Department of Health 2011), given these considerable costs – it is estimated that between £8 and £13 billion of NHS spending in England is due to comorbid mental health problems and long-term conditions (Naylor et al. 2012) – and burden to society (Department of Health 2011). Comorbidity is also associated with lower quality of life: utilizing data from the World Health Surveys, Moussavi et al. (2007) showed that people who suffered from depression as well as a long-term condition reported lower quality of life scores compared to people who only suffered from long-term conditions (Fig. 1).

## Provision of Mental Health Care: How Is Care Delivered?

### Who Delivers Care: Medical Professionals, Unpaid Caregivers

The occurrence of mental illness does not always require a need for treatment (Bebbington 1990). Nevertheless, much like its determinants, the treatment can be categorized into biological (such as psychotropic drugs), psychological

(or psychotherapies), and psychosocial (like case management and family interventions) (Lehtinen et al. 2007, p. 128).

Antipsychotics basically control the production of dopamine, the main neurotransmitter in the brain – the excess of which may play a part in producing hallucinations, delusions, and thought disorder and hence are mainly used for the treatment of schizophrenia. Older antipsychotics, such as haloperidol and chlorpromazine, depending on their dosage, have side effects which include stiffness and shakiness. In comparison, newer drugs, the most popular of which are clozapine and olanzapine, have side effects which include sleepiness and slowness, weight gain, sexual problems, increased risk of diabetes, and some risk of Parkinson's disease; long-term use can produce movements of the face and, rarely, of the arms and legs. Both are administered in the form of a pill. Increasingly, the use of depot antipsychotics, where medication is given as an injection every 2–4 weeks, has become more common: medication is hence released slowly over the course of time. Depots are usually administered at the local GP surgery, at a community mental health center, at a special outpatient clinic or by a nurse at home (Royal College of Psychiatrists Public Education Editorial Board 2012).

Antidepressants are also frequently administered; their main effect is to stimulate the amount of serotonin and/or noradrenaline in the brain (Lehtinen et al. 2007, p. 131). Due to their potentially adverse side effects, “new” antidepressants were introduced in the 1990s to curb these. Overall, the uptake of antidepressants has been on the rise in the last decades.

Cross-country variations are also apparent, with the USA leading the pack in terms of drug prescribing. More recent data show a continuation in this positive trend especially with regard to antidepressants in the USA (Olfson and Marcus 2009), New Zealand (Exeter et al. 2009), and Italy (Deambrosis et al. 2010) as well as antipsychotics in various countries (Verdoux et al. 2010). A study of the trends of antipsychotic prescribing in the USA for anxiety disorders among a representative sample of visits to office-based psychiatrists (Comer et al. 2011) found an increase from

10.6% (1996–1999) to 21.3% (2004–2007). Data from England showed annual increases from 1998 to 2010 of 6.8% on average: antidepressant prescriptions rose by 10% per year, while antipsychotics grew by 5.1%.

Costs of antipsychotics overtook those of antidepressants as the most costly psychiatric drug, with costs rising by 22% (Ilyas and Moncrieff 2012). Similarly, data from the USA in recent years show that antipsychotics, antidepressants, and drugs for attention deficit hyperactivity disorder have been consistently ranked as the most expensive prescription drugs (IMS Health 2010).

The first point of contact for mental health care in many countries is usually primary health care, and a majority of countries allow primary health care (PHC) doctors to prescribe and/or continue prescribing medicines for mental and behavioral disorders either without restrictions (56%) or with some legal restrictions (40%), such as allowing prescriptions only in certain categories of medicines or only in emergency settings. In other cases, psychiatrists or neurologists would take responsibility for prescribing for patients with more severe or treatment-resistant symptoms. Only 3% of respondent countries in a WHO survey did not allow any form of prescription by PHC doctors (World Health Organization 2011).

Treatment, care, and support for people with mental health problems are managed by primary, secondary (and tertiary) health-care settings, with a lot of treatment and care delivered in the community by non-medics. The most comprehensive form of mental health care, which comprises a balance between hospital and community-based services, has only been achieved in a few high-income countries (Saxena et al. 2007). Only half the countries in Africa, the eastern Mediterranean, and southeast Asia provide community-based care (World Health Organization 2005). Within-country differences also exist in terms of the availability of community-based care: this type of care is restricted to only a few areas in China, India, Paraguay, and Zambia. In general, about 52% of low-income countries and about 97% of high-income countries provide

community-based care (Saxena et al. 2007; World Health Organization 2005).

Hospital inpatient beds were the mainstay of mental health provision in many high-income countries for many decades and remain crucially important, but in many countries, the specialist (institutional) asylums are being or have been closed.

The global median number of facilities per 100,000 population is 0.61 outpatient facilities, 0.05 day treatment facilities, 0.01 community residential facilities, and 0.04 mental hospitals. In terms of psychiatric beds in general hospitals, the global median is 1.4 beds per 100,000 population. Higher income countries typically have more facilities and higher admission/utilization rates.

*Deinstitutionalization* is the process of shifting the care and support for patients with mental illness from custodial asylums to community-based settings and saw its real beginnings in the USA and then in England in the 1970s (Shorter 1997). This period also saw a shift in treatment, in terms of becoming demedicalized, as non-physician specialists begin to assume a role (Shorter 2007, pp. 21, 22).

In England, generally, studies have demonstrated that deinstitutionalization has had positive outcomes for service users (see the TAPS studies, e.g.). However, systematic data on the preferences and situations of people with mental health problems is gravely missing, with no existing European overview (Anderson et al. 2007). Data from the UK show that although the majority of people with mental health problems live in mainstream housing (Boardman 2010; Social Exclusion Unit 2004), many live in residential care homes (Health and Social Care Information Centre 2013) or in supported housing services or in independent flats where they receive “floating support” (Centre for Housing Research 2013), which is support for a set number of hours a week within a person’s home.

Data from Priebe et al. (2005) show that in fact, in most of the nine selected European countries (Austria, Denmark, England, Germany, Republic of Ireland, Italy, the Netherlands, Spain, and Switzerland), the number of psychiatric hospital beds

has been generally on the decline between 2002 and 2006. On the other hand, forensic bed spaces have been on the increase (except in Ireland, Italy, and Switzerland), as well as places in supported and supportive housing (except in Ireland and Switzerland) and in prisons. More specifically, in Iceland, Italy, and Sweden, psychiatric hospitals no longer exist and care is provided in beds in general hospitals or in community-based facilities (Medeiros et al. 2008).

However, community-based residential services are not available in all countries. Turkey and most cantons in Switzerland do not possess such facilities. Deinstitutionalization is advancing at different paces in different countries, mainly due to national traditions and the sociocultural context, the availability of resources, as well as financial incentives (Fakhoury and Priebe 2002).

Within Europe, the rates of the closure of asylums have been uneven between countries, and sometimes gaps have been reported between the closure of institutions and the provision of alternative services (Medeiros et al. 2008).

Research conducted in 2000 has shown that, for example, in Asia, and specifically in Japan (Kuno and Asukai 2000) and Hong Kong (Yip 2000), deinstitutionalization has yet to occur. In Japan, Kuno and Asukai (2000) comment that deinstitutionalization is unlikely to happen in the near future since people with mental health problems are not valued as members of society. More recently in Japan, the Sasagawa Project (Mizuno et al. 2005) aimed to make the transition of people with mental health problems from hospital to community living; this project claimed to be the first of its kind in the country. The study on the closure of Sasagawa hospital and the subsequent relocation of patients into Sasagawa “village” reported positive outcomes; however, there is much to say about the segregation of people with mental health problems.

Deinstitutionalization is currently under way in several South American countries (Larrobla and Botega 2000). In Australia, Moxham and Pegg (2000) commented that the shift to community care was not met with systematic and adequate planning and the delivery of appropriate housing services or placements.

Recently, a new project – EMERALD – was launched to improve mental health outcomes in health systems performance and identify its potential barriers, specifically in low- and mid-income countries (EMERALD 2014); results of this program have yet to be disseminated.

---

## Financing Mental Health Services: How Is Care Financed?

Though the cost of poor mental health has been estimated to be between 3% and 4% of GDP even in many European countries, no countries dedicate a proportionate level of resources to treating mental health disorders (Gabriel and Liimatainen 2000). Just over two thirds of countries across the world have a budget that is specifically dedicated to mental health, and many countries spend less than 1% of their total health budget on mental health-care services (Thornicroft and Maingay 2002). According to the 2005 WHO Mental Health Atlas, South East Asia had the highest proportion of countries with a specified budget for mental health care (90%); the Western Pacific had the lowest proportion of countries (59%). European countries often allocate funds specifically for mental health, despite not necessarily always having a specific line item within their national budgets (World Health Organization 2005a).

Generally, mental health budget information is scarce in low-income countries (Raja et al. 2010). A study by Raja et al. (2010) found however that national ring-fenced budgets for mental health as a percentage of national health spending for 2007–2008 were less than 4% in Sri Lanka, Ghana, and Kerala (India) and less than 7% in Uganda. Even in countries that dedicate substantial resources to mental health, coverage for mental health-care services may be more limited than other health-care services (Knapp et al. 2007).

Worldwide, government funds such as those generated by taxes are the most common source of mental health financing (World Health Organization 2003b). In countries where the government pays for the bulk of mental health care, care

is often financed in a similar way to the mechanism of funding general health care in that particular country. Out-of-pocket payments are also an important source of funding for mental health care in some countries, particularly outside of Europe. Even so, nearly half of western European countries levy user charges for specialist mental health-care services, even within their publicly funded system (Knapp et al. 2006). Generally, voluntary health insurance does not play a major role in funding mental health care. However, in some countries like the UK and Germany, there has been some expansion of mental health-care coverage within voluntary health insurance (Knapp 2007). In the USA there have also been recent efforts to ensure that private health insurers cover mental health conditions no differently than they cover physical conditions.

Naturally, each country allocates different levels of funding to the treatment of mental health. Historically, spending has been directed toward psychiatric hospitals; for example, three quarters of spending in Sri Lanka, Ghana, Kerala (India), and Uganda were on psychiatric hospital care. Recently though, there have been shifts in many countries toward allocating funds to community-based services as opposed to psychiatric hospitals. As a result of this move into the realm of social and community care, in some cases there has also been a trend to shift mental health-care funding away from health budgets and onto social protection budgets. This intersectoral approach to financing mental health care is not exclusive to high-income countries; for example, according to the WHO, the Burundi Ministry of Finance requested a social sector loan from the World Bank for work on early childhood development, which had an explicit mental health component (World Health Organization 2003b).

---

## Key Policy Dimensions/Recent Policies and Trends

Several key policy dimensions have dominated the global conversation on mental health.

## Personalization and Empowerment

Service user involvement is becoming increasingly commonplace, through patient-centered care, shared decision-making, patient participation, as well as the recovery model (Storm and Edwards 2013).

Personalized care and services are said to empower individuals and improve their quality of life. A novel method by which personalization is translated is through “direct payments” or “personal budgets” and has been introduced in a few countries, namely, England, Scotland, and the Netherlands (Knapp and McDaid 2007, p. 93), as well as the USA. In England more specifically, direct payments are “cash payments made to individuals who have been assessed as needing services, in lieu of social service provisions” (Department of Health 2008). They are aimed at giving recipients greater control over their own lives, enabling them to purchase services other than those provided by a local council, including novel solutions in terms of services and activities; the money a person receives is still decided on following an assessment of need. However, the uptake of direct payments among mental health service users has been slow, and it has been reported that they may have great difficulty of access possibly due to a lack of awareness, or as reported by staff, difficulties in managing payments (Davey et al. 2007). Still, despite low uptake rates, there was great diversity in their use, ranging from support with regard to personal care and transport, to everyday activities (Spandler and Vick 2004).

Individual budgets (subsequently called personal budgets) were later introduced in the UK, promising greater personalized purchasing and freedom in the selection of the chosen type of care and support (Department of Health 2006), and were to be delivered as a single transparent sum allocated to a person in their name and held on their behalf (like a bank account), allowing the individual to either then choose to take the funds out in cash (as a direct payment) or as a mixture of cash and services up to the value of their individual budget.

A Cochrane review of advance treatment directives (Campbell and Kisely 2009), which is a document which specifies a person’s preference for treatment, should they lose the capacity to make such decisions in the future, found two RCTs which included 321 people with severe mental illness. Although concluding that too little data was available to make robust conclusions, authors felt that more intensive forms of advance treatment directives seemed promising.

Recently, an RCT conducted in the UK (Thornicroft et al. 2013) aimed to test whether JCP (joint crisis planning) was associated with better outcomes among mental health service users, compared to the control group, who was receiving treatment as usual. No significant differences were found with regard to the primary outcome measure, namely, compulsory admissions; however, a modest improvement was found with regard to the therapeutic relationship. Qualitative analyses revealed that some patients in the intervention group reported positive experience; however, there was concern among others about how clinical services struggled to put JCP into practice.

In the Netherlands, an RCT comparing the quality aspects of crisis plans drawn up with the help of patient advocates compared to those by clinicians (Ruchlewska et al. 2012). The quality aspects checklist was devised specifically for this study and comprised four domains: (1) relapse indicators/daily functioning, (2) advance statements on what to do during a crisis, (3) medical information, and (4) information on contacts. The study concluded that, in terms of completeness and specificity, crisis plans drawn up by patients advocates were of better quality than those completed by clinicians.

## Carer and Family Impact

The impact of mental health problems can be felt through several mediums, namely, mortality, suicide, and crime and also by family. Schizophrenia, for example, can have enormous personal consequences for people with the illness and their families, as well as tremendous economic



consequences for them, as well as for governments and society as a whole. For example, in England, the total cost of schizophrenia has been estimated at almost £12 billion, and this includes a cost to the public sector of more than £7 billion (Andrew et al. 2012). Many relatives or other unpaid carers of people with schizophrenia may give up employment (4.8% of carers) or take time off work (15.5% took a mean 12.5 days off) in order to provide care and support. In economic terms, this translates into a loss of £517 (in 2011/2012 prices) per individual with schizophrenia living in a household (Mangalore and Knapp 2007).

The WHO has emphasized that more support is required for unpaid (sometimes called informal) carers, as usually their expenses as well as their opportunity costs (e.g., from lost employment) are not covered by the State or by insurance (World Health Organization 2003a). In addition to the emotional strain of caring, relatives can also be exposed to the stigma and discrimination associated with mental ill health. This in turn often translates into social isolation and exclusion from their communities, friends, and relatives.

### **Prevention, Promotion, Public Mental Health (e.g., Campaigning)**

Given the huge psychological, economic, and societal burdens, much emphasis has been placed on the prevention and promotion of mental health (World Health Organization 2004). (Much of the discussion on prevention can be found in Jané-Llopis and Anderson (2007).) In addition to targeted interventions, the WHO distinguishes macro-strategies that may reduce risk and improve quality of life. These include improving nutrition (especially in impoverished countries); improving housing and its quality; improving access to education; reducing economic insecurity; strengthening community networks through, for example, the Communities That Care program, already in force in the USA, the Netherlands, the UK, and Australia (Hawkins et al. 2002); and reducing the harm from addictive substances, through interventions such as taxes

and bans on underage drinking. For example, a comprehensive anti-smoking campaign can reduce smoking by up to 6% (Saffer 2000).

More specifically, Jané-Llopis and Anderson (2007, pp. 191–192) carefully lay out an integrated policy framework for the promotion of mental health and the prevention of mental disorders. These are subdivided by age categories: childhood and adolescence, adulthood and older groups, as well as by type of approach, whether public or mental health policy. Starting with fetal development, it is important to raise awareness among expectant mothers of the risk of substance use during pregnancy, for example, smoking while pregnant doubles the risk of lower birth weight. Educational programs in some countries to help pregnant women cease smoking have had immediate and long-term mental health gains on infants (Institute of Medicine 2001). Other interventions during childhood include parenting interventions. These target basic reading skills or other parenting skills and are said to improve literacy as well as emotional and language growth (Jané-Llopis and Anderson 2007, p. 193). Indeed, poor school performance increases the risk of social and mental health problems. School prevention programs involve general cognitive, problem-solving, and social skill-building, resulting in 50% reductions in depressive symptoms (Greenberg et al. 2001). However, and unfortunately, most low-income countries lack appropriate child and adolescent mental health services (Patel et al. 2008).

Funk et al. (2012) also focused on similar aspects to those of the WHO (2010) with regard to mental health interventions to improve development, by employing targeted poverty-alleviation programs in order to break the cycle between mental illness and poverty.

Funk et al. (2012) discuss many interventions, including pharmacological, psychosocial, and care-management strategies for schizophrenia, depression, alcohol misuse, epilepsy, and suicide prevention that have been associated with positive outcomes across the world, regardless of wealth. Suicide prevention should be highlighted through comprehensive public health programs and should at least comprise the following

interventions in low- and middle-income countries (LMICs): reducing the access to means for suicide, responsible and deglamorized media reporting, and early identification and treatment of people with mental and substance use disorders.

An important point to consider in working-age adults is employment and associated stress factors that may lead to anxiety, depression, or stress-related problems. Interventions to improve mental health in the workplace have centered on task and technical interventions (e.g., lowering workload or ergonomic improvements) and clarifying job role expectations as well as improving social environment (e.g., conflict resolution) (Price and Kompier 2006). There is now evidence that many of these prevention and promotion initiatives can be not only effective but also cost-effective. Andrew et al. (2012) assessed the various interventions in schizophrenia in terms of effectiveness and cost-effectiveness. One intervention, where authors found strong evidence for cost-effectiveness, was individual placement and support, which aims to help people with schizophrenia find competitive employment.

## Aging and Dementia

With the world population aging rapidly, and people living longer, the prevalence rate of age-related disorders is increasing. One such disorder is dementia, which often has an overwhelming effect on the individual with the illness, their family, and society more generally, prompting the WHO to promote it as a major public health priority (World Health Organization 2012). Dementia is a chronic and progressive syndrome, caused by a variety of brain illnesses and affecting memory, thinking, behavior, and ability to perform everyday activities. The latest figures from the WHO (2012) estimated the total number of people with dementia worldwide in 2010 to be 35.6 million, and this number is projected to nearly double every 20 years, to 65.7 million in 2030 and 115.4 million in 2050. The worldwide annual incidence rate of dementia is nearly 7.7 million, implying one new case every 4 s.

Total estimated worldwide costs of dementia were US\$ 604 billion in 2010. In high-income countries, informal care (45%) and formal social care (40%) make up the majority of costs, in comparison to direct medical costs (15%) which are much lower. In low-income and lower-middle-income countries, direct social care costs are small, and the costs of unpaid care provided by the family dominate (World Health Organization 2012). Given the expected growth over the coming decades in the number of people with dementia, the costs of supporting and treating them can also be expected to increase rapidly too. For example, a study comparing future dementia costs in Italy, Spain, the UK, and Germany suggested that the proportion of GDP spent on long-term care would more than double between 2000 and 2050 (Comas-Herrera et al. 2006).

These projected future trends have prompted much discussion and also some real action across many countries. One of the first countries to develop such a plan was Canada in 1999, and their “Alzheimer Strategy – Preparing for our future” runs till 2014. A good example of an integrated action plan for dementia comes from France, which was one of the first European countries to launch such a program (in 2008). Based on 44 measures to combat dementia and related disorders (République Française 2013), the key aims are to improve diagnosis, to provide better treatment and support through establishing “coordinators” throughout the country and through encouraging treatment at home by skilled support staff, and to provide more effective help through developing and diversifying respite structure and through the use of technology (such as a telephone line or a website). A final aim was to create a foundation for scientific cooperation to stimulate and coordinate research through memory clinics and diagnostics centers, with a lesser reliance on antipsychotic drugs (République Française 2008). It also aimed to change the way dementia is viewed, by raising awareness at the national and international level. The plan pledged 1.6 billion Euros over this period.

More recently (December 2013), the Health Ministers from the G8 countries met in London for a Dementia Summit, following which they jointly issued a declaration and communique, spelling out clearly the challenges so often experienced by family and other carers of people with dementia and the need for action. Further joint action is planned to tackle what has become a major global mental health challenge.

### Employment

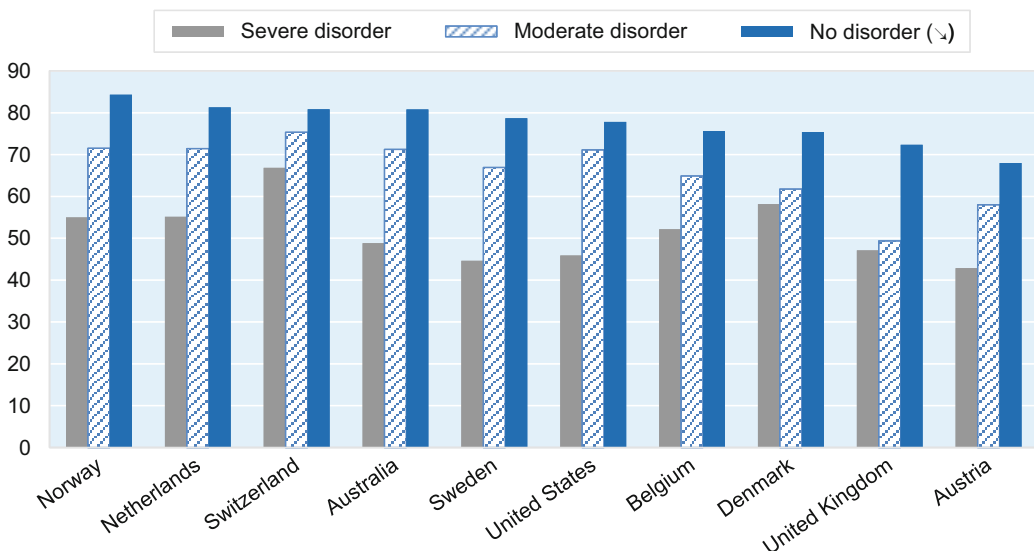
Previous studies have shown the enhancing effect employment can have within the mental health population. However, poor-quality jobs can be detrimental to mental health. This is problematic, due to the fact that people with mental health problems often find themselves in low-skilled jobs, which can add strain to their emotional well-being, as well as not being suitable to their needs and preferences (OECD 2011).

However, despite these gains in outcomes, employment rates among people with mental health problems vary by diagnosis severity and

are relatively low compared to the general population.

People with a history of mental health problems face problems in the open employment market, including stigma, a reluctance from employers to give them a job (Manning and White 1995), with some even alluding to their perceived risk of violence (Roberts et al. 2004). A recent study using Eurobarometer surveys of 2006 and 2010 has demonstrated that the economic crisis has widened the gap even more in terms of unemployment rates between people with and without mental health problems; those who were particularly affected were men and people with lower educational attainment (Evans-Lacko et al. 2013b). Additionally, people living within countries with higher levels of stigmatizing attitudes toward people with mental health problems were particularly more vulnerable to unemployment in 2010.

Nevertheless, it is generally agreed that motivation to work has a significant influence on whether people with severe mental illness gain competitive employment (Catty et al. 2008); anti-psychotic medication also plays a role here as



Employment/population ratio (employed people as a proportion of the working-age population), by severity of mental disorder, ten OECD countries, latest available year (late 2000s) (OECD 2011)

well. Another reported and important disincentive to work is a risk of losing entitlements. Innovative solutions have come from OECD countries. For example, in Luxembourg, people who were on benefits and find a job get a permanent payment to supplement any loss in earnings (OECD 2010).

Previous studies have found that people with mental health problems are generally interested in pursuing employment opportunities (Grove 1999; Secker et al. 2001) but felt that their mental health was an important barrier to doing so (Bond et al. 1997; Sainsbury Centre for Mental Health 2004; Secker et al. 2001).

Most people with mental health problems can achieve competitive employment (Bond et al. 1997). Studies of supported employment schemes in the USA have shown that employment may lead to improvements in outcomes, in terms of mental health treatment (Cuyun Carter et al. 2011), through increasing self-esteem and improving quality of life (Van Dongen 1996), as well as alleviating psychiatric symptoms and reducing dependency (Cook and Razzano 2000). A Cochrane review of vocational rehabilitation found that supported employment schemes to be more effective than various type of prevocational training in obtaining and maintaining employment (Crowther et al. 2001). An evidence-based refinement of the supported employment approach, individual placement and support, has had positive outcomes in the UK and the USA (Leff and Warner 2006, p. 134). A report commissioned by the cross government Health Work and Wellbeing Programme on mental health and the workplace (Lelliott et al. 2008) concluded that IPS has the strongest evidence base of interventions aimed at helping people with severe mental illness back into employment. However, authors added that IPS can only be beneficial among people who believed they were ready for paid employment. Other limitations included the fact that most people ended up in mostly part-time, entry-level jobs; the long-term outcomes and economic benefits are unknown.

### Individual Placement and Support (IPS)

IPS is based on an integrated approach to seeking and maintaining employment. It starts with the principle that no mental health service users are excluded due to a poor previous work history, lack of “work-readiness,” frequent hospital admissions, or apparent symptoms. Vocational programs should be integrated as part of the mental health agency or team. The achievement of competitive employment, one that also takes into consideration preferences, work and education history (if applicable), strengths, and weaknesses, is key. Rapid job search and placement is preferred to prior assessment, skills training, and vocational counseling. The most valuable assessment is made *after* obtaining employment, as well as providing support and services for a sustained period of time. Job coaches should also be used to help service users understand the complex rules governing disability benefits and assisting them in making the best employment decisions (Bond 1998).

The OECD (2011) has also recently highlighted the promotion of good mental health in the workplace, as well as its linkages to well-being and productivity. The OECD reports that most people with mental ill health are in work yet that productivity losses are extensive.

These may include short- and long-term absenteeism, early retirement, reduced employment opportunities, presenteeism, days out-of-role, and reduced lifetime productivity due to premature mortality (McDaid et al. 2008). Data from the OECD shows sharp increases in absenteeism and presenteeism with poorer mental health. So far the OECD has released several country reports, with others due in 2014. For example, in Sweden, Denmark, Norway, and Belgium, a common recommendation was to tackle the issue of possible future unemployment from an early age, during the school years, and to minimize dropouts, as well as focus on cases that were deemed at risk.

Another commonality was to tackle the issue of mental ill health within the workplace in a systematic and rapid fashion, before potential absenteeism or job loss.

Other techniques may involve cognitive behavioral therapy within the workplace, and several US studies demonstrated positive outcomes, in terms of better mental health outcomes and higher rates of job retention (Wang et al. 2006, 2007). In France, the Electricite de France and Gaz de France are major companies and employers that implemented a program “APRAND” (Action de Prévention des Rechutes des troubles Anxieux et Dépressifs), which focused mostly on prevention, by encouraging company health physicians, primary care doctors, and social workers to identify anxiety and depression problems early among employees. The implementation of preventative activities among people on long-term sick leave who had been identified as having anxiety problems had better outcomes in terms of recovery or remission compared to people who received regular treatment (Godard et al. 2006).

A recent report released by the OECD (2014) stated that, in an international comparison, the UK is among the most advanced countries in terms of awareness of the costs of mental illness for society and the benefits that employment may bring to people with mental health problems. The OECD recognized that stricter eligibility criteria for benefits as well as large-scale reassessments were steps in the right direction. The report stated some key recommendations as well which focused on prevention (and early intervention to avoid sickness benefit becoming a disability benefit) and the expansion of psychological therapies for people with CMDs, more awareness of employability within the benefits system, and building on the better integration of health and employment services. Other recommendations include to invest in labor markets more generally, to be able to provide better support for people with mental health problems, and to focus on outcome payments for employers in to promote better employment outcomes.

## New Advancements in Treatments and Technologies

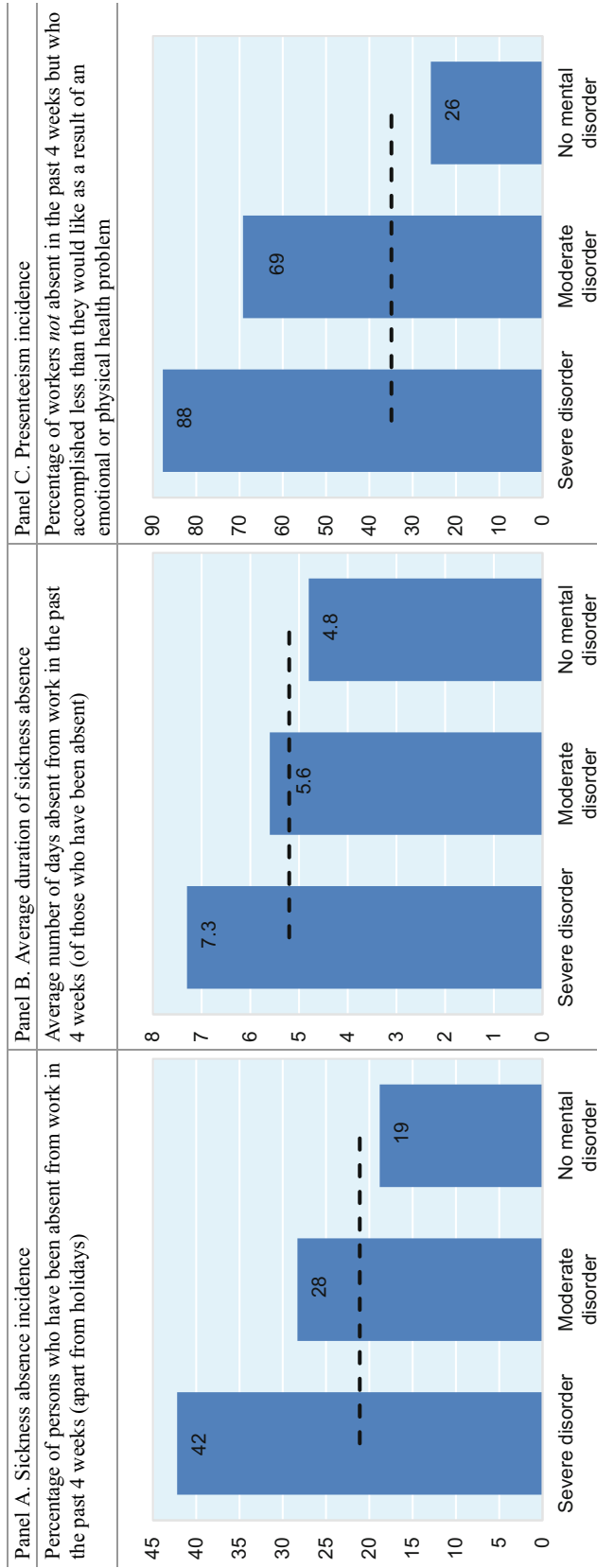
Innovative and cost-effective treatments and technologies with regard to mental health have started being developed, as traditional services struggle to cope with the growing demand and variety of needs of its users, at least in the UK (Limb 2012). Video games, online and social network tools, as well as mobile phone apps have been quick to respond. The internet does not only provide greater control and power over services (Gournay 2000), but can also be a powerful therapeutic tool (Foroushani et al. 2011), and a way for service users to voice their opinions. For instance, an online system for people with eating disorders in Germany (Moessner and Bauer 2012), which makes use of e-mailing and forums, generated high user satisfaction and, more importantly, increased the probability of users seeking professional help. In the UK, a Google search for “suicide” generates links for the Samaritans and other support groups first (Wicks 2012). Otherwise, PlayMancer is an EU initiative to develop a video game prototype to treat specific mental health problems, such as impulse control disorders. This game uses biofeedback to help people learn relaxation skills and develop better self-control and emotion regulation strategies. Initial results are positive, and patients have started to show new coping styles and more self-control (Fernandez-Aranda et al. 2012). In England, the NHS has developed a mobile phone app “My Journey” as part of their early interventions in psychosis, to monitor young persons’ mood, keep track of medication, set goals, and be a source of advice if needed (<http://www.sabp.nhs.uk/eiip/app>).

---

## Discussion

Mental health problems present tremendous personal and financial burdens to the individual, to their carers, and to society as a whole. However, resources allocated by different countries to the care and support of these individuals vary. The WHO has put mental health promotion at the heart of its policy agenda; huge inequalities still remain.

Incidence of absenteeism and presenteeism (in percentage) and average absence duration (in days), by mental health status, average over 21 European OECD countries in 2010



Note: Averages are represented by dashed lines  
 Source: OECD calculations based on Eurobarometer (2010)

Stigma seemed to be pervasive across the board: people in low-income and maybe middle-income countries still experience basic human rights violations, while in high-income countries, policies are more focused on social inclusion and integration. Indeed, the promotion of personalized care and service user empowerment, although producing mixed results, are steps in the right direction.

Deinstitutionalization has yet to occur in many countries, although the process has been under way in some parts of Europe and the USA since the 1970s. Regardless of the financial resources, and funding arrangements in place, perhaps stigma and discrimination do play a role in the continued existence of asylums and institutions, alluding to the so-called NIMBY phenomenon (Thornicroft 2007).

Mixed evidence exists with regard to anti-stigma campaigns, and it could be that a more integrated approach to mental health promotion should be adopted somehow, with a focus on prevention as well.

E-technologies may prove to be innovative and perhaps more importantly, cost-effective solutions but should still be regarded as complementary therapies.

## References

- Alem A, Jacobsson L, Araya M, Kebede D, Kullgren G. How are mental disorders seen and where is help sought in a rural Ethiopian community? A key informant study in Butajira, Ethiopia. *Acta Psychiatr Scand*. 1999;397:40–7.
- Al-Krenawi A, Graham JR. Culturally sensitive social work practice with Arab clients in mental health settings. *Health Soc Work*. 2000;25(1):9–22.
- Anderson R, Wynne R, McDaid D. Housing and employment. In: Knapp M, McDaid D, Mossialos E, Thornicroft G, editors. *Mental health policy and practice across Europe*. Maidenhead: Open University Press/McGraw-Hill Education; 2007.
- Andrew A, Knapp M, McCrone P, Parsonage M, Trachtenberg M. *Effective interventions in schizophrenia: the economic case: a report prepared for the Schizophrenia Commission*. London: Rethink Mental Illness; 2012.
- Angermeyer MC, Matschinger H. Reporting of isolated violent attacks by people with schizophrenia in the media changes attitudes towards people with mental illness. *Soc Sci Med*. 1996;43(12):1721–8.
- Angermeyer MC, Matschinger H. Causal beliefs and attitudes to people with schizophrenia. Trend analysis based on data from two population surveys in Germany. [Research Support, Non-U.S. Gov't]. *Br J Psychiatry*. 2005;186:331–4.
- Ayuso-Mateos JL. Global burden of bipolar disorder in the year 2000. Geneva: WHO; 2000a.
- Ayuso-Mateos JL. Global burden of obsessive-compulsive disorder in the year 2000. Geneva: WHO; 2000b.
- Ayuso-Mateos JL. Global burden of panic disorder in the year 2000: version 1 estimates. Geneva: WHO; 2000c.
- Ayuso-Mateos JL. Global burden of schizophrenia in the year 2000: version 1 estimates. Geneva: WHO; 2000d.
- Baldwin ML, Marcus SC. Stigma, discrimination, and employment outcomes among persons with mental health disabilities. In: Schultz IZ, Rogers ES, editors. *Work accommodation and retention in mental health*. New York: Springer; 2011.
- Bebbington P. Population surveys of psychiatric disorder and the need for treatment. *Soc Psychiatry Psychiatr Epidemiol*. 1990;25(1):33–40.
- Bennett D. The value of work in psychiatric rehabilitation. *Soc Psychiatry*. 1970;5(4):224–30.
- Boardman J. How are people with mental health problems excluded? In: Boardman J, Currie A, Killaspy H, Mezey G, editors. *Social inclusion and mental health*. London: Royal College of Psychiatrists; 2010.
- Bond GR. Principles of individual placement and support. *Psychiatr Rehabil J*. 1998;27:345–59.
- Bond GR, Drake RE, Mueser KT, Becker DR. An update on supported employment for people with severe mental illness. *Psychiatr Serv*. 1997;48(3):335–46.
- Brohan E, Elgie R, Sartorius N, Thornicroft G, GAMIAN-Europe Study Group. Self-stigma, empowerment and perceived discrimination among people with schizophrenia in 14 European countries: the GAMIAN-Europe study. *Schizophr Res*. 2010;122(1–3):232–8.
- Catty J, Lissouba P, White S, Becker T, Drake RE, Fioritti A, et al. Predictors of employment for people with severe mental illness: results of an international six-centre randomised controlled trial. *Br J Psychiatry*. 2008;192:224–31.
- Centre for Housing Research. *Supporting people*. 2013. From <https://supportingpeople.st-andrews.ac.uk/index.cfm>
- Chapman DP, Perry GS, Strine TW. The vital link between chronic disease and depressive disorders. *Prev Chronic Dis*. 2005;3(2):1–3.
- Cimpean D, Drake RE. Treating co-morbid chronic medical conditions and anxiety/depression. [Review]. *Epidemiol Psychiatr Sci*. 2011;20(2):141–50.
- Comas-Herrera A, Wittenberg R, Costa-Font J, Gori C, Di Maio A, Patxot C, et al. Future long-term care expenditure in Germany, Spain, Italy and the United Kingdom. *Ageing Soc*. 2006;26:285–302.
- Cook J, Razzano L. Vocational rehabilitation for persons with schizophrenia: recent research and implications for practice. *Schizophr Bull*. 2000;26(1):87–103.

- Crosby C, Barry M, Carter MF, Lowe CF. Psychiatric rehabilitation and community care: resettlement from a North Wales Hospital. *Health Soc Care.* 1993;1:355–63.
- Crowther RE, Marshall M, Bond GR, Huxley P. Helping people with severe mental illness to obtain work: systematic review. *Br Med J.* 2001;322:204–8.
- Cuyun Carter GB, Milton DR, Ascher-Svanum H, Faries DE. Sustained favorable long-term outcome in the treatment of schizophrenia: a 3-year prospective observational study. *BMC Psychiatry.* 2011;11:143.
- Davey V, Fernández J-L, Knapp M, Vick N, Jolly D, Swift P, et al. Direct payments: a national survey of direct payments policy and practice. London: London School of Economics; 2007.
- Deambrosis P, Chinellato A, Terrazzani G, Pullia G, Giusti P, Skaper SD, et al. Antidepressant drug prescribing patterns to outpatients of an Italian local health authority during the years 1998 to 2008. [Comparative Study Letter Research Support, Non-U.S. Gov't]. *J Clin Psychopharmacol.* 2010;30(2):212–5.
- Department for Education and Employment. Towards full employment in a modern society. 2001.
- Department of Health. Our health, our care, our say: a new direction for community services. 2006. Retrieved from [http://webarchive.nationalarchives.gov.uk/+www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH\\_4127453](http://webarchive.nationalarchives.gov.uk/+www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4127453)
- Department of Health. Direct payments. 2008. From [http://www.dh.gov.uk/en/SocialCare/Socialcarereform/Personalisation/Directpayments/DH\\_080273](http://www.dh.gov.uk/en/SocialCare/Socialcarereform/Personalisation/Directpayments/DH_080273)
- Department of Health. No health without mental health: a cross-government mental health outcomes strategy for people of all ages. 2011.
- Drew N, Funk M, Pathare S, Swartz L. Mental health and human rights. In: Herrman H, Saxena S, Moodie R, editors. *Promoting mental health: concepts, emerging evidence, practice.* Geneva: World Health Organisation; 2005.
- Ellis PM, Smith DA, beyond blue: the national depression initiative. Treating depression: the beyond blue guidelines for treating depression in primary care. “Not so much what you do but that you keep doing it”. [Guideline Meta-Analysis Practice Guideline Research Support, Non-U.S. Gov't]. *Med J Aust.* 2002;176(Suppl):S77–83.
- EMERALD. EMERALD. 2014.
- Eurobarometer. Mental health eurobarometer. 2010.
- Evans-Lacko S, Henderson C, Thornicroft G. Public knowledge, attitudes and behaviour regarding people with mental illness in England 2009–2012. *Br J Psychiatry.* 2013a;202:51–7.
- Evans-Lacko S, Knapp M, McCrone P, Thornicroft G, Mojtabai R. The mental health consequences of the recession: economic hardship and employment of people with mental health problems in 27 European countries. [Research Support, Non-U.S. Gov't]. *PLoS One.* 2013b;8(7):e69792.
- Evans-Lacko S, Malcolm E, West K, Rose D, London J, Rusch N, et al. Influence of Time to Change's social marketing interventions on stigma in England 2009–2011. [Evaluation Studies Research Support, Non-U.S. Gov't]. *Br J Psychiatry Suppl.* 2013c;55:s77–88.
- Fakhoury W, Priebe S. The process of deinstitutionalisation: an international overview. *Curr Opin Psychiatry.* 2002;15:187–92.
- Fakhoury W, Priebe S. Deinstitutionalization and reinstitutionalization: major changes in the provision of mental healthcare. *Psychiatry.* 2007;6(8):313–6.
- Fernandez-Aranda F, Jimenez-Murcia S, Santamaria JJ, Gunnard K, Soto A, Kalapanidas E, et al. Video games as a complementary therapy tool in mental disorders: PlayMancer, a European multicentre study. *J Ment Health.* 2012;21(4):364–74.
- Foroushani PS, Schneider J, Assareh N. Meta-review of the effectiveness of computerised CBT in treating depression. [Comparative Study Research Support, Non-U.S. Gov't Review]. *BMC Psychiatry.* 2011;11:131.
- Funk M, Saraceno B, Drew N. Global perspective on mental health policy and service development issues. In: Knapp M, McDaid D, Mossialos E, Thornicroft G, editors. *Mental health policy and practice across Europe: the future direction of mental health care.* Maidenhead: Open University Press; 2005.
- Gabriel P, Liimatainen M-R. Mental health in the workplace. Geneva: International Labour Organisation; 2000.
- Godard C, Chevalier A, Lecrubier Y, Lahon G. APRAND programme: an intervention to prevent relapses of anxiety and depressive disorders – first results of a medical health promotion intervention in a population of employees. *Eur Psychiatry.* 2006;21(7):451–9.
- Goffman E. *Stigma: notes on the management of spoiled identity.* Englewood Cliffs: Prentice-Hall; 1963.
- Gournay K. Commentaries and reflections on mental health nursing in the UK at the dawn of the new millennium: commentary 2. *J Ment Health.* 2000;9(6): 621–3.
- Greenberg MT, Domitrovich C, Bumbarger B. The prevention of mental disorders in school-aged children: current state of the field. *Prev Treat.* 2001;4.
- Grove B. Mental health and employment. Shaping a new agenda. *J Ment Health.* 1999;8:131–40.
- Hawkins JD, Catalano RF, Arthur MW. Promoting science-based prevention in communities. *Addict Behav.* 2002;27(6):951–76.
- Health and Social Care Information Centre. Community care statistics, social services activity – England, 2011–12, Final release. 2013. From <https://catalogue.ic.nhs.uk/publications/social-care/activity/comm-care-soci-serv-act-eng-11-12-fin/comm-care-stat-eng-2011-12-soci-serv-act-rep.pdf>
- Institute of Medicine. *Clearing the smoke.* Washington, DC: National Academy Press; 2001.
- Jané-Llopis E, Anderson P. A policy framework for the promotion of mental health and the prevention of



- mental disorders. In: Knapp M, McDaid D, Mossialos E, Thornicroft G, editors. *Mental health policy and practice across Europe*. Maidenhead: McGraw-Hill; 2007.
- Kessler RC, Aguilar-Gaxiola S, Alonso J, Chatterji S, Lee S, Ormel J, et al. The global burden of mental disorders: an update from the WHO World Mental Health (WMH) surveys. [Research Support, N.I.H., Extramural]. *Epidemiol Psychiatr Soc*. 2009;18(1): 23–33.
- Killaspay H. From the asylum to community care: learning from experience. *Br Med Bull*. 2006;79:245–58.
- Knapp M. *Mental health policy and practice across Europe: the future direction of mental health care*. Maidenhead: Open University Press; 2007.
- Knapp M, McDaid D. Financing and funding mental health care services. In: Knapp M, McDaid D, Mossialos E, Thornicroft G, editors. *Mental health policy and practice across Europe*. Maidenhead: McGraw-Hill/Open University Press; 2007.
- Knapp M, McDaid D, Amaddeo F. Financing arrangements for mental health in Western Europe. *J Ment Health*. 2006.
- Knapp M, McDaid D, Amaddeo F, Constantopoulos A, Oliveira MD, Salvador-Carulla L, Zechmeister I, the MHEEN Group. Financing mental health care in Europe. *J Ment Health*. 2007;16(2):167–80.
- Kohn R, Saxena S, Levav I, Saraceno B. The treatment gap in mental health care. [Research Support, Non-U.S. Gov't Review]. *Bull World Health Organ*. 2004;82(11):858–66.
- Kuno E, Asukai N. Efforts toward building a community-based mental health system in Japan. *Int J Law Psychiatry*. 2000;23(3–4):361–73.
- Larrobla C, Botega NJ. Psychiatric care policies and deinstitutionalization in South America. *Actas Esp Psiquiatr*. 2000;28(1):22–30.
- Lawrie SM. Newspaper coverage of psychiatric and physical illness. *Psychiatr Bull*. 2000;24:104–6.
- Leff J. The outcome for long-stay non-demented patients. In: Leff J, editor. *Care in the community: illusion or reality?* London: Wiley; 1997.
- Leff J, Warner R. *Social inclusion of people with mental illness*. Cambridge: Cambridge University Press; 2006.
- Leff J, Dayson D, Gooch C, Thornicroft G, Wills W. Quality of life of long stay patients discharged from two psychiatric institutions. *Psychiatr Serv*. 1996; 47:62–7.
- Lehtinen V, Katschnig H, Kovess-Masfety V, Goldberg D. *Mental health policy and practice across Europe*. Maidenhead: McGraw Hill; 2007.
- Lelliott P, Tulloch S, Boardman J, Harvey S, Henderson M, Knapp M. *Mental health and work*. London: Royal College of Psychiatrists; 2008.
- Limb M. Digital technologies offer new ways to tackle mental health problems. *Br Med J*. 2012;345: e5163.
- Mangalore R, Knapp M. Cost of schizophrenia in England. [Research Support, Non-U.S. Gov't]. *J Ment Health Policy Econ*. 2007;10(1):23–41.
- Manning C, White PD. Attitudes of employers to the mentally ill. *Psychiatr Bull*. 1995;19:541–3.
- Marrone J, Golowka E. If work makes people with mental illness sick, what do unemployment, poverty, and social isolation cause? *Psychiatr Rehabil J*. 2000; 23(2):187–93.
- McCourt CA. Life after hospital closure: users' views of living in residential 'resettlement' projects. A case study in consumer-led research. *Health Expect*. 2000;3:192–202.
- McDaid D, Knapp M, Medeiros H, the MHEEN Group. *Employment and mental health: assessing the economic impact and the case for intervention*. London: Personal Social Services Research Unit; 2008.
- Medeiros H, McDaid D, Knapp M, the MHEEN Group. *Shifting care from hospital to the community in Europe: economic challenges and opportunities*. London: Personal Social Services Research Unit; 2008.
- Mental Disability Advocacy Centre. *Guardianship and human rights in Bulgaria: analysis of law, policy and practice*. 2007a.
- Mental Disability Advocacy Centre. *Guardianship and human rights in Russia: analysis of law, policy and practice*. 2007b.
- Mindout for mental health. *Working minds: making mental health your business*. 2000.
- Mizuno M, Sakuma K, Ryu Y, Munakata S, Takebayashi T, Murakami M, et al. The Sasagawa project: a model for deinstitutionalisation in Japan. *Keio J Med*. 2005; 54(2):95–101.
- Moessner M, Bauer S. Online counselling for eating disorders: reaching an underserved population? *J Ment Health*. 2012;21(4):336–45.
- Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *Lancet*. 2007;370(9590):851–8.
- Moxham LJ, Pegg SA. Permanent and stable housing for individuals living with a mental illness in the community: a paradigm shift in attitude for mental health nurses. *Aust N Z J Ment Health Nurs*. 2000;9(2):82–8.
- National Centre for Social Research. *Adult psychiatric morbidity in England, 2007. Results of a household survey*. 2007. Retrieved from [http://www.ic.nhs.uk/webfiles/publications/mental%20health/other%20mental%20health%20publications/Adult%20psychiatric%20morbidity%2007/APMS%2007%20\(FINAL\)%20Standard.pdf](http://www.ic.nhs.uk/webfiles/publications/mental%20health/other%20mental%20health%20publications/Adult%20psychiatric%20morbidity%2007/APMS%2007%20(FINAL)%20Standard.pdf)
- Naylor C, Parsonage M, McDaid D, Knapp M, Fossey M, Galea A. *Long-term conditions and mental health. The cost of co-morbidities*. London: The King's Fund and Centre for Mental Health; 2012.
- OECD. *Sickness, disability and work: breaking the barriers*. 2010.
- OECD. *Sick on the job*. 2011. From <http://www.oecd.org/els/emp/sickonthejob2011.htm>
- OECD. *Sick on the job? Myths and realities about mental health and work*. OECD Publishing; 2012.

- Ohara T, Doi Y, Ninomiya T, Hirakawa Y, Hata J, Iwaki T, et al. Glucose tolerance status and risk of dementia in the community: the Hisayama study. [Comparative Study Research Support, Non-U.S. Gov't]. *Neurology*. 2011;77(12):1126–34.
- Olfson M, Marcus SC. National patterns in antidepressant medication treatment. [Comparative Study Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.]. *Arch Gen Psychiatry*. 2009;66(8):848–56.
- Oxford Dictionaries. Stigma. Oxford Dictionaries. 2010. From <http://oxforddictionaries.com/definition/stigma>
- Paykel ES, Hart D, Priest RG. Changes in public attitudes to depression during the Defeat Depression Campaign. [Research Support, Non-U.S. Gov't]. *Br J Psychiatry*. 1998;173:519–22.
- Price R, Kompier M. Work stress and unemployment: risks, mechanisms, and prevention. In: Hosman C, Jané-Llopis E, Saxena S, editors. *Prevention of mental disorders: effective strategies and policy options*. Oxford: Oxford University Press; 2006.
- Priebe S, Badescony A, Fioritti A, Hansson L, Kilian R, Torres-Gonzales F, et al. Reinstitutionalisation in mental health care: comparison of data on service provision from six European countries. *Br Med J*. 2005;330:123–6.
- Raja S, Wood SK, de Menil V, Mannarath SC. Mapping mental health finances in Ghana, Uganda, Sri Lanka, India and Lao PDR. *Int J Ment Heal Syst*. 2010;4:11.
- Read J, Baker S. Not just sticks and stones: a survey of the discrimination experienced by people with mental health problems. 1996.
- République Française. National plan for “Alzheimer and related diseases” 2008–2012. 2008. Available from [http://www.plan-alzheimer.gouv.fr/IMG/pdf/Plan\\_Alzheimer\\_2008-2012\\_uk.pdf](http://www.plan-alzheimer.gouv.fr/IMG/pdf/Plan_Alzheimer_2008-2012_uk.pdf)
- République Française. 44 measures in order to fight Alzheimer’s disease and related disorders. 2013. From <http://www.plan-alzheimer.gouv.fr/44-measures.html>
- Roberts S, Heaver C, Hill K, Rennison J, Stafford B, Howat N, et al. Disability in the workplace: employers and service providers’ response to the Disability Discrimination Act in 2003 and preparation for 2004 changes. 2004.
- Royal College of Psychiatrists Public Education Editorial Board. Antipsychotics. 2012. From <http://www.rcpsych.ac.uk/mentalhealthinfo/treatments/antipsychoticmedication.aspx>
- Saffer H. Tobacco advertising and promotion. In: Jha P, Chaloupka F, editors. *Tobacco control in developing countries*. Oxford: Oxford Medical Publications; 2000. p. 215–36.
- Sainsbury Centre for Mental Health. Briefing 27. Benefits and work for people with mental health problems: a briefing for mental health workers. 2004. Retrieved from [http://www.centreformentalhealth.org.uk/pdfs/briefing\\_27.pdf](http://www.centreformentalhealth.org.uk/pdfs/briefing_27.pdf)
- Sartorius N, Jablensky A, Korten A, Ernberg G, Anker M, Cooper JE, et al. Early manifestations and first-contact incidence of schizophrenia in different cultures. A preliminary report on the initial evaluation phase of the WHO Collaborative Study on determinants of outcome of severe mental disorders. *Psychol Med*. 1986;16(4):909–28.
- Savaya R. Attitudes towards family and marital counseling among Israeli Arab women. *J Soc Serv Res*. 1995;21(1):35–51.
- Secker J, Grove B, Seebom P. Challenging barriers to employment, training and education for mental health service users: the service user’s perspective. *J Ment Health*. 2001;10(4):395–404.
- Sederer LI, Silver L, McVeigh KH, Levy J. Integrating care for medical and mental illnesses. [Comment]. *Prev Chronic Dis*. 2006;3(2):A33.
- Shepherd G. *Institutional care and rehabilitation*. London: Longman; 1984.
- Shepherd G, Muijen M, Dean R, Cooney M. *Inside residential care*. London: The Sainsbury Centre for Mental Health; 1995.
- Social Exclusion Unit. *Mental health and social exclusion social exclusion unit report*. 2004. Retrieved from <http://www.socialinclusion.org.uk/publications/SEU.pdf>
- Spandler H, Vick N. *Direct payments, independent living and mental health*. London: Health and Social Care Advisory Service; 2004.
- Tansella M. Community psychiatry without mental hospitals – the Italian experience: a review. *J R Soc Med*. 1986;79:664–9.
- Thornicroft G. *Shunned. Discrimination against people with mental illness*. Oxford: Oxford University Press; 2007.
- Thornicroft G, Bebbington PE. Deinstitutionalisation – from hospital closure to service development. *Br J Psychiatry*. 1989;155:739–53.
- Thornicroft G, Maingay S. The global response to mental illness – an enormous health burden is increasingly being recognised. *Br Med J*. 2002;325(7365):608–9.
- Thornicroft G, Brohan E, Rose D, Sartorius N, Leese M. Global pattern of experienced and anticipated discrimination against people with schizophrenia: a cross-sectional survey. *Lancet*. 2009;373:408–15.
- Van Dongen CJ. Quality of life and self-esteem in working and nonworking persons with mental illness. *Community Ment Health J*. 1996;32(6):535–48.
- Velayudhan L, Poppe M, Archer N, Proitsi P, Brown RG, Lovestone S. Risk of developing dementia in people with diabetes and mild cognitive impairment. [Research Support, Non-U.S. Gov't]. *Br J Psychiatry*. 2010;196(1):36–40.
- Verdoux H, Tournier M, Begaud B. Antipsychotic prescribing trends: a review of pharmaco-epidemiological studies. [Review]. *Acta Psychiatr Scand*. 2010;121(1):4–10.
- Wahl OF. *Telling is risky business*. New Brunswick: Rutgers University Press; 1999.
- Wang PS, Patrick A, Avorn J, Azocar F, Ludman E, McCulloch J, et al. The costs and benefits of enhanced

- depression care to employers. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. *Arch Gen Psychiatry*. 2006;63(12):1345–53.
- Wang PS, Simon GE, Avorn J, Azocar F, Ludman EJ, McCulloch J, et al. Telephone screening, outreach, and care management for depressed workers and impact on clinical and work productivity outcomes – a randomized controlled trial. *JAMA*. 2007;298(12):1401–11.
- Warner R. Recovery from schizophrenia: psychiatry and political economy. Oxford: Oxford University Press; 1994.
- Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet*. 2013;382(9904):1575–86.
- Wicks P. E-mental health: a medium reaches maturity. [Comment Editorial]. *J Ment Health*. 2012;21(4):332–5.
- Wing L, Gould J, Gillberg C. Autism spectrum disorders in the DSM-V: better or worse than the DSM-IV? *Res Dev Disabil*. 2011;32(2):768–73.
- World Health Organisation. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and clinical guidelines. Geneva: WHO; 1992.
- World Health Organization. The world health report 2001: mental health: new understanding, new hope. Geneva: World Health Organization; 2001.
- World Health Organisation. Investing in mental health. Geneva: WHO; 2003a.
- World Health Organisation. Mental health financing. 2003b.
- World Health Organisation. Prevention of mental disorders. Effective interventions and policy options. Geneva: WHO; 2004.
- World Health Organisation. Mental health atlas 2005. Geneva: WHO; 2005a.
- World Health Organisation. WHO resource book on mental health, human rights and legislation. Stop exclusion, dare to care. Geneva; 2005b.
- World Health Organisation. Scaling up care for mental, neurological, and substance use disorders. Geneva: World Health Organisation; 2008.
- World Health Organisation. Eurobarometer 73.2. Mental health. Brussels: WHO; 2010.
- World Health Organisation. Mental health atlas 2011. Geneva: WHO; 2011.
- World Health Organisation. Dementia a public health priority. Geneva: WHO; 2012.
- World Health Organisation. Mental health action plan 2013–2020. Geneva: WHO; 2013.
- Xu W, Qiu C, Gatz M, Pedersen NL, Johansson B, Fratiglioni L. Mid- and late-life diabetes in relation to the risk of dementia: a population-based twin study. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Twin Study]. *Diabetes*. 2009; 58(1):71–7.
- Yip KS. Have psychiatric services in Hong Kong been impacted by the deinstitutionalization and community care movements? *Adm Policy Ment Health*. 2000;27(6):443–9.