

# Textbook of Clinical Epidemiology

For Medical Students

Chongjian Wang  
Fen Liu  
*Editors*

 郑州大学出版社

 Springer


# Textbook of Clinical Epidemiology

Chongjian Wang • Fen Liu  
Editors

# Textbook of Clinical Epidemiology

For Medical Students

 Springer

 郑州大学出版社

*Editors*

Chongjian Wang  
Department of Epidemiology and  
Biostatistics, College of Public Health  
Zhengzhou University  
Zhengzhou, Henan, China

Fen Liu  
Department of Epidemiology and Health  
Statistics, School of Public Health  
Capital Medical University  
Beijing, China

ISBN 978-981-99-3621-2      ISBN 978-981-99-3622-9 (eBook)  
<https://doi.org/10.1007/978-981-99-3622-9>

Jointly published with Zhengzhou University Press  
The print edition is not for sale in China (Mainland). Customers from China (Mainland) please order the  
print book from: Zhengzhou University Press.

© Zhengzhou University Press 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of 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 publishers, 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 publishers nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publishers remain neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.  
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Foreword

As a newly developed basic discipline of modern medicine, clinical epidemiology attempts to apply epidemiological and statistical theories and methods to resolve the various problems in medical practice and guides clinical practice. With the rapid development of social mobility, computer availability, readily accessible powerful software, and the development of statistical methods, clinical epidemiology has increasingly become an important approach to clinical medicine and practice. This is because of its ingenious application of theories and methods of epidemiology and health statistics to clinical research to continuously enrich and optimize methodology for clinical research. Clinical epidemiology also provides tools used to obtain observable evidence from clinical trials and contributes to enhancing the development of clinical diagnosis and treatment. Clinical epidemiology is a useful tool for clinical practitioners undertaking clinical practice and scientific research, adequately learning and applying its principles will help clinicians enhance their knowledge and increase their efficiency through acquiring reliable information that is needed for decision-making.

The most important feature of the knowledge economy era is the continuous acquisition and updating of information. This book was therefore compiled in order to reflect the need to nurture students in this new era, meet the demand for talented individuals in contemporary society, and promote the exchange of science and technology between China and the West through “the Belt and Road Initiative.” Several topics pertaining to more than one aspect of science are discussed in various sections of this text. The first ten chapters of this book concentrate on the basic principles, concepts, and methodology used in clinical epidemiology while the remaining chapters are composed of its practical applications. Each chapter starts with a few key points and ends with short questions.

This book aims to provide an overview of the principles of clinical epidemiology, which is not only as a reference but also as a tool for several daily tasks. When writing a paper or reviewing articles and reports, it can aid in checking the appropriateness and the implications of concepts and words; when teaching and lecturing, it may assist in preparing notes and visual aids. Individual chapters may provide a

fresh perspective on familiar topics. Furthermore, this book can be used as a textbook for graduate and undergraduate students in medical schools, as well as a reference book for medical teachers and practitioners. Although we have attempted to be as accurate as possible, we acknowledge that any work of this scope could contain mistakes and omissions, thus, any suggestions on the improvement of this book would be appreciated.

Department of Epidemiology and  
Biostatistics, College of Public Health,  
Zhengzhou University, Zhengzhou,  
Henan, People's Republic of China

Guangcai Duan

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
	Chongjian Wang	
1.1	Brief History of Clinical Epidemiology	1
1.1.1	Early Clinical Epidemiology	2
1.1.2	Clinical Epidemiology	2
1.1.3	Modern Clinical Epidemiology	3
1.2	Definition of Clinical Epidemiology	4
1.3	Roles of Clinical Epidemiology	5
1.3.1	To Provide Scientific Ideas and Methods for Clinical Medical Research	5
1.3.2	To Provide Scientific Methods and Means for the Evaluation of Clinical Diagnosis and Treatment	6
1.3.3	To Provide a Scientific Methodology and Evidence for Clinical Decision-Making and Practice of Evidence-Based Medicine	6
1.3.4	It is Possible to Train Clinicians and Medical Scientists with Excellent Knowledge, Skills, and Quality Under the Modern Medical Model	7
1.4	Methodology of Clinical Epidemiologic Study	7
1.4.1	Design	8
1.4.1.1	Clarification of Study Aim	8
1.4.1.2	Determination of Study Methods	8
1.4.1.3	Identification of the Study Subjects	9
1.4.1.4	Determination of the Groups	10
1.4.1.5	Determination of Study Indicators	10
1.4.1.6	Determination Methods for Data Collection and Analysis	11
1.4.1.7	Determination of Study Quality Control Methods	11

1.4.2	Measurement . . . . .	11
1.4.3	Evaluation . . . . .	12
	1.4.3.1 Evaluate the Validity and Reliability of the Study Results . . . . .	12
	1.4.3.2 Evaluate the Importance of Study Results . . . . .	12
1.5	Characteristics of Clinical Epidemiology . . . . .	13
1.5.1	Group . . . . .	13
1.5.2	Comparison . . . . .	13
1.5.3	Probability Theory and Mathematical Statistics . . . . .	14
1.5.4	Social Psychology . . . . .	14
1.5.5	Integrating Medicine . . . . .	14
1.5.6	Development . . . . .	15
<b>2</b>	<b>Distribution of Disease . . . . .</b>	<b>17</b>
	Shan Zheng	
2.1	Measures of Disease Frequency . . . . .	18
2.1.1	Frequency Measures . . . . .	18
	2.1.1.1 Ratio . . . . .	18
	2.1.1.2 Proportion . . . . .	18
	2.1.1.3 Rate . . . . .	18
2.1.2	Morbidity Frequency Measures . . . . .	18
	2.1.2.1 Incidence Rate . . . . .	18
	2.1.2.2 Attack Rate . . . . .	19
	2.1.2.3 Secondary Attack Rate . . . . .	20
	2.1.2.4 Prevalence . . . . .	20
2.1.3	Mortality Frequency Measures . . . . .	21
	2.1.3.1 Mortality Rate . . . . .	21
	2.1.3.2 Case Fatality Rate . . . . .	22
	2.1.3.3 Survival Rate . . . . .	22
2.2	Epidemic Disease Occurrence . . . . .	22
2.2.1	Sporadic . . . . .	22
2.2.2	Epidemic . . . . .	23
2.2.3	Outbreak . . . . .	23
2.3	Distribution of Disease by Time, Place, and Person . . . . .	23
2.3.1	Time . . . . .	24
	2.3.1.1 Rapid Fluctuation . . . . .	24
	2.3.1.2 Seasonality . . . . .	25
	2.3.1.3 Periodicity . . . . .	25
	2.3.1.4 Secular Trend . . . . .	25
2.3.2	Place . . . . .	26
	2.3.2.1 Comparisons Among and Within Countries . . . . .	28
	2.3.2.2 Urban-Rural Comparisons . . . . .	29
	2.3.2.3 Endemic Clustering . . . . .	30
	2.3.2.4 Endemic Disease . . . . .	30



- 2.3.3 Person . . . . . 30
  - 2.3.3.1 Age . . . . . 31
  - 2.3.3.2 Gender . . . . . 32
  - 2.3.3.3 Ethnic and Racial Group . . . . . 33
  - 2.3.3.4 Occupation . . . . . 33
  - 2.3.3.5 Marital Status . . . . . 34
  - 2.3.3.6 Behavior and Lifestyles . . . . . 34
- 2.3.4 Combinations . . . . . 35
- 3 Descriptive Study . . . . . 37**

Zhenxing Mao and Wenqian Huo

  - 3.1 Introduction . . . . . 38
    - 3.1.1 Concept . . . . . 38
    - 3.1.2 Characteristics of Descriptive Studies . . . . . 38
    - 3.1.3 Application . . . . . 38
  - 3.2 Case and Case Series Report . . . . . 39
    - 3.2.1 Concept . . . . . 39
    - 3.2.2 Application . . . . . 39
      - 3.2.2.1 Identifying New Diseases . . . . . 39
      - 3.2.2.2 Establishing the Diagnosis . . . . . 39
      - 3.2.2.3 Forming an Etiological Hypothesis . . . . . 40
      - 3.2.2.4 Identifying Early Disease Outbreaks and Epidemics . . . . . 40
    - 3.2.3 Case . . . . . 40
      - 3.2.3.1 Estrogen Chemical Bisphenol a and Breast Cancer . . . . . 40
      - 3.2.3.2 Occupational Exposure to Vinyl Chloride and Hepatic Hemangioma . . . . . 41
      - 3.2.3.3 AIDS Discovery Case . . . . . 41
    - 3.2.4 Bias . . . . . 41
    - 3.2.5 Limitation . . . . . 42
  - 3.3 Cross-Sectional Study . . . . . 42
    - 3.3.1 Concept . . . . . 42
    - 3.3.2 Application . . . . . 43
    - 3.3.3 Classification . . . . . 43
      - 3.3.3.1 Census . . . . . 43
      - 3.3.3.2 Sampling Survey . . . . . 45
    - 3.3.4 Design and Implementation . . . . . 46
      - 3.3.4.1 Clarifying the Purpose and Type of Investigation . . . . . 46
      - 3.3.4.2 Identifying Subjects . . . . . 46
      - 3.3.4.3 Determining Sample Size . . . . . 46
      - 3.3.4.4 Determining the Sampling Method . . . . . 49
      - 3.3.4.5 Data Collection, Collation, and Analysis . . . . . 51

3.3.5	Bias and Control . . . . .	53
3.3.5.1	Selection Bias . . . . .	54
3.3.5.2	Information Bias . . . . .	54
3.3.5.3	Confounding Bias . . . . .	54
3.3.6	Strengths and Limitations . . . . .	55
3.3.6.1	Strengths . . . . .	55
3.3.6.2	Limitations . . . . .	55
3.3.7	Cases . . . . .	55
3.3.7.1	Purpose and Type of Study . . . . .	56
3.3.7.2	Subjects and Sample Size . . . . .	56
3.3.7.3	Research Content and Data Collection, Collation, and Analysis . . . . .	56
3.3.7.4	Conclusion . . . . .	57
3.4	Ecological Study . . . . .	57
3.4.1	Concept . . . . .	57
3.4.2	Type of Study Design . . . . .	57
3.4.2.1	Ecological Comparison Study . . . . .	57
3.4.2.2	Ecological Trend Study . . . . .	57
3.4.3	The Main Application . . . . .	58
3.4.4	Bias . . . . .	58
3.4.5	Strengths and Limitations . . . . .	59
3.4.5.1	Strengths . . . . .	59
3.4.5.2	Limitations . . . . .	59
3.4.6	Case . . . . .	60
<b>4</b>	<b>Cohort Study . . . . .</b>	<b>61</b>
	Li Liu	
4.1	Introduction . . . . .	61
4.1.1	Definition . . . . .	61
4.1.1.1	Observational Study . . . . .	63
4.1.1.2	Setting up a Comparison Group . . . . .	63
4.1.1.3	From “Cause” to “Outcome” . . . . .	63
4.1.2	Types of Cohort Study . . . . .	63
4.1.2.1	Prospective Cohort Study . . . . .	63
4.1.2.2	Retrospective Cohort Study (Historical Cohort Study) . . . . .	63
4.1.2.3	Ambispective Cohort Study . . . . .	64
4.2	Design of a Cohort Study . . . . .	65
4.2.1	Selection of the Cohort . . . . .	65
4.2.1.1	Choice of the Exposure Population . . . . .	65
4.2.1.2	Choice of Control Population . . . . .	66
4.2.2	Determine the Sample Size . . . . .	68
4.2.2.1	Matters to Be Considered when Calculating Sample Size . . . . .	68
4.2.2.2	Four Factors Affecting Sample Size . . . . .	68
4.2.2.3	Calculation of Sample Size . . . . .	68

- 4.2.3 Follow-Up . . . . . 69
  - 4.2.3.1 Purpose of Follow-Up . . . . . 69
  - 4.2.3.2 Follow-up Methods . . . . . 69
  - 4.2.3.3 Follow-up Contents . . . . . 69
  - 4.2.3.4 Endpoint of Observation . . . . . 70
  - 4.2.3.5 Follow-Up Interval . . . . . 70
  - 4.2.3.6 The Termination Time of Observation . . . . . 70
- 4.2.4 Quality Control . . . . . 70
  - 4.2.4.1 Selection and Training of the Investigators . . . . . 70
  - 4.2.4.2 Preparation of an Investigator’s Handbook . . . . . 71
  - 4.2.4.3 Supervision during the Follow-Up . . . . . 71
- 4.3 Data Collection and Analysis . . . . . 71
  - 4.3.1 Data Collection . . . . . 71
  - 4.3.2 Measures of Outcome Frequency . . . . . 72
    - 4.3.2.1 The Basic  $2 \times 2$  Tables Summarizing the Results of a Cohort Study . . . . . 72
    - 4.3.2.2 Person-Time . . . . . 73
    - 4.3.2.3 Standardized Mortality Ratio (SMR) . . . . . 75
    - 4.3.2.4 Statistical Tests . . . . . 75
  - 4.3.3 Measures of Association . . . . . 76
    - 4.3.3.1 Relative Risk (RR) . . . . . 76
    - 4.3.3.2 Attributable Risk (AR) and Attributable Fraction (AF) . . . . . 77
    - 4.3.3.3 Population Attributable Risk (PAR) and Population Attributable Fraction (PAF) . . . . . 78
    - 4.3.3.4 Dose-Effect Relationship . . . . . 79
- 4.4 Common Bias and Controlling . . . . . 79
  - 4.4.1 Selection Bias . . . . . 79
  - 4.4.2 Information Bias . . . . . 80
  - 4.4.3 Confounding . . . . . 80
- 4.5 Advantages and Disadvantages of Cohort Studies . . . . . 81
  - 4.5.1 Advantages of Cohort Studies . . . . . 81
  - 4.5.2 Disadvantages of Cohort Studies . . . . . 81
- 4.6 Example of a Cohort Study . . . . . 82
- 5 Case-Control Studies . . . . . 83**
  - Qian Wu
  - 5.1 Overview of Case-Control Studies . . . . . 83
    - 5.1.1 History . . . . . 84
    - 5.1.2 Definition . . . . . 84
    - 5.1.3 Type of Design Case-Control Studies . . . . . 85
    - 5.1.4 Characteristics of Case-Control Study . . . . . 86
    - 5.1.5 Application . . . . . 86
      - 5.1.5.1 Example of a Case-Control Study . . . . . 86

5.2	Design of Case-Control Studies . . . . .	87
5.2.1	Basic Principles . . . . .	87
5.2.2	Selection of Cases . . . . .	88
5.2.3	Selection of Controls . . . . .	88
	5.2.3.1 Population Controls . . . . .	89
	5.2.3.2 Hospital or Disease Registry Controls . . . . .	89
5.2.4	Matching . . . . .	90
	5.2.4.1 Matching Type . . . . .	90
	5.2.4.2 Overmatching . . . . .	91
5.2.5	Exposure . . . . .	91
5.2.6	Sample Size . . . . .	92
5.3	Data Collection and Analysis . . . . .	93
5.3.1	Main Analysis Objectives . . . . .	94
5.3.2	Descriptive Analysis . . . . .	94
5.3.3	Statistical Inference . . . . .	94
	5.3.3.1 Unmatched (Frequency Matching) Design . . . . .	95
	5.3.3.2 Matched Design . . . . .	98
5.4	Common Bias and Controlling . . . . .	100
5.4.1	Selection Bias . . . . .	100
	5.4.1.1 Prevalence-Incidence Bias . . . . .	101
	5.4.1.2 Unmasking Bias . . . . .	101
	5.4.1.3 Subject Refuses Participation . . . . .	101
5.4.2	Information Bias . . . . .	102
5.4.3	Confounding Bias . . . . .	102
5.5	Strengths and Weaknesses of Case-Control Studies . . . . .	103
5.5.1	Advantage of the Case-Control Study . . . . .	103
5.5.2	Disadvantage of the Case-Control Study . . . . .	104
<b>6</b>	<b>Experimental Epidemiology . . . . .</b>	<b>105</b>
	Xing Liu	
6.1	Basic Ideas of Experimental Study . . . . .	106
6.1.1	Study Question . . . . .	106
6.1.2	Choice of Intervention . . . . .	107
6.1.3	Choice of Control . . . . .	107
	6.1.3.1 Standard Control . . . . .	107
	6.1.3.2 Placebo Control . . . . .	108
	6.1.3.3 Self-Control . . . . .	108
	6.1.3.4 Cross-Over Control . . . . .	108
6.1.4	Randomization . . . . .	109
	6.1.4.1 The Randomization Process . . . . .	110
	6.1.4.2 Simple Randomization . . . . .	110
	6.1.4.3 Blocked Randomization . . . . .	110
	6.1.4.4 Stratified Randomization . . . . .	111

6.1.5	Blinding . . . . .	111
6.1.5.1	Single-Blind . . . . .	112
6.1.5.2	Double-Blind . . . . .	112
6.1.5.3	Triple-Blind . . . . .	112
6.1.6	Data Analysis . . . . .	112
6.1.7	Sample Size . . . . .	113
6.1.7.1	Sample Size Calculation for Dichotomous Response Variables . . . . .	114
6.1.7.2	Sample Size Calculation for Continuous Response Variables . . . . .	114
6.1.7.3	Sample Size Calculation for “Time to Failure” . . . . .	114
6.2	Clinical Trial . . . . .	114
6.2.1	Basic Ideas of Clinical Trial . . . . .	114
6.2.2	Phases of Clinical Trial . . . . .	115
6.2.2.1	Phase I Studies . . . . .	115
6.2.2.2	Phase II Studies . . . . .	115
6.2.2.3	Phase III Studies . . . . .	116
6.2.2.4	Phase IV Studies . . . . .	116
6.2.3	Case Study of Clinical Trial . . . . .	116
6.3	Field Trial . . . . .	117
6.3.1	Basic Ideas of Field Trial . . . . .	117
6.3.2	Design and Implementation . . . . .	117
6.3.2.1	A specified Question . . . . .	117
6.3.2.2	Inclusion and Exclusion Criteria . . . . .	118
6.3.2.3	Choice of Intervention . . . . .	118
6.3.2.4	Time and Interval of Follow-up . . . . .	118
6.3.3	Case Study of Field Trial . . . . .	118
6.4	Community Trial . . . . .	119
6.4.1	Basic Ideas of Community Trial . . . . .	119
6.4.2	Case Study . . . . .	119
<b>7</b>	<b>Screening and Diagnostic Tests . . . . .</b>	<b>123</b>
	Fen Liu	
7.1	Design a Screening or Diagnostic Test . . . . .	124
7.1.1	Gold Standard (Reference Standard) . . . . .	124
7.1.2	Study Subjects . . . . .	124
7.1.3	Sample Size . . . . .	125
7.2	Evaluation of a Screening Test . . . . .	125
7.2.1	Validity of a Screening Test . . . . .	126
7.2.1.1	Sensitivity . . . . .	126
7.2.1.2	Specificity . . . . .	126
7.2.1.3	Youden’s Index . . . . .	127
7.2.1.4	Likelihood Ratio . . . . .	127

- 7.2.2 Evaluation of the Reliability of a Test . . . . . 128
  - 7.2.2.1 Coefficient of Variation . . . . . 128
  - 7.2.2.2 Agreement Rate and Kappa Statistic . . . . . 128
- 7.2.3 Predictive Value . . . . . 129
- 7.2.4 Determination of Cutoff Point for a Screening Test . . . . . 132
  - 7.2.4.1 ROC Curve . . . . . 133
  - 7.2.4.2 The Area under ROC Curve . . . . . 134
- 7.3 Improving the Efficiency of Screening and Diagnostic Tests . . . . . 134
  - 7.3.1 Selecting Population with a High Prevalence . . . . . 134
  - 7.3.2 Use of Multiple Tests . . . . . 134
    - 7.3.2.1 Simultaneous Testing . . . . . 134
    - 7.3.2.2 Sequential Testing . . . . . 135
- 7.4 Potential Bias in Screening Tests . . . . . 136
  - 7.4.1 Volunteer Bias . . . . . 136
  - 7.4.2 Lead-Time Bias . . . . . 136
  - 7.4.3 Length-Time Bias . . . . . 137
- 8 Bias . . . . . 139**
  - Lu Long
  - 8.1 Introduction of Bias . . . . . 139
  - 8.2 Selection Bias . . . . . 140
    - 8.2.1 Definition . . . . . 140
    - 8.2.2 Classification . . . . . 140
      - 8.2.2.1 Self-Selection Bias . . . . . 140
      - 8.2.2.2 Berksonian Bias . . . . . 141
      - 8.2.2.3 Detection Signal Bias . . . . . 142
      - 8.2.2.4 Neyman Bias . . . . . 143
      - 8.2.2.5 Loss of Follow-Up . . . . . 143
    - 8.2.3 Control . . . . . 143
      - 8.2.3.1 Scientific Research Design . . . . . 143
      - 8.2.3.2 Develop Strict Inclusion and Exclusion Standards . . . . . 144
      - 8.2.3.3 Maximize Response Rates . . . . . 144
      - 8.2.3.4 Randomization Principle . . . . . 144
  - 8.3 Information Bias . . . . . 145
    - 8.3.1 Definition . . . . . 145
    - 8.3.2 Classification . . . . . 145
      - 8.3.2.1 Differential Misclassification . . . . . 145
      - 8.3.2.2 Nondifferential Misclassification . . . . . 146
    - 8.3.3 Control . . . . . 147
      - 8.3.3.1 Material Collection . . . . . 147
      - 8.3.3.2 Objective Research Indicators . . . . . 148
      - 8.3.3.3 Investigation Skills . . . . . 148

8.4	Confounding Bias . . . . .	148
8.4.1	Definition . . . . .	148
8.4.2	Confounding . . . . .	149
8.4.3	Control . . . . .	150
8.4.3.1	Random Allocation . . . . .	150
8.4.3.2	Restrict . . . . .	150
8.4.3.3	Matching . . . . .	151
8.4.3.4	Data Analysis . . . . .	151
<b>9</b>	<b>Cause of Disease and Causal Inference . . . . .</b>	<b>153</b>
	Li Ye	
9.1	Introduction . . . . .	153
9.2	Cause of Disease in Epidemiology . . . . .	154
9.2.1	The Concept of Cause in Epidemiology and its Development History . . . . .	154
9.2.2	Classification of Cause . . . . .	155
9.2.3	Causation Models . . . . .	156
9.2.3.1	Triangle Model . . . . .	156
9.2.3.2	Wheel Model . . . . .	157
9.2.3.3	Chain of Causation Model . . . . .	157
9.2.3.4	Web of Causation Model . . . . .	158
9.2.4	Sufficient Cause and Necessary Cause . . . . .	158
9.3	Epidemiologic Methods of Causation . . . . .	160
9.3.1	Epidemiologic Study Designs for Causation . . . . .	160
9.3.1.1	Descriptive Studies . . . . .	160
9.3.1.2	Analytical Studies (Case-Control Studies, Cohort Studies) . . . . .	161
9.3.1.3	Experimental Studies . . . . .	161
9.3.2	Mill’s Canons-the Logical Basis of Causation . . . . .	162
9.3.2.1	Method of Agreement . . . . .	162
9.3.2.2	Method of Difference . . . . .	162
9.3.2.3	Joint Methods of Agreement and Difference . . . . .	163
9.3.2.4	Method of Concomitant Variations . . . . .	163
9.3.2.5	Method of Residue . . . . .	164
9.4	Causal Inference . . . . .	164
9.4.1	Association Vs. Causation . . . . .	164
9.4.1.1	Chance Association . . . . .	165
9.4.1.2	Spurious Association . . . . .	165
9.4.1.3	Noncausal Association . . . . .	166
9.4.1.4	Causal Association . . . . .	166
9.4.2	Evaluating Causal Association—Hill’s Criteria . . . . .	167
9.4.2.1	Temporal Relationship . . . . .	167
9.4.2.2	Strength of Association . . . . .	168

- 9.4.2.3 Dose-Response Relationship . . . . . 168
- 9.4.2.4 Consistency . . . . . 169
- 9.4.2.5 Biologic Plausibility . . . . . 169
- 9.4.2.6 Reversibility . . . . . 169
- 9.4.2.7 Specificity . . . . . 169
- 9.4.2.8 Analogy . . . . . 170
- 9.4.2.9 Experimental Evidence . . . . . 170
- 9.4.3 An Example of Causal Inference Using Hill’s Criteria . . . . . 170
  - 9.4.3.1 Temporality of Association . . . . . 171
  - 9.4.3.2 Strength of Association . . . . . 171
  - 9.4.3.3 Consistency . . . . . 171
  - 9.4.3.4 Dose-Response Relationship . . . . . 171
  - 9.4.3.5 Biologic Plausibility . . . . . 171
  - 9.4.3.6 Experimental Evidence . . . . . 171
- 10 Disease Prevention and Surveillance . . . . . 173**
  - Chunhua Song
    - 10.1 Prevention Strategies and Measures . . . . . 173
      - 10.1.1 Strategy and Implementation for Prevention . . . . . 173
      - 10.1.2 Disease Prevention . . . . . 174
        - 10.1.2.1 The Definition of Disease Prevention . . . . . 174
        - 10.1.2.2 The Development of Disease . . . . . 174
        - 10.1.2.3 The Three Levels of Prevention of Disease . . . . . 175
      - 10.1.3 Health Protection and Promotion . . . . . 177
        - 10.1.3.1 Health Protection . . . . . 177
        - 10.1.3.2 Health Education . . . . . 178
        - 10.1.3.3 Health Management . . . . . 178
        - 10.1.3.4 Health Promotion . . . . . 179
        - 10.1.3.5 Global Health Strategies and Practice . . . . . 179
    - 10.2 Public Health Monitoring . . . . . 181
      - 10.2.1 Introduction of Public Health Surveillance . . . . . 181
        - 10.2.1.1 The Basic Concept of Public Health Surveillance . . . . . 181
        - 10.2.1.2 The Purpose and Application of Public Health Surveillance . . . . . 182
      - 10.2.2 Categories of Public Health Surveillance . . . . . 184
        - 10.2.2.1 Surveillance of Disease . . . . . 184
        - 10.2.2.2 Symptom Surveillance . . . . . 186
      - 10.2.3 Methods of Public Health Surveillance . . . . . 187
        - 10.2.3.1 Surveillance Methods . . . . . 187
        - 10.2.3.2 Surveillance Methods and Techniques . . . . . 189
        - 10.2.3.3 Attention in Public Health Surveillance . . . . . 191



10.2.4	Procedures and Assessment of Public Health Surveillance . . . . .	192
10.2.4.1	Basic Procedures of Public Health Surveillance . . . . .	192
10.2.4.2	Evaluation of Public Health Surveillance System . . . . .	194
<b>11</b>	<b>Communicable Diseases Epidemiology . . . . .</b>	<b>197</b>
	Rongguang Zhang	
11.1	Infection Process . . . . .	198
11.1.1	Infection Process . . . . .	198
11.1.2	Spectrum of Infection . . . . .	198
11.2	Epidemic Process . . . . .	199
11.2.1	Definition . . . . .	199
11.2.2	Three Links in the Epidemic Process . . . . .	199
11.2.2.1	Sources of Infection . . . . .	199
11.2.2.2	Routes of Transmission . . . . .	200
11.2.2.3	Herd Susceptibility . . . . .	201
11.2.3	Two Factors Affecting the Epidemic Process . . . . .	201
11.2.3.1	Natural Factors . . . . .	202
11.2.3.2	Social Factors . . . . .	202
11.2.4	Epidemic Focus and Epidemic Process . . . . .	202
11.2.4.1	Epidemic Focus . . . . .	202
11.2.4.2	Epidemic Process . . . . .	203
11.3	Strategy and Implementation . . . . .	203
11.3.1	Strategies for Control of Communicable Diseases . . . . .	203
11.3.1.1	Population Strategy . . . . .	204
11.3.1.2	High-Risk Strategy . . . . .	204
11.3.2	Measures for Control of Communicable Diseases . . . . .	205
11.3.2.1	Surveillance of Communicable Diseases . . . . .	205
11.3.2.2	Measures on Sources of Infection . . . . .	205
11.3.2.3	Measures on Routes of Infection . . . . .	206
11.3.2.4	Measures on Susceptible Populations . . . . .	206
11.4	Immunization Program and Effectiveness . . . . .	207
11.4.1	Immunization . . . . .	207
11.4.2	Immunization Program . . . . .	208
11.4.3	Evaluation of Immune Effectiveness . . . . .	208
11.5	Emerging Communicable Diseases . . . . .	209
11.5.1	Definition . . . . .	209
11.5.2	Main Emerging Communicable Diseases . . . . .	209
11.6	Summary . . . . .	209
<b>12</b>	<b>Epidemiology of Noncommunicable Diseases . . . . .</b>	<b>213</b>
	Jie Yang and Man Li	
12.1	Introduction . . . . .	213
12.1.1	Definition . . . . .	213
12.1.2	The Influence of NCDs on Health and Society . . . . .	214

12.2	Epidemiological Features . . . . .	214
12.2.1	Overall Global NCDs Outlook . . . . .	214
12.2.2	Epidemiological Features of the Risk Factors of NCDs . . . . .	216
12.2.2.1	Tobacco Use . . . . .	216
12.2.2.2	Alcohol Use . . . . .	217
12.2.2.3	Unhealthy Diet . . . . .	217
12.2.2.4	Physical Inactivity . . . . .	217
12.2.2.5	Raised Blood Pressure . . . . .	218
12.2.2.6	Overweight/Obesity . . . . .	218
12.3	Risk Factors of Several Common NCDs . . . . .	218
12.3.1	Cardiovascular and Cerebrovascular Disease . . . . .	218
12.3.1.1	Stroke . . . . .	218
12.3.1.2	Coronary Heart Disease . . . . .	219
12.3.2	T2DM . . . . .	220
12.3.2.1	Genetic Factor . . . . .	220
12.3.2.2	Overweight/Obesity . . . . .	221
12.3.2.3	Physical Inactivity . . . . .	221
12.3.2.4	Unhealthy Diet . . . . .	221
12.3.2.5	Malnutrition . . . . .	222
12.3.2.6	Impaired Glucose Tolerance (IGT) . . . . .	222
12.3.2.7	Insulin Resistance . . . . .	222
12.3.2.8	Maternal Diabetes . . . . .	222
12.3.3	Cancer . . . . .	223
12.3.3.1	Physical Factors . . . . .	223
12.3.3.2	Tobacco Use . . . . .	223
12.3.3.3	Alcohol Use . . . . .	223
12.3.3.4	Dietary Factors . . . . .	223
12.3.3.5	Occupational Exposures . . . . .	224
12.3.3.6	Biological Factors . . . . .	224
12.3.3.7	Genetic Factors . . . . .	224
12.3.3.8	Other Factors . . . . .	224
12.4	Prevention and Control of NCDs . . . . .	225
12.4.1	Prevention Strategy . . . . .	225
12.4.2	Prevention Measures . . . . .	226
<b>13</b>	<b>Epidemiology of Public Health Emergencies . . . . .</b>	<b>227</b>
	Hong Zhu	
13.1	Basic Conception of Public Health Emergencies . . . . .	228
13.1.1	Definition of Public Health Emergencies . . . . .	228
13.1.2	Characteristics of Public Health Emergencies . . . . .	228
13.1.3	Classification of Public Health Emergencies . . . . .	229
13.1.4	Phases of Public Health Emergencies . . . . .	230
13.1.5	Harm Caused by Public Health Emergencies . . . . .	231

13.2	Basic Principles and Application of Epidemiology in Public Health Emergencies . . . . .	232
13.2.1	Role of Epidemiology in Public Health Emergencies . . . . .	232
13.2.2	Key Epidemiological Indicators . . . . .	232
13.2.3	Outbreak Investigation . . . . .	233
13.2.3.1	Purpose of Outbreak Investigation . . . . .	233
13.2.3.2	Three Elemental Epidemiological Designs in an Outbreak Investigation . . . . .	234
13.2.3.3	Key Steps in Carrying Out Outbreak Investigation . . . . .	234
13.2.4	Disaster Investigation . . . . .	236
13.2.4.1	Purpose of Disaster Investigation . . . . .	236
13.2.4.2	Key Steps in Carrying Out Disaster Investigation . . . . .	237
13.3	Public Health Emergency Preparedness . . . . .	238
13.3.1	Definition of Public Health Emergency Preparedness . . . . .	238
13.3.2	Significance of Public Health Emergency Preparedness . . . . .	239
13.3.3	The Main Activities of Public Health Emergency Preparedness . . . . .	239
13.4	Public Health Emergency Response . . . . .	242
13.4.1	Definition of Public Health Emergency Response . . . . .	242
13.4.2	Significance of Public Health Emergency Response . . . . .	242
13.4.3	The Main Activities of Public Health Emergency Response . . . . .	242
13.4.3.1	Ensuring Availability of Preventive and Emergency Medical Treatment . . . . .	242
13.4.3.2	Preventing Secondary Public Health Emergencies After Disaster . . . . .	243
13.4.3.3	Interrupting the Route of Transmission . . . . .	243
13.4.3.4	Remediating of Environmental Health Conditions . . . . .	243
13.4.3.5	Performing Laboratory Analyses to Support Epidemiology and Surveillance . . . . .	244
13.4.3.6	Communicating with Media and Delivering Message to the Public . . . . .	244
13.5	Summary . . . . .	244
<b>14</b>	<b>Molecular Epidemiology . . . . .</b>	<b>247</b>
	Hui Wang	
14.1	Introduction . . . . .	247
14.1.1	Concept . . . . .	247
14.1.2	Characteristic . . . . .	248
14.2	Classes of Biomarkers . . . . .	248
14.2.1	Biomarkers of Exposure . . . . .	248
14.2.2	Biomarkers of Effects . . . . .	249

14.2.3	Biomarkers of Susceptibility . . . . .	250
14.2.4	Biomarker Selection . . . . .	250
14.3	Main Research Methods Used in Molecular Epidemiology . . . . .	252
14.3.1	Study Design in Molecular Epidemiology . . . . .	252
14.3.1.1	Cross-Sectional Studies . . . . .	252
14.3.1.2	Case-Control Studies . . . . .	253
14.3.1.3	Cohort Studies . . . . .	253
14.3.2	Main Molecular Methods Used in Molecular Epidemiology . . . . .	254
14.3.2.1	Electrophoretic Mobility Shift Assay (EMSA) . . . . .	254
14.3.2.2	Dual Luciferase Reporter Assay . . . . .	255
14.3.2.3	The Comet Assay . . . . .	256
14.3.2.4	Micronucleus (MN) Assay . . . . .	256
14.3.3	Genome-Wide Association Studies (GWAS) . . . . .	257
14.3.4	Mendelian Randomization (MR) . . . . .	258
14.4	Application and Prospecption . . . . .	259
14.4.1	Control and Prevention of Infectious Diseases . . . . .	259
14.4.1.1	Outbreak Investigation . . . . .	259
14.4.1.2	Trace Dissemination of a Specific Subtype of Pathogen Across Time and Space . . . . .	260
14.4.1.3	Determine the Origin of an Epidemic . . . . .	260
14.4.1.4	Follow the Emergence and Spread of New Infections . . . . .	261
14.4.1.5	Identify Previously Unknown or Uncultivable Infectious Microbes . . . . .	261
14.4.2	Control and Prevention of Chronic Diseases . . . . .	262
14.4.2.1	Improving the Understanding of Mechanism of Pathogenesis . . . . .	262
14.4.2.2	Evaluating the Susceptibility of Individual and Defining the Risk Population . . . . .	263
14.4.3	Conclusions . . . . .	263
<b>15</b>	<b>Pharmacoepidemiology</b> . . . . .	<b>265</b>
	Xiaotian Liu and Jian Hou	
15.1	A Brief History and Definition . . . . .	265
15.1.1	A Brief History of Pharmacoepidemiology . . . . .	265
15.1.2	Definition of Pharmacoepidemiology . . . . .	267
15.1.3	Drug-Related Concepts . . . . .	267
15.2	Main Research Contents . . . . .	268
15.2.1	Drug Safety Evaluation . . . . .	268
15.2.2	Drug Effectiveness Evaluation . . . . .	269
15.2.3	Drug Utilization Study . . . . .	269
15.2.4	Pharmacoeconomic Study . . . . .	269

15.3	Aims and Significances . . . . .	270
15.3.1	Aims of Pharmacoepidemiology . . . . .	270
15.3.2	Significances of pharmacoepidemiology . . . . .	271
15.3.2.1	Improve the Quality of Premarketing Clinical Trials . . . . .	271
15.3.2.2	Post-Marketing Study of Drug . . . . .	271
15.4	Methods of Pharmacoepidemiology . . . . .	272
15.4.1	Case Report and Case Series Study . . . . .	272
15.4.2	Ecological Study . . . . .	272
15.4.3	Cross-Sectional Study . . . . .	274
15.4.4	Case-Control Study . . . . .	274
15.4.5	Cohort Study . . . . .	274
15.4.6	Experimental Study . . . . .	275
15.4.7	Systematic Review and Meta-Analysis . . . . .	276
15.4.8	Real-world Study . . . . .	276
15.4.9	Newly Derived Study Design . . . . .	276
15.5	Data Collection and Analysis . . . . .	277
15.5.1	Data Collection . . . . .	277
15.5.1.1	Routine Data . . . . .	277
15.5.1.2	The Literature . . . . .	278
15.5.1.3	ADR Monitoring and Reporting System . . . . .	278
15.5.2	Data Processing and Analysis . . . . .	278
15.5.2.1	Mining and Analysis of ADR Monitoring Database . . . . .	279
15.5.2.2	Mining and Analysis of Prescription Database . . . . .	279
<b>16</b>	<b>Evidence-Based Medicine and Systematic Review . . . . .</b>	<b>281</b>
	Qi Gao and Huiping Zhu	
16.1	Evidence-Based Medicine . . . . .	281
16.1.1	Concept . . . . .	281
16.1.2	Development of EBM . . . . .	282
16.1.3	Categories of EBM Practice . . . . .	284
16.1.4	Procedures of Practicing EBM . . . . .	285
16.2	Systematic Review and Meta-Analysis . . . . .	287
16.2.1	Systematic Review . . . . .	287
16.2.1.1	Cochrane Systematic Review . . . . .	288
16.2.1.2	Importance of Systematic Review . . . . .	288
16.2.1.3	The Difference Between Systematic Review and Traditional Review . . . . .	289
16.2.1.4	How to Do a Systematic Review? . . . . .	289
16.2.1.5	Evaluation and Application of Systematic Review . . . . .	294
16.2.1.6	Methods of Evaluating System Review . . . . .	298
16.2.1.7	Application of Systematic Review . . . . .	299

- 16.3 Meta-Analysis . . . . . 301
  - 16.3.1 Introduction . . . . . 301
    - 16.3.1.1 Basic Concepts . . . . . 301
  - 16.3.2 Steps to Perform a Meta-Analysis . . . . . 301
    - 16.3.2.1 Data Extraction . . . . . 301
    - 16.3.2.2 Data Types and Effect Size . . . . . 302
    - 16.3.2.3 Heterogeneity Test . . . . . 302
    - 16.3.2.4 Combining Effect Size Estimates and Hypothesis Testing . . . . . 303
  - 16.3.3 Fixed Effect Model and Random Effect Model . . . . . 303
    - 16.3.3.1 Fixed Effect Model . . . . . 303
    - 16.3.3.2 Random Effect Model . . . . . 304
  - 16.3.4 Evaluating the Result of a Meta-Analysis . . . . . 304
    - 16.3.4.1 Heterogeneity Test . . . . . 304
    - 16.3.4.2 Robustness of Meta-Analysis Results . . . . . 304
    - 16.3.4.3 Applicability of the Meta-Analysis Results . . . . . 305
- 17 Disease Prognosis . . . . . 307**
  - Fang Wang
  - 17.1 Basic Concepts . . . . . 308
    - 17.1.1 Concept of Prognosis . . . . . 308
    - 17.1.2 Natural History and Clinical Course of Disease . . . . . 308
    - 17.1.3 Prognostic Factors . . . . . 309
  - 17.2 Design of a Prognosis Study . . . . . 310
    - 17.2.1 Research Methods . . . . . 310
    - 17.2.2 The Patient Sample . . . . . 310
    - 17.2.3 Determine the Starting Time Point . . . . . 311
    - 17.2.4 Determine the Outcomes . . . . . 311
    - 17.2.5 Sample Size . . . . . 312
    - 17.2.6 Follow-Up . . . . . 312
  - 17.3 Describing Prognosis . . . . . 312
    - 17.3.1 Case Fatality . . . . . 312
    - 17.3.2 Remission Rate . . . . . 312
    - 17.3.3 Recurrence Rate . . . . . 313
    - 17.3.4 Disability Rate . . . . . 313
    - 17.3.5 Quality of Life . . . . . 313
  - 17.4 Analysis for Prognosis Study Data . . . . . 314
    - 17.4.1 Calculating Survival Rate . . . . . 314
      - 17.4.1.1 Five-Year Survival . . . . . 314
      - 17.4.1.2 Life Tables . . . . . 314
      - 17.4.1.3 Kaplan–Meier Analysis . . . . . 315
    - 17.4.2 Several Points About Interpreting Survival Curves . . . . . 316
    - 17.4.3 Comparison of the Survival Curves . . . . . 316
    - 17.4.4 Dealing with Multiple Prognostic Factors . . . . . 317

17.5	Common Bias and Controlling . . . . .	317
17.5.1	Bias in Prognosis Study . . . . .	317
17.5.1.1	Assembly Bias . . . . .	317
17.5.1.2	Migration . . . . .	317
17.5.1.3	Loss to Follow-Up . . . . .	318
17.5.1.4	Survival Cohort Bias . . . . .	318
17.5.1.5	Zero Bias . . . . .	318
17.5.1.6	Measurement Bias . . . . .	318
17.5.2	Control of Bias . . . . .	319
17.5.2.1	Randomization . . . . .	319
17.5.2.2	Matching . . . . .	319
17.5.2.3	Restriction . . . . .	319
17.5.2.4	Stratification . . . . .	320
17.5.2.5	Standardization . . . . .	320
17.5.2.6	Multivariable Analysis . . . . .	320
17.5.2.7	Other Methods to Control Bias in Prognosis . . . . .	321
<b>18</b>	<b>Nosocomial Infections . . . . .</b>	<b>323</b>
	Zhijiang Zhang	
18.1	Introduction . . . . .	323
18.2	Definition and Diagnostic Standards . . . . .	324
18.3	Nosocomial Infection Sites . . . . .	325
18.3.1	Surgical Sites . . . . .	325
18.3.2	Respiratory System . . . . .	325
18.3.3	Bacteremia . . . . .	325
18.3.4	Urinary Tract . . . . .	326
18.4	Microorganisms . . . . .	326
18.4.1	Normal Microorganisms in Nosocomial Infections . . . . .	326
18.4.1.1	Bacteria . . . . .	326
18.4.1.2	Viruses . . . . .	326
18.4.1.3	Parasites and Fungi . . . . .	326
18.4.2	Antimicrobial Resistance and Nosocomial Infections . . . . .	327
18.5	Categories of Nosocomial Infections . . . . .	327
18.5.1	Endogenous Infections . . . . .	327
18.5.2	Exogenous Infections . . . . .	327
18.6	Epidemic Process of Nosocomial Infection . . . . .	328
18.6.1	Source of Infection . . . . .	328
18.6.2	Route of Transmission . . . . .	328
18.6.3	Susceptible Population . . . . .	328
18.7	Prevention of Nosocomial Infections . . . . .	329
18.7.1	Preventing Human-to-Human Transmission . . . . .	329
18.7.1.1	Hand Decontamination . . . . .	329
18.7.1.2	Clothing . . . . .	329
18.7.1.3	Masks . . . . .	329

- 18.7.1.4 Gloves . . . . . 329
- 18.7.1.5 Safe Injection and Other Skin-Piercing Practice . . . . . 330
- 18.7.2 Preventing Transmission from Environment . . . . . 330
  - 18.7.2.1 Routine Cleaning . . . . . 330
  - 18.7.2.2 Disinfection of Equipment . . . . . 330
  - 18.7.2.3 Sterilization . . . . . 331
- 18.8 Surveillance of Nosocomial Infections . . . . . 331
  - 18.8.1 Objectives of Surveillance Programs . . . . . 331
  - 18.8.2 Implementation of Surveillance Programs . . . . . 331
  - 18.8.3 Evaluation of Surveillance Program . . . . . 332
    - 18.8.3.1 Strategy Evaluation . . . . . 332
    - 18.8.3.2 Feedback Evaluation . . . . . 332
    - 18.8.3.3 Evaluation of Data Quality . . . . . 333
- 19 Epidemiology Design in Clinical Research . . . . . 335**
  - Yi Wang
  - 19.1 Design and Implementation of Clinical Research . . . . . 336
    - 19.1.1 Forming Research Questions . . . . . 336
    - 19.1.2 Commonly Used Epidemiological Design in Clinical Research . . . . . 337
    - 19.1.3 Collection and Analysis of Clinical Research Data . . . . . 338
      - 19.1.3.1 Data Collection . . . . . 339
      - 19.1.3.2 Data Analysis . . . . . 342
    - 19.1.4 Preparing Papers for Publication . . . . . 343
      - 19.1.4.1 Choose Target Journal(s) . . . . . 343
      - 19.1.4.2 Choose a Clear Message . . . . . 344
      - 19.1.4.3 Achieve High Quality in Writing . . . . . 345
    - 19.1.5 Common Problems in Clinical Research Design . . . . . 345
  - 19.2 Reporting Guidelines for Clinical Research Reports . . . . . 346
    - 19.2.1 Observational Studies Reporting Guidelines . . . . . 346
    - 19.2.2 Diagnostic/Prognostic Studies Reporting Guidelines . . . . . 347
    - 19.2.3 Clinical Trials Reporting Guidelines . . . . . 347
    - 19.2.4 Systematic Reviews Reporting Guidelines . . . . . 348
  - 19.3 Real-World Study . . . . . 348
    - 19.3.1 Definition of Real-World Study . . . . . 349
    - 19.3.2 The Difference Between RWS and RCT . . . . . 349
- References . . . . . 351**



# Chapter 1

## Introduction



Chongjian Wang

### Key Points

- Clinical epidemiology concerns with the application of epidemiological principles and methods in specified populations to observe, analyze, and explain issues in clinical medicine such as diagnosis, screening, treatment, prognosis as well as the cause of disease with the purpose of providing scientific evidence for clinical decision.
- The methods of clinical epidemiological studies include descriptive studies, analytical studies, experimental studies, and mathematical statistics
- The core contents of clinical epidemiology are design, measurement, and evaluation in clinical research.

### 1.1 Brief History of Clinical Epidemiology

Clinical epidemiology is a newly developed basic science that integrates clinical medicine with epidemiology by concentrating on the scientific research of clinical medicine. The development of clinical epidemiology occurred over a long period of time, with numerous doctors and persons contributing to its progress. When disease occurred in the population, the people always attempted to identify the cause of the disease and tried to control the spread of the disease. These efforts contributed to the development of clinical epidemiology.

---

C. Wang (✉)  
College of Public Health, Zhengzhou University, Zhengzhou, China

### ***1.1.1 Early Clinical Epidemiology***

Many observations were made in early epidemiology. Today, these observations may seem simple or not rigorous enough in design, but they provided very useful information toward describing diseases and their control in those days.

Hippocrates (circa 400 B.C.), the pioneer of epidemiology, attempted to explain disease occurrence from a rational rather than a supernatural viewpoint. In his essay entitled “On Airs, Waters, and Places,” Hippocrates described the epidemics of disease and suggested that environmental and host factors such as behaviors might influence the development of disease.

In the mid-fifteenth century, ships from outside Europe were asked to stay in an isolated area for 40 days before arrival at the port to protect the community against the Black Death (plague) in Europe, which was the origin of “quarantine.”

John Graunt, a London haberdasher, published his landmark analysis of mortality data in 1662. He was the first to quantify patterns of birth, death, and disease occurrence, noting men-women disparities, high infant mortality, urban-rural differences, and seasonal variations. He put forward the need to establish a comparative group when studying the rule of death and the quality of death data. His contribution was to introduce statistics into epidemiology.

### ***1.1.2 Clinical Epidemiology***

Later, in 1747, the British surgeon James Lind found that vitamin C deficiency was the cause of scurvy by dividing 12 sailors with scurvy into six groups for a comparative treatment trial, which was the first epidemiological experiment and marked the beginning of clinical epidemiology in human history.

In 1796, Edward Jenner, a British doctor, carried out a vaccination in order to prevent smallpox, which effectively controlled the spread of smallpox and pioneered active immunization for the control of infectious diseases. In the 18th century, the French Revolution had a profound impact on the development of epidemiology. Pierre Charles Alexandre Louis, one of the pioneers of modern epidemiology, explored the curative effect of bloodletting therapy on inflammatory diseases through comparative observation and studied the genetic effect on tuberculosis using life tables. In 1838, Pierre Charles Alexandre Louis and his student, William Farr, considered the father of modern vital statistics and surveillance, began to systematically collect and analyze Britain’s mortality data. They developed many of the basic practices used today in vital statistics and disease classification, extended the epidemiologic analysis of morbidity and mortality data, and looked at the effects of marital status, occupation, and attitude. They also developed many epidemiologic concepts and techniques, such as life tables and the standardization of rates that are still in use today.

During the mid-1880s, “the father of field epidemiology” named John Snow conducted a series of investigations using a dot distribution map method of case distribution to explore the prevalence of cholera in the Broad Street of London. He also analyzed the mortality rates of cholera in different water supply areas. He was the first to put forward the famous scientific statement that “cholera is transmitted through water,” and successfully controlled further spread of the epidemic by stopping the water supply from the suspected pump. Twenty years before the development of the microscope, Snow’s studies of cholera outbreaks led to the discovery of contaminated drinking water as the cause of the disease and brought about effective measures to prevent its recurrence. This became a classic example of epidemiological field investigation, analysis, and control.

### ***1.1.3 Modern Clinical Epidemiology***

Later in the 1800s, many researchers in Europe and the United States began to apply epidemiological methods to investigate disease occurrence. At that time, most investigators focused on acute infectious diseases. Around World War II and later, epidemiologists extended their methods to chronic noncommunicable diseases such as cancer and cardiovascular diseases. A series of studies, including case-control and prospective cohort studies by Richard Doll and Austin Bradford Hill, suggested a very significant association between cigarette smoking and lung cancer. The Framingham Heart Study is another classic example, which was started in 1948 when heart disease had become the leading cause of death in the United States. Finally, during the 1960s and early 1970s, health workers applied epidemiological methods to eradicate smallpox worldwide. This was an unprecedented achievement in applied epidemiology.

In 1951, Jerome Cornfield put forward the relative risk, odds ratio, and other measurement indicators. In 1959, Nathan Mantel and William Haenszel proposed stratified analysis, which was one of the most cited epidemiological study methods. In the discipline of infectious diseases, field trials of the polio vaccine, organized by Jonas Edward Salk in 1954, involving more than 1.5 million children in grades 1 to 3 in the United States, Canada, and Finland, not only confirmed the protective effect of the vaccine but also laid the foundation for the eventual eradication of polio. In 1979, Sackett summarized 35 possible biases that might occur in analytical studies. In 1985, Miettinen proposed a bias classification that included comparison, selection, and information bias.

However, the study of etiology does not solve all the problems of disease prevention and treatment. For example, epidemiologic studies neglected many problems in clinical medicine, such as research on medical and health needs, evaluation of clinical treatment effects, screening and early diagnosis of diseases, prediction of natural disease history and prognosis, and so on. In this context, many clinicians began to focus on rigorous design, measurement, and evaluation (DME) in clinical medical research, and epidemiologists collaborated with clinicians,

meanwhile, randomized control trials (RCTs) were proposed in a clinical study. Selection bias and confounding bias were eliminated using the randomization of group. The blinding principles of intervention drugs or preventive measures can eliminate the information bias in the trial process, and then ensure the authenticity of research results, thus becoming the signature method of clinical epidemiology research. As the most reliable way to assess causality in a population, RCTs became the gold standard for evaluating the effectiveness of medical interventions. Since then, as an independent discipline, clinical epidemiology began to step into modern medicine. Several representative clinical epidemiology textbooks were also published, such as *Clinical Trials: A Practical Approach* (Stuart J Pocock, 1983), *Clinical Epidemiology: The Architecture of Clinical Research* (Alvan R Feinstein, 1985), *Clinical Epidemiology* (David L Sackett, 1985) and *Clinical Trials: Design, Conduct, and Analysis* (Curtis L Meinen, 1986).

Today, epidemiological methods of investigation have become tools for answering questions in medicine and public health regarding their biological and social facets. Analyses of large databases and complicated calculations have become feasible due to the collaborations and integration with other disciplines, especially the use of computers. Clinical epidemiology has contributed to the understanding of diseases in the population, the study of etiology, and controls of some health problems, including prevention or treatment of important diseases such as cardiovascular diseases, especially ischemic heart disease, asthma, and some cancers. Regarding the identification of the possible causal risk factors for some emerging infectious diseases such as acquired immunodeficiency syndrome (AIDS), severe acute respiratory syndrome (SARS), and coronavirus disease 2019 (COVID-19), epidemiology plays a significant role.

In summary, how to systematically summarize the evidence and make clinical and prevention decisions based on the current best research results under the circumstances of limited resources is imminent, which also creates opportunities for the development of clinical epidemiology.

## 1.2 Definition of Clinical Epidemiology

The term “clinical epidemiology” is derived from the combination of clinical medicine and epidemiology. It is “clinical” because it seeks to answer clinical questions and guide clinical decision-making with the best available evidence. It is “epidemiology” because many epidemiological methods are used to answer these questions and the care of individual patients is seen in the context of the larger population of which the patient is a member. Many definitions have been proposed, but the following definition from Siyan Zhan captures the underlying principles and the public health aspect of clinical epidemiology: “clinical epidemiology is the science of the application of epidemiological principles and methods in specified populations to observe, analyze, and explain issues in clinical medicine such as

diagnosis, screening, treatment, prognosis, as well as the cause of disease with the purpose of providing scientific evidence for clinical decision-making.”

### **1.3 Roles of Clinical Epidemiology**

Clinical epidemiology is based on clinical practice and aims to resolve clinical questions. Following the gradual development of scientific technology, especially the explosive development of biology and information science, new diagnostic techniques and therapeutic strategies are constantly emerging. Clinical epidemiology creatively applies the theories and methods of epidemiology and health statistics to clinical research and thus continuously enriches and optimizes the methodology for clinical research. Clinical epidemiology also provides tools for obtaining optimal evidence from clinical trials and contributes to enhancing the level of clinical diagnosis and treatment. Clinical epidemiology is a useful tool for clinical practitioners undertaking clinical practice and scientific research; correctly learning and applying its principles will help clinicians improve their academic level and increase their efficiency in acquiring reliable information for decision-making.

#### ***1.3.1 To Provide Scientific Ideas and Methods for Clinical Medical Research***

Humans are the research object of medicine, which has dual attributes of nature and society. For a difference in genetic traits, growth, and living environments, the clinical manifestations of one disease can vary greatly, and the drug of choice may either be effective or ineffective in a particular patient groups, which leads to endless difficulties in diagnosis and treatment. The following problems need to be solved through clinical medical research: how to improve the level of diagnosis and clinical differential diagnosis; how to strengthen the evaluation of drug safety and effectiveness, and improve the level of clinical treatment. Clinical epidemiology is to provide clinical workers with scientific research methods from three aspects of design, measurement, and evaluation. DME is the core content of clinical epidemiology, summarized by clinical epidemiologists at MacMaster University in Canada, and has been recognized by peers around the world.

### ***1.3.2 To Provide Scientific Methods and Means for the Evaluation of Clinical Diagnosis and Treatment***

Problems in clinical practice include four aspects: etiology, diagnosis, treatment, and outcome. While traditional epidemiology studies focus on etiology, RCTs have solved the important problem of evaluation of clinical diagnosis and treatment, thus driving the overall development of clinical research methodology and promoting the emergence of clinical epidemiology. RCTs serve as the gold standard for exploring causal relationship and evaluating clinical outcomes in populations in epidemiology studies due to their effective control of selection, information, and confounding biases.

### ***1.3.3 To Provide a Scientific Methodology and Evidence for Clinical Decision-Making and Practice of Evidence-Based Medicine***

The core explanation of evidence-based practice and clinical decision-making according to David L Sackett, a Canadian academic is "Evidence-based medicine is the conscious, unambiguous, and deliberate utilization of the best available evidence to make decisions about the care of individual patients. Practicing evidence-based medicine means doctors need to take into account the best available research evidence, clinical experience, and patient opinion." Clinical epidemiology, with RCTs as the basic method, provides a scientific methodology for solving various clinical problems. By the 1970s, a number of RCTs had been completed, and new studies were still being published. However, how to systematically summarize and disseminate the results of these RCTs, use the evidence to guide medical practice, and improve the quality and efficiency of medical and health services became a great challenge for medical workers at that time. Therefore, clinical epidemiologists propose that clinicians should continuously obtain evidence from published clinical research papers or generate evidence through their own research to support clinical decision-making and improve the ability of literature retrieval, analysis, evaluation, and correct use of the latest research results. It is also further proposed that: how to propose problems that need to be solved clinically; how to retrieve and collect the best scientific evidence; how to evaluate the quality of this evidence; whether the effect is good or bad, and extrapolation of the results; how to formulate a reasonable patient diagnosis and treatment plan combining existing evidence and reference for other related factors, and according to the effect of the practice continuously improve the diagnosis and treatment plan, which is a complete evidence-based decision-making based on scientific thought. The development of clinical epidemiology has not only catalyzed and followed the development of evidence-based medicine theory and practice but has also triggered a medical practice revolution.

### ***1.3.4 It is Possible to Train Clinicians and Medical Scientists with Excellent Knowledge, Skills, and Quality Under the Modern Medical Model***

The modern medical model has changed from the traditional socio-psychological and biomedical model to the environmental ecological public health model. The core of this model requires modern doctors to have a comprehensive decision-making ability. In order to improve the scientific nature of clinical decision-making, it is necessary to take various clinical probabilities as the basis, guided by the theory of probability and applied strategy theory, through certain analysis and calculation, to quantify complex clinical problems, and then choose a reasonable diagnosis and treatment plan. At the same time, complex factors such as bioethics, health economics, and social value orientation should be considered to make safe, effective, and affordable clinical diagnosis and treatment decisions. Clinical epidemiology is based on clinical medicine and epidemiology, which is characterized by: under the environmental ecological public health model, it permeates and integrates with epidemiology, biostatistics, health economics, and social medicine based on clinical practice. The research object is expanded from focusing on individual cases to the corresponding whole disease population. The study site is extended from individual patients in the hospital to the comprehensive prevention and treatment of diseases in the community. The research content is transformed from the research and discussion of early detection, diagnosis, and treatment of diseases to the law of disease occurrence, development, and outcome to formulate complete clinical research ideas and improve the clinical diagnosis and treatment level. Scientific method and thinking of clinical epidemiology not only improve the ability of clinical doctors in medical research but also make them master clinical decision-making thoughts and methods, which is beneficial to clinical medicine development, improving the diagnosis and treatment level, and training a group of high-quality clinical doctors.

## **1.4 Methodology of Clinical Epidemiologic Study**

Clinical epidemiology investigates disease in the population by observing and inquiring, describing the frequency and distribution of disease, developing hypotheses through induction, synthesis, and analysis, and then testing the hypothesis through analytical study, and finally verifying the hypothesis through experimental study. After understanding the occurrence of the disease, a mathematical model is used to predict the incidence or prevalence of the disease. Thus, clinical epidemiology may be classified into three general categories: observational study, experimental study, and mathematical statistics.

The core of clinical epidemiology is to design, measure, and evaluate (DME). Any scientific study requires rigorous design, accurate measurement, and reasonable

evaluation; this is the essence of clinical scientific study, and it applies to any subject. The basic aspects of DME are represented below.

### ***1.4.1 Design***

Clinical research should have clear study aims. Additionally, research hypotheses should be put forward according to study aims, and appropriate research objects and methods should be determined to verify or test the hypotheses. This process is the clinical study design. It generally includes the following.

#### **1.4.1.1 Clarification of Study Aim**

The study's aim is the core basis of the design. It may be the problems encountered in clinical work, the unsolved problems of previous work, the scientific enlightenment, and problems concluded from a literature review, and some clinical problems that need to be solved by superiors. A clear and specific aim is the foundation of all designs.

#### **1.4.1.2 Determination of Study Methods**

In clinical research, an appropriate design method is extremely important according to different study aims and the nature of the clinical research topic. As all research methods have both advantages and disadvantages, appropriate and practical methods should be selected according to different clinical subject areas and goals.

##### **① Observational Study**

In an observational study, the study factors of the groups are predetermined variables, not controlled or influenced by the epidemiologist. Past exposure to risk factors, lifestyle, personal behaviors, environmental factors, immunization status, and genetics all affect the status of health and susceptibility to disease in the groups, and these factors are not influenced by the epidemiologist. That is to say, the investigator in an observational study measures the factor or exposure but does not intervene. Observational studies include cross-sectional study, surveillance study, ecological study, case-control study (retrospective study), and cohort study (prospective or follow-up study).

Observational investigations may be divided into descriptive studies or analytical studies. A descriptive study (cross-sectional study, surveillance study, and ecological study) is the first step in an epidemiological investigation and is used to validate the measurement of the health conditions and health-related characteristics of populations, typically in terms of person, place, and time. This information provides essential contextual information with which to develop hypotheses,



study design, and interpret results and serves as the foundation for studying populations. As a particular type of descriptive study, surveillance can monitor the changes in the occurrence of disease over time. Descriptive approaches are also useful in clinical epidemiology, which includes assessing the performance of diagnostic and screening approaches and clinical decision-making. Analytical epidemiology (case-control study and cohort study) is often used to systematically evaluate the suspected relationships between an exposure and a health outcome, and provide stronger evidence about particular relationships between exposure variables and health status in the general population or in a specific population.

### ② **Experimental Study**

In an experimental study, the subjects are randomly assigned to either an experimental group or a control group. In the experimental group, an active attempt is adopted to change a disease, condition, or death determinant, such as an exposure or behavior, or the progress of a disease through treatment, and then assess the cause-and-effect relationships statistically. The experimental study design includes randomized controlled trials using patients as subjects (clinical trial), field trials, and community trials in which the participants are general people.

An experimental study is useful in establishing a sound cause-and-effect relationship between a factor, intervention, or agent for therapy or prevention of a disease, condition, or death. However, the implementation of experimental studies often involves practical and ethical issues. Analytical epidemiologic studies can offer a realistic approach to testing hypotheses of exposure-disease relationships, and provide more accurate information and useful insights into the effects of diseases or conditions.

### ③ **Mathematical Statistics**

After understanding the occurrence rule of the disease, we can use a mathematical model (theoretical epidemiology) to predict the incidence or prevalence of the disease. Figure 1.1 summarizes the hierarchy of clinical epidemiological study design.

The appropriate design applies not only to attaining study goals but also allows for making effective use of human effort, materials, and time. It reflects the scientific accuracy of the observed results and therefore should also aim to be cost-effective and scientifically rigorous, accurate, and reliable.

#### **1.4.1.3 Identification of the Study Subjects**

The study subjects include population and sample. The research population is the whole group of subjects determined according to the study aim. The sample is a representative part selected from the whole population, which is often used in practical work. This requires sampling randomization, a sufficient number of samples, and clear diagnostic criteria for samples (cases). Three criteria can guarantee the reliability of the research.

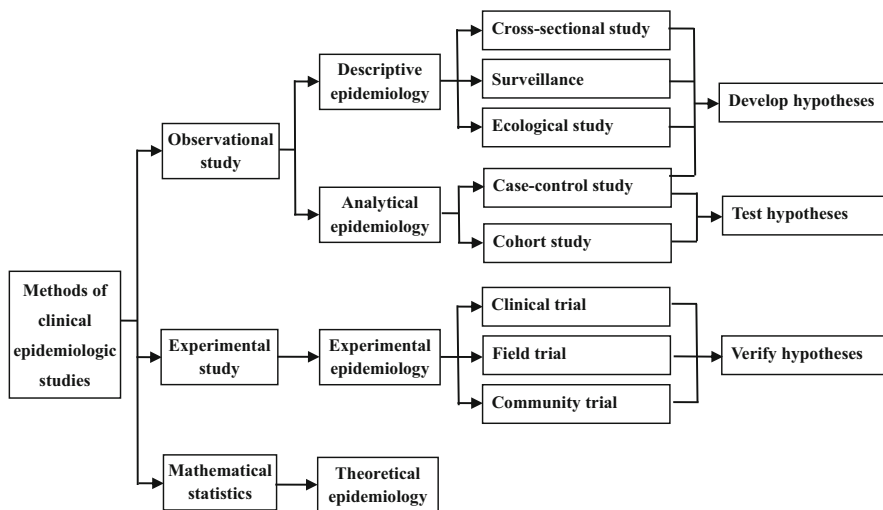


Fig. 1.1 Hierarchy of clinical epidemiological studies design

#### 1.4.1.4 Determination of the Groups

In clinical scientific research, the subjects are divided into experimental and control groups for comparison. Experimental groups can be new diagnostic methods, new drugs, or preventive measures. Comparison is one of the characteristics of clinical epidemiology. The method of grouping can be random or non-random, such as grouping by different times, places, or by certain features. However, randomization allows for an equal chance of selection among all qualified subjects, and guarantees that samples are representative and free of selection bias.

#### 1.4.1.5 Determination of Study Indicators

Study indicators are determined according to study aims. In order to evaluate clinical diagnostic trials, it is necessary to select a recognized index of clinical diagnostic method or equipment as the gold standard. To evaluate the safety and effectiveness of clinical new drugs, clinical indicators, such as effective rate, improvement rate, fatality rate, and incidence of adverse reactions, are used to verify the effect. With the aim of evaluating the effect of vaccine prevention, the protection rate and the rate of change of serum antibody level should be the evaluation criteria. The complication rate, disability rate, and recovery rate should be considered as the indicators of disease prognosis. When measuring these indicators, the measurement methods used and the authenticity and reliability of indicators need to be carefully considered and clearly defined in the research and design stage.

#### **1.4.1.6 Determination Methods for Data Collection and Analysis**

In clinical scientific research, the results usually involve the confounding influence of known and unknown factors that are not the direct object of study. To attain scientifically rigorous results and conclusions, the study design should comply with some basic principles, e.g., control, randomization, and blinding principles.

**Control principle:** The setting of control is a core idea of epidemiology. Proper identification can be derived only through comparison.

**Randomization principle:** Randomization is an important method and a basic principle of clinical studies. Randomization allows for an equal chance of selection among all qualified subjects and guarantees that samples are representative and free of selection bias.

**Blinded principle:** This principle avoids the influence of subjectivity on the part of the investigator and prevents the investigator from knowing the details of grouping and intervention. This practice safeguards the accuracy of the clinical conditions and observed results. Based on the level of understanding regarding the grouping details and relationships amongst the subject, investigator, and data analyzer, this can be divided into single-blinded (only the subject does not know what the intervention is), double-blinded (neither the subject nor the investigator performing the study knows how the intervention is grouped), and triple-blinded trials (neither the subject nor the investigator nor the data analyzer knows how the intervention is grouped).

#### **1.4.1.7 Determination of Study Quality Control Methods**

The common biases in clinical epidemiology include selection bias due to inconsistent diagnostic criteria when patients are enrolled, information bias due to collection of clinical information, and confounding bias due to non-strict randomization of grouping. In addition to the above bias, the inconsistent equipment, different batches of diagnostic reagents, inconsistent information acquisition time (such as blood pressure measurement), inconsistent recognition of diagnostic standards for different doctors, and different adherence of the patient to the doctor's orders, will bring some bias for the results of the study. Therefore, it is very important to adopt quality control methods against the above possible bias at the study design stage.

### ***1.4.2 Measurement***

In clinical studies, the effect of drugs needs to be evaluated through the measurement of certain indices, which can be quantitatively and qualitatively measured. However, high sensitivity and specificity are required for any method. Some can be objectively expressed in terms of specific units or values and are called hard indices or objective

indices, which include factors such as frequency indices (incidence, prevalence, mortality, fatality rate, etc.), effect indices (absolute risk, relative risk, and attributable risk, dose-response relationship, etc.), and objective indices (obtained from objective methods or instruments, such as heart rate, blood pressure, height, weight, morbidity, and death, etc.). Others are difficult to express in terms of specific measurement units and are termed soft indices or subjective indices. These include factors such as pain, quality of life, nausea, dizziness, fatigue, and some psychological determinations. Subjective indices are obtained from the individual impressions of the subject while objective indices are obtained from clinical observation or measurement. Objective indices are not influenced by subjective factors and are therefore more accurate and reliable.

Usually, measurements in clinical research are conducted by different medical staff, so there are many possible influencing factors resulting in bias. For example, some bias due to the inconsistent equipment, different batches of diagnostic reagents, inconsistent recognition of diagnostic standards for different doctors, and variation adherence of the patient to the doctor's orders will bring uncertainty to the results of the study.

### ***1.4.3 Evaluation***

Evaluation is the application of basic theory and methods of epidemiology to assess the accuracy and scientific basis of clinical data and results through statistical analysis. It involves the final selection of valid results and the discarding of false results. The content of the evaluation is mainly reflected in the following aspects.

#### **1.4.3.1 Evaluate the Validity and Reliability of the Study Results**

Clinical epidemiological methods are used to evaluate the design, the accuracy of diagnostic methods, the short-term and long-term efficacy of treatments, the prevention and control measures of bias, the source, representativeness, and compliance of study subjects, etc., to test the validity and reliability.

#### **1.4.3.2 Evaluate the Importance of Study Results**

##### **① Evaluate the Clinical Significance of the Results**

According to the strict evaluation standards of etiology, diagnosis, prevention, treatment, the prognosis of clinical epidemiology and evidence-based medicine, and related indicators of clinical significance, combined with professional and clinical practice, the clinical value of the results is evaluated for improving clinical medical level.

##### **② Evaluate the Statistical Significance of the Results**

If the results of the study have clinical significance, the correct statistical methods must be used to test the significance of the results to evaluate the true degree of clinical differences, i.e., the probability of true positive and true negative results as well as the level and confidence interval (CI) range of test effectiveness, so as to obtain the evaluation of the true degree of clinical study results.

### ③ **Evaluate the Health Economic Implications of the Results**

The results of clinical medical research should evaluate the social and economic benefits (including the cost-effectiveness, cost-benefit, and cost-utility) by applying the principles and methods of health economics, compare, and evaluate to affirm those research results with low cost and good effect so that they can be popularized and applied.

In summary, the main research content and methods of clinical epidemiology are design, measurement, and evaluation. Effective control of various biases should be adopted to ensure the validity and reliability of clinical research results.

## **1.5 Characteristics of Clinical Epidemiology**

As a field of inquiry, clinical epidemiology is a logical discipline that proceeds by way of a sequence of reasoning from empirical data. Clinical epidemiology involves a systematic set of methods and procedures for developing knowledge about health-related events and the relationships between them. It effectively deals with chance, bias, and error. Clinical epidemiology has the following characteristics.

### ***1.5.1 Group***

The research object of clinical medicine is the individual diagnosis, treatment, and outcome, while the object of clinical epidemiology is a group of populations with a particular clinical disease according to the purpose of the study.

### ***1.5.2 Comparison***

Epidemiology is a comparative discipline. It asks questions such as how much disease is in different populations, places, and times and why the disease is distributed this way. It likewise compares the frequency of possible risk factors between groups that have a particular disease (cases) and those without the disease (controls). Key comparative measurements in epidemiology are prevalence, incidence, and risk ratio. Clinical epidemiology also requires the comparison of diagnostic methods or

drugs used in the two groups of cases so as to judge the effectiveness of the diagnostic methods or drugs.

### ***1.5.3 Probability Theory and Mathematical Statistics***

Epidemiology uses frequency, rather than absolute numbers, to describe the distribution of disease because absolute numbers do not show the intensity of disease or the risk of death in the population. Epidemiology, particularly clinical epidemiology, emphasizes that probability or frequency is actually a probability needing the right denominator. In addition, epidemiological work requires a reasonable sample size, and the final sample depends on statistical principles and varies from case to case.

### ***1.5.4 Social Psychology***

Health is closely related to environmental factors. The occurrence of disease is not only related to the health status of the human body but it is also affected by the natural environment and social environment. The biological, psychological, and social conditions of people should be taken into consideration when studying the etiology and risk factors of disease.

### ***1.5.5 Integrating Medicine***

At present, chronic non-communicable diseases (NCDs), such as cardiovascular and cerebrovascular diseases, tumors, diabetes mellitus, and respiratory diseases, are the main health problems affecting most clinical patients. Chronic NCDs not only rank first among the causes of death in the Chinese population but also endanger the health of the labor force and cause the rapid rise of medical expenses. Therefore, controlling the rapid rise of NCDs cannot be achieved by only adhering to the approach of clinical treatment. This requires the integration of prevention and treatment: the integration of residents' health and disease management in a community hospital, with disease management in superior general and specialized hospitals, and the establishment of benign referral; the integration of etiology prevention, early diagnosis and early treatment with active treatment, rehabilitation, reduction or delay of complications, and prevention of disability; the integration of control of behavioral risk factors with clinical drug therapy; the integration of physiological, pathological and psychological therapy; the integration of traditional Chinese medicine with modern Western medicine treats the patient as an organism (to treat both symptoms and root causes) and applies high and new technology to improve the

quality of life. All of these will be reflected in the study of clinical epidemiology and will play an increasingly important role.

### ***1.5.6 Development***

The definition and tasks of epidemiology are developed according to the major health problems during different periods. In recent years, the epidemiological methods have also improved with the development of other disciplines. Traditional epidemiology focuses on the study of three links and two factors of infectious diseases, modern epidemiology focuses on the study of social, psychological, and environmental factors of disease, while clinical epidemiology concentrates on design, measurement, and evaluation. From ecological research of macro epidemiology to molecular biology of micro epidemiology, from RCTs in clinical epidemiology to production, evaluation, and use of medical scientific evidence in evidence-based medical research. All these indicate that development is one of the characteristics of clinical epidemiology.

# Chapter 2

## Distribution of Disease



Shan Zheng

### Key Points

- Frequency measurement is an effective method to quantitatively study the characteristics of disease distribution. Common measures of disease frequency include morbidity, prevalence, and mortality.
- The level of disease is usually expressed by sporadic, outbreak, epidemic, and pandemic, which refers to the change of incidence rate and the association between cases in a certain population during a certain period of time.
- The distribution of disease is used for describing the status of disease in specified populations, areas and time, which will reveal the epidemic characteristics of the disease and its potential risk factors.

By describing the incidence, prevalence, and death, the epidemiological characteristics of diseases may be presented based on the combination of the distribution by people, time, and place (commonly known as the three-dimensional distribution in epidemiology). The study of the distribution of disease is the foundation of finding possible risk factors or understanding etiology, which may have an important contribution in determining the core problems in public health and high-risk groups. Of course, the application of this knowledge would provide scientific evidence for the planning and evaluation of healthcare systems. Briefly, the study of the distribution of disease is not only the starting point and basis of epidemiological research but also an indispensable part of studying epidemic patterns and etiology of diseases. In this chapter, some basic elements, concepts, and tools of epidemiology are discussed: the basics of measurement and comparison, the level of disease, and distribution of disease.

---

S. Zheng (✉)  
School of Public Health, Lanzhou University, Lanzhou, China  
e-mail: [zhengsh@lzu.edu.cn](mailto:zhengsh@lzu.edu.cn)



## 2.1 Measures of Disease Frequency

### 2.1.1 Frequency Measures

#### 2.1.1.1 Ratio

The value obtained by dividing one quantity by another. In a ratio, the values of  $x$  and  $y$  may be completely independent, or  $x$  may be included in  $y$ .

$$\text{Ratio} = \frac{x}{y} \text{ (} x \text{ is completely independent of } y \text{ or } x \text{ is part of } y \text{)} \quad (2.1)$$

#### 2.1.1.2 Proportion

A type of ratio in which the numerator is included in the denominator.

$$\text{Proportion} = \frac{x}{y} \text{ (} x \text{ is part of } y \text{)} \quad (2.2)$$

#### 2.1.1.3 Rate

A measure of the frequency of occurrence of a phenomenon. In epidemiology, a rate is an expression of the frequency with which an event occurs in a defined population, usually in a specified period.

$$\text{Rate} = \frac{\text{Number of cases or events occurring during a given time period}}{\text{Population at risk during the same time period}} \times K \quad (2.3)$$

### 2.1.2 Morbidity Frequency Measures

#### 2.1.2.1 Incidence Rate

The incidence rate is the number of new cases per population at risk in a given time period. The numerator is the number of new events in a defined period or other physical spans. The denominator is the population at risk of experiencing the event during this period. The calculating formula for incidence rate follows:

$$\text{Incidence rate} = \frac{\text{Number of new cases in specified period}}{\text{Average number of person exposed to risk during this period}} \times K$$

$K = 100\%, 1000\text{‰}, 10,000 \text{ per } 10,000 \dots\dots$

(2.4)

To calculate incidence rate, a year is usually chosen as a study period. You can also define the study period by the characteristics of diseases or events.

The numerator of an incidence rate should be the new cases of disease which occurred or were first diagnosed during the observation period. Those cases which occurred or were diagnosed earlier should not be included in the numerator. If a person has multiple episodes of illness during the observation period, they should be all counted as new cases, such as influenza and diarrhea, which can occur more than once in a year.

Notice that the denominator is the population at risk, which means that the denominator is the number of people exposed and at risk for the observed disease in the population of an area during the observation period. Persons who are already ill and are not likely to become new cases during the observation period should not be included in the exposed population. For example, in calculating the incidence of measles, people who already have measles cannot be included in the denominator. In addition, theoretically, people who have received the measles vaccine and gained immunity should not be included in the denominator, but it is not easy to divide in practice. The denominator is usually the average population of the area during the observation period.

Incidence rates are useful in the study of disease etiology because they are informative about the risk of a disease process in different population groups. By comparing the difference in incidence rates, some possible or potential causation could be found and proposed. It is usually used to measure the risk of acute disease or conditions but is also used for chronic diseases.

### 2.1.2.2 Attack Rate

An attack rate is a variant of an incidence rate applied to an outbreak of disease among a narrowly defined population during a short period of time. It is calculated by the same formula as incidence rate, but it is observed over a shorter period of time. The attack rate is often used in outbreaks and epidemics of food poisoning, occupational poisoning, or infectious diseases.

Attack rate

$$= \frac{\text{Number of new cases among the population during the period}}{\text{Population at risk at the beginning of the period}} \times 100\% \quad (2.5)$$

**2.1.2.3 Secondary Attack Rate**

A secondary attack rate is a measure of the frequency of new cases of a disease among the contacts of known cases. The formula is as follows:

$$\begin{aligned} &\text{Secondary attack rate} \\ &= \frac{\text{Number of cases among contacts of primary cases during the period}}{\text{Total number of contacts}} \times 100\% \end{aligned} \tag{2.6}$$

This index is always used to measure the contagiousity of infectious diseases and to evaluate the effectiveness of prevention measures. It is worth noting that the cases occurring without the incubation period following exposure to a primary case are generally not included in the numerator and denominator.

**2.1.2.4 Prevalence**

Prevalence is a proportion of persons in a population who have a particular disease (including new cases and pre-existing cases) at a specified point in time (point prevalence) or over a specified period (period prevalence). The formula for prevalence is:

$$\begin{aligned} \text{Prevalence} &= \frac{\text{New cases and pre-existing cases during a given time period}}{\text{Population during the same time period}} \times K \\ K &= 100\%, 1000\%, 10,000 \text{ per } 10,000 \dots \dots \end{aligned} \tag{2.7}$$

The prevalence of a disease in the population is influenced by many factors; the common factors are shown in Table 2.1. Among these factors, incidence and duration of the disease play the most significant effect on prevalence. When the incidence and duration of a disease in a certain place are stable for a long time, the relationship between prevalence, incidence, and duration of the disease is shown as: Prevalence = incidence × duration of disease.

**Table 2.1** Factors affecting prevalence

Increased by	Decreased by
Longer duration of the disease	Shorter duration of the disease
Prolongation of life of patients without a cure	Increased case-fatality rate
Increase in new cases (increase in incidence)	Decrease in new cases (decrease in incidence)
In-migration of cases	In-migration of healthy people
Out-migration of healthy people	Out-migration of cases
In-migration of susceptible people	Improved cure rate of cases
Improved diagnostic level	—
Better reporting rate	

The prevalence is usually used to describe the epidemiological characteristics of diseases, especially for the chronic diseases with a longer duration. It can be used to estimate the impact of the diseases to human health and to evaluate medical and health work and health resource allocation conditions, such as hospital beds turnover rate, sanitation, and the demand and supply of human resources, etc.

### 2.1.3 Mortality Frequency Measures

#### 2.1.3.1 Mortality Rate

A mortality rate represents the proportion of total deaths in a population in a given period of time. The numerator is the number of persons dying during the period; the denominator is the average population during the same period. The annual death rate or mortality rate from all causes in a population is generally calculated by the following formula:

$$\begin{aligned} & \text{Annual mortality rate for all causes} \\ &= \frac{\text{Number of deaths from all causes in 1 year}}{\text{Average population in the same period}} \times K \quad (2.8) \\ & K = 100\%, 1000\%, 10,000 \text{ per } 10,000 \dots \end{aligned}$$

Generally, a mortality rate for all causes is also called the crude death rate. When comparing mortality rates in different regions, mortality rates need to be standardized. Similarly, comparisons of prevalence or incidence rates across regions need to be standardized.

In addition, the crude death rate is an unadjusted mortality rate which shows all causes of death for a population. When we calculate the mortality rate by a specific disease, age, gender, race, and so on, this mortality rate will be named as cause-specific mortality rate. For example, an age-specific mortality rate is defined as:

$$\frac{\text{Total number of deaths occurring in a specific age group of the population in a defined area during a specified period}}{\text{Estimated total population of the same age group of the population in the same area during the same period}} \times K \quad (2.9)$$

Mortality rate can be an indicator reflecting the total death level of a population, which is usually used to measure the death risk of a population in a certain period and a certain area. The specific mortality rate can provide information on the variation of the mortality rate of a disease in population, time, and region and can be used to explore the etiology and evaluate the prevention and treatment measures.

### 2.1.3.2 Case Fatality Rate

Case fatality rate is used to measure the severity of disease and is defined as the proportion of cases with a specified disease or condition who die within a specified time.

$$\text{Case fatality rate} = \frac{\text{Number of deaths from diagnosed cases in a specified time}}{\text{Number of diagnosed cases of the disease in the same period}} \times 100\% \quad (2.10)$$

### 2.1.3.3 Survival Rate

The proportion of patients who received treatment or those with a certain disease who were still alive after following up for several years.

$$\text{Survival rate} = \frac{\text{Number of cases who were still alive after following up } N \text{ years}}{\text{Number of all cases after following up } N \text{ years}} \times 100\% \quad (2.11)$$

Where  $N$  is always equal to 1, 3, 5, and 10 years.

Survival rates are used to evaluate the severity and long-term outcomes of chronic diseases such as tumors, cardiovascular diseases, and chronic infectious diseases (like tuberculosis and AIDS).

## 2.2 Epidemic Disease Occurrence

The level of disease is defined as the variations in the trend of incidence of a particular disease and the association between the cases in a given population during a certain period. Terms of the level of disease include sporadic, epidemic, and outbreak.

### 2.2.1 Sporadic

Sporadic means the incidence of a particular disease is in the range of the average incidence of the disease over the past 3 years in that population in that area, which refers to a disease that occurs infrequently and irregularly. For example, plague occurs by chance in a grazing area.

The cause of sporadic occurrence includes four aspects.

- ① The level of immunity in a population has increased as a large portion of the population in the area has been vaccinated, or people have been infected with the disease before.
- ② The disease is an inapparent infectious disease, such as poliomyelitis and Japanese encephalitis.
- ③ The transmission mechanism of the disease is not easy to achieve.
- ④ Diseases with a long incubation period, such as leprosy.

### **2.2.2 Epidemic**

Epidemic means the incidence of a disease is above the average incidence of the disease over the past several years in that population in that area. When a disease shows epidemic, it means there is an obvious association between all cases in time and space. For example, in 2009, the epidemic of influenza A (H1N1) showed an obvious characteristic, including person-to-person transmission and dissemination in different areas. In addition, when an epidemic occurs on a scale which crosses international boundaries, it is called a pandemic. For example, the 1918–1919 influenza pandemic.

### **2.2.3 Outbreak**

As a group of people are all exposed to an infectious agent or a toxin from the same source, an increase in the number of cases of a disease appear suddenly. The epidemic is always limited to a localized area, e.g., in a village, town, or closed institution.

## **2.3 Distribution of Disease by Time, Place, and Person**

The distribution of disease is used for describing the status of disease by time, place, and person, which will reveal the epidemic characteristics of the disease and potential risk factors and even propose the hypothesis for the pathogenesis of some disease process. Distribution of disease is the core of descriptive study and the basis of analytical study.

### 2.3.1 Time

Whether it is an infectious disease or a non-communicable disease, their epidemic characteristics will change by time lapse. Analyzing the time variation in disease risk is used to provide the basis for studying the cause of disease and drawing up prevention and control measures. There are four forms of time variation used to describe the time distribution of disease, including rapid fluctuation, seasonality, periodicity, and secular trend.

#### 2.3.1.1 Rapid Fluctuation

This refers to a phenomenon where the number of cases of a disease increases suddenly in a certain institution or a fixed population. It is similar to an outbreak, but an outbreak is suitable for a small population, while rapid fluctuations apply to large population.

The causes of rapid fluctuation are generally clear, and the main cause is that a number of people are exposed to the same pathogenic factors at the same time or continually contacted, such as food poisoning in collective canteens or the outbreak or epidemic of dysentery and measles. Besides, rapid fluctuation of a disease might result from natural disasters, environmental pollution, or social politics.

Figure 2.1 shows the time distribution of viral meningitis outbreak in a school as the pollution of drinking water. The first case occurred on June 26, and the number of cases increased quickly on June 30 and decreased on July 2. The epidemic stopped on July 18 and lasted 23 days.

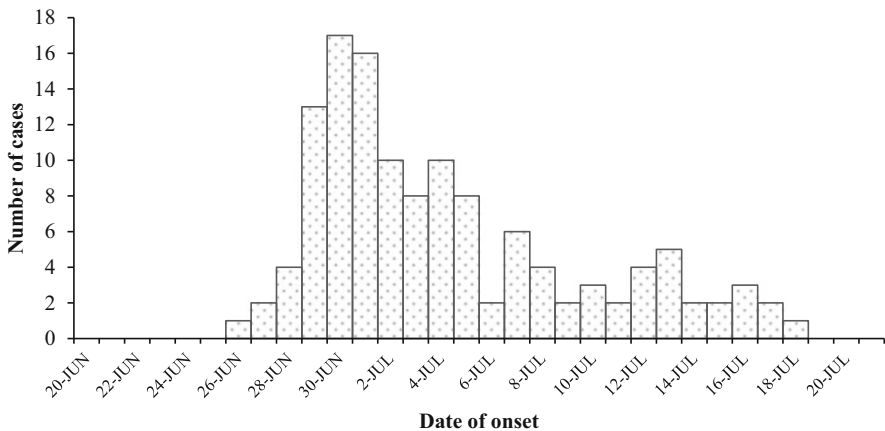


Fig. 2.1 Time distribution of viral meningitis outbreak in a school

### 2.3.1.2 Seasonality

Seasonality refers to a phenomenon that the incidence or mortality of a disease increases in a certain season, which includes two different characteristics.

#### Strict Seasonality

The new cases of some infectious diseases only occur in some specific months of the year; however, in other months, no cases are occurring. The epidemic of malaria and encephalitis B is usually seen in summer and autumn, which is related to the life cycle of the vector that causes disease transmission.

#### Seasonal Rise

Some diseases occur all year round, but the incidence increases only in certain month (s). For example, enteric infectious diseases and respiratory infectious diseases occur all year, but the incidence of enteric infectious diseases is most common during the summer and autumn months, while the incidence of respiratory infectious diseases always rises in spring and winter. Moreover, some non-communicable diseases also show a characteristic seasonal rise, such as pollinosis occurs at the end of spring and the beginning of summer; the incidence of stroke always increases in winter.

### 2.3.1.3 Periodicity

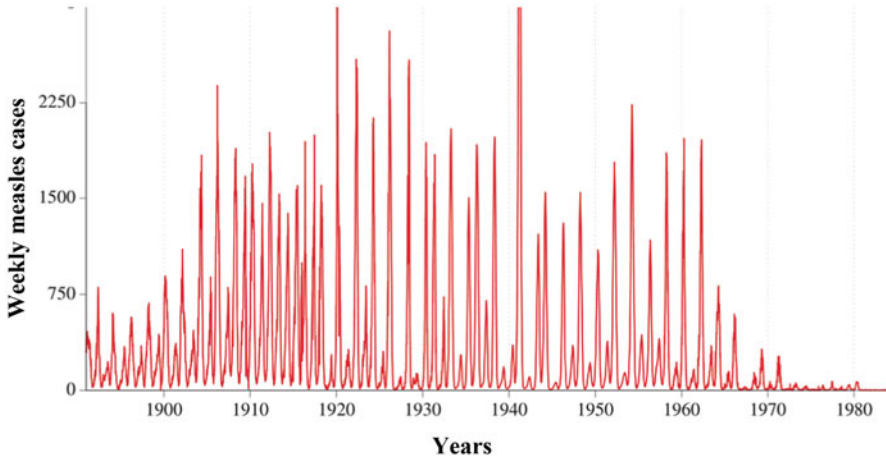
Periodicity refers to a phenomenon where the frequency of a disease increases in a certain interval. Some respiratory infectious diseases often have regular periodic epidemics. As shown in Fig. 2.2, the weekly measles cases in New York City showed a clear periodicity from 1891 to 1963. There were two peaks during this period, on the weekend of 31 January 1920 (4671 cases) and 22 March 1941 (5016 cases). But with the invention of the first measles vaccine in 1963, measles vaccines have been used widely to immunize susceptible population against measles. So the number of measles cases in this region declined rapidly.

The reasons for periodic epidemics of disease are associated with immunization level of the population, duration of immunization, accumulation of susceptible population, variation of pathogens, etc.

### 2.3.1.4 Secular Trend

Secular trend (Synonyms: temporal trend or long-term trend) denotes that the annual cases or rate of a disease over a long period, generally years or decades, shows





**Fig. 2.2** Trends in measles cases in New York City from 1891 to 1984. Reprinted with permission from Hempel, Karsten; Earn, David J D. A century of transitions in New York City's measles dynamics. *J R Soc Interface*. 2015 May 6;12(106):20150024. doi: 10.1098/rsif.2015.0024

long-term or secular trends in the occurrence of the disease. The probable causes of secular trends include changes in etiology or pathogenic factors, changes in pathogens, improvements in medical treatment and prevention, and changes in social policies. Therefore, these trends are commonly used to suggest or predict the future incidence of a disease, to evaluate programs or policy decisions, and to reveal some potential causes in the occurrence of a disease.

Figure 2.3 shows mortality rates for six diseases in men from 2000 to 2016. According to the WHO, trachea, bronchus, and lung cancer showed a significant decline during the 16-year period, the death rate (Age-standardized) was 33.41 per 100,000 population in 2016 with a  $-38.4\%$  of change relative to 2000. The reason for this downward trend may be related to the decline in smoking rates among men, and the implementation of smoking-related policies such as banning smoking in public areas and raising tobacco taxes may also influence the change in the trend. In addition, we also see slight increases in diabetes mellitus and pancreatic cancer, which may also be related to changes in the lifestyle of the population.

### 2.3.2 Place

The occurrence of a disease is associated with natural and social environment, so analyzing data by place can provide an important clue for revealing the cause of a disease and also provide a basis for proposing prevention measures. The variation of pathogenic factors in different places can result in the variation of the disease distribution. Some natural environmental factors, e.g., special geographical location, topography and geomorphology, meteorological conditions, and some social

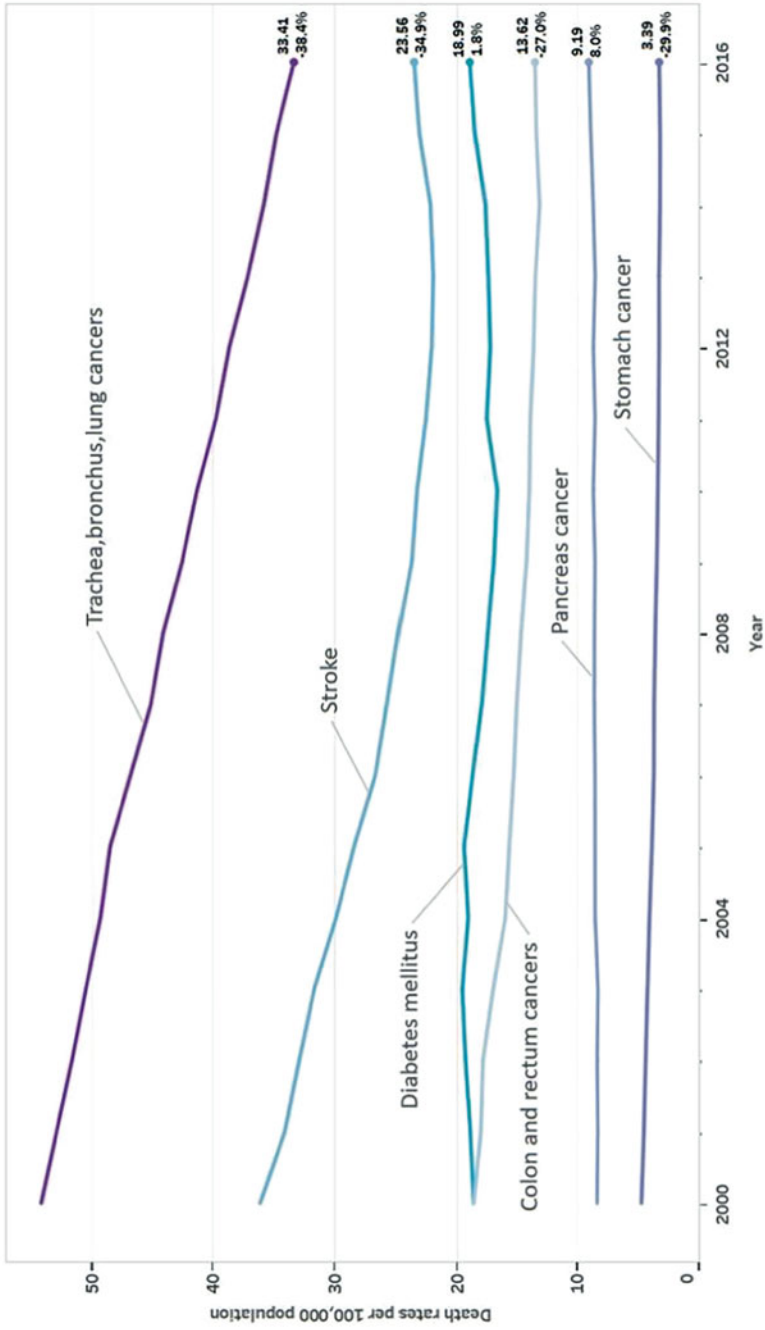


Fig. 2.3 Trends in age-adjusted death rates from six diseases for males, United States, 2000–2016

environmental factors, e.g., living habits and sociocultural background, can play a significant role on disease distribution by place.

To study the distribution by place, there are two methods used to analyze the data which are political boundaries and natural boundaries. The method of political boundaries is used to divide places by continent, state, or region at the global level and by province, city, or urban and rural in a country. This method is easy and convenient, but it generally does not match the distribution of natural environmental factors, which may cover up the ecological relationship between natural environmental factors and the distribution of a disease. Another method used to analyze data is natural boundaries, which is used to divide places by mountains, plains, lakes, rivers, or grasslands, and so on. This method is better at revealing the correlation between natural environmental factors and disease, but it is difficult to collect the data and carry it out.

### **2.3.2.1 Comparisons Among and Within Countries**

#### The Distribution of a Disease in Different Countries

Some diseases are spread all over the world, but the distribution is not uniform, and the morbidity or mortality rate of these diseases may vary greatly. Both infectious diseases and non-communicable diseases present various distributions between countries. For example, yellow fever is only an epidemic in South America and Africa. Cholera is common in India. The mortality rate of stomach cancer is higher in countries such as Japan and Chile, while it is lower in Australia and the United States. Liver cancer is common in Asia and Africa, but breast cancer and colon cancer are common in Europe and North America.

According to the Global cancer statistics (2012) [1], prostate cancer is the second most frequently diagnosed cancer in men worldwide. It is clear that the most frequently diagnosed cancer among men was in more developed countries. Incidence rates vary by more than 25-fold worldwide and are higher in Australia/New Zealand, Northern America, Northern, and Western Europe, and some Caribbean nations, and lower in Asia.

#### The Distribution of a Disease in Different Areas within a Country

The incidence or mortality rates of diseases also vary in different areas within a country. For example, schistosomiasis only occurs in some provinces, south of the Yangtze River in China, which is associated with the distribution of oncomelania in the surrounding environment. Nasopharyngeal cancer is most common in Guangdong Province, China.

Based on the third national retrospective sampling survey of death causes in China from 2004 to 2005, the highest breast cancer mortality rates were found in Shanghai (5.21 per 100,000) and Heilongjiang Province (5.69 per 100,000), while

the mortality rates in Liaoning, Jilin, Shandong, Guangxi, and Hunan (ranging from 4.53 to 4.84 per 100,000) were higher than most provinces.

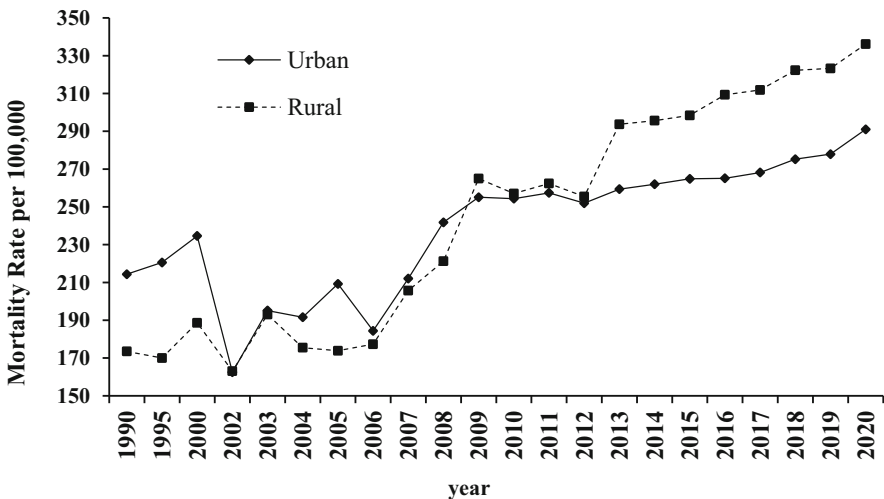
### 2.3.2.2 Urban-Rural Comparisons

The variation of population density, natural environment, health level, living conditions, economic level, and population mobility shows an obvious difference in kinds of the disease, cause of death, incidence, and fatality of the disease between urban and rural areas.

In urban areas, respiratory infections, such as chicken pox, mumps, and influenza, are mostly common due to higher population density, small living space, traffic congestion, and large mobility. In addition, more serious environmental pollution in urban areas can increase the incidence and mortality rates of some diseases related to air pollution.

In rural areas, intestinal infections and vector-borne infectious diseases are mostly common due to poor sanitary conditions and local natural environment. Respiratory infectious diseases are difficult to spread because of lower population density, but it can also cause an outbreak once the infection sources enter.

According to the China Health Statistical Yearbook (2003, 2021) [2, 3], cardiovascular disease mortality rates among urban and rural residents in China increased from 1990 to 2020 (Fig. 2.4). While the mortality rate increased rapidly in rural areas, it ultimately surpassed the rate in urban areas by 2009. In recent years, differences in the distribution of disease between urban and rural areas are narrowing and getting bigger after 2013 due to economic development and urbanization in China.



**Fig. 2.4** Trends in cardiovascular diseases mortality among urban and rural residents in China, 1990–2020

### 2.3.2.3 Endemic Clustering

Endemic clustering refers to the phenomenon of the morbidity or mortality of disease in a locality being significantly higher than in the surrounding areas. A disease with the characteristic of endemic clustering suggests that there are some specific pathogenic factors which have an adverse effect on human health in a certain place. The term is normally used to discuss the cause of disease and evaluate the intervention effects of some prevention measures. Ebola hemorrhagic fever has been found in the Ebola River of southern Sudan and the Congo since 1976. At present, the disease presents endemic clustering, which is confined to the central African rainforest and the tropical savannah of southeast Africa. In February 2014, an Ebola hemorrhagic fever broke out in West Africa and gradually spread to areas such as Sierra Leone and Liberia. The process presents obvious regional aggregation.

### 2.3.2.4 Endemic Disease

Endemic disease presents that some diseases are confined to a certain area due to the influence of natural environmental and social factors.

The standards for determining whether a disease is endemic include several aspects.

- ① The incidence of the disease is high among the residents of the area.
- ② The incidence of the disease among similar populations in other regions is low and even nonpathogenic.
- ③ After immigrating to the area, the incidence of the disease in immigrants is consistent with that in the local population.
- ④ After the population has moved out of the area, the incidence is reduced or the symptoms of the disease are reduced or self-healing.
- ⑤ Besides people, local susceptible animals have the same disease.

### 2.3.3 Person

The differences in morbidity and mortality are more influenced by personal characteristics (age, sex, ethnic and racial group, occupation, behavior, etc.), and these characteristic variables or their variation can be considered as the foremost risk factors for many diseases. Therefore, the study of these demographic characteristics is of great significance to explore the risk factors or epidemic characteristics of disease or health status, which can provide clues to confirm high-risk population and many hypotheses, importantly, provide evidence for policy-making.

### 2.3.3.1 Age

Age is one of the most important demographic variables. The occurrence and development of almost all diseases are associated with age. The age-specific patterns also differ between and within diseases. Generally, the incidence of chronic diseases increases with age, while the incidence of acute infectious diseases decreases. However, some chronic diseases present a peak at younger ages as the change of pathogenic factors, such as malignant tumor, diabetes, high blood pressure, etc. Sometimes, the incidence and prevalence of diseases are not the same in the different age groups. The incidence of acute lymphocytic leukemia is higher in children, and the Hodgkin's disease has two peaks at youth and old age. The incidence of gastric cancer shows a rising trend with age. The effects of age are most commonly ascribed to an individual's cumulative exposure to environmental factors over a lifetime or to the decline in immunological defenses.

There are two ways to analyze the association between diseases and age. One is cross-sectional analysis, and another is birth cohort analysis.

1. Cross-sectional analysis is used to examine the age-specific disease rates of a population at a particular period, which would include individuals born at different time periods. For example, the age-specific incidence rates of some acute infectious diseases are presented in a cross-sectional age curve at a particular period for different generations. The potential problem associated with a cross-sectional analysis is that individuals in different age groups at a particular calendar time were born in different years. Therefore, these individuals belong to different generations or birth cohorts. Different generations always have different disease risks due to different types or levels of exposure to disease risk factors. Thus, the cross-sectional analysis, especially for chronic disease, may or may not validly explain the association between disease and age.
2. Birth cohort analysis is based on age-specific incidence rates for a particular birth cohort observed at successive points in calendar time as the cohort grows older, which includes all persons born within a specified period. People from a generation or a birth cohort would carry, even throughout their lives, a relatively higher or lower risk for certain diseases compared to other birth cohorts. If disease risk changes with successive birth cohorts due to the increase or decrease in exposure to risk factors, the age-birth cohort curves would be progressively higher or lower than the previous birth cohorts. Thus, birth cohort analysis can reflect a real change in disease rate with age through examining the disease rates from people born in the same or different time periods as age increases, if disease risk indeed varied by birth cohorts.

Figure 2.5 shows the relationship between age and lung cancer occurrence in a city for 5-year periods (1995–1999, 2000–2004, 2005–2009, 2010–2014, and 2015–2019) in the cross-sectional age curves, a classic example of examining age effect on disease. According to these curves, the incidences of lung cancer at first increase with age, then the incidences are at peak in the 65–69 age groups and

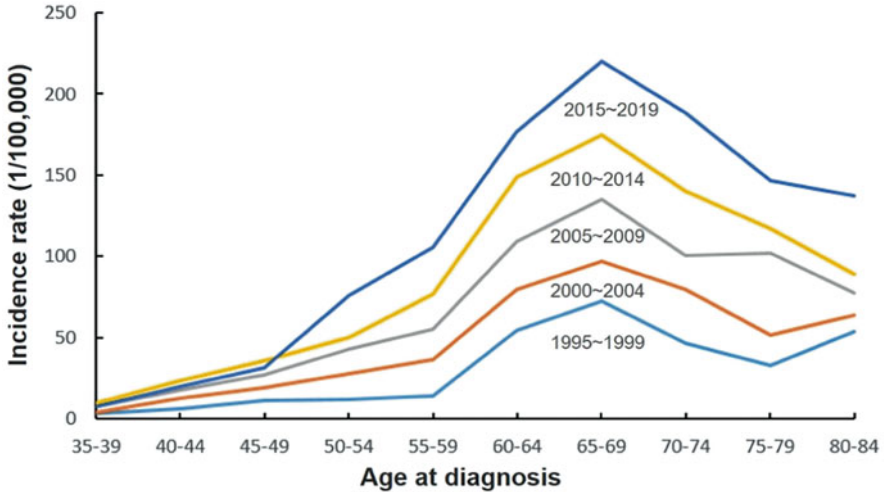


Fig. 2.5 Lung cancer incidence rates in a city based on cross-sectional age curves

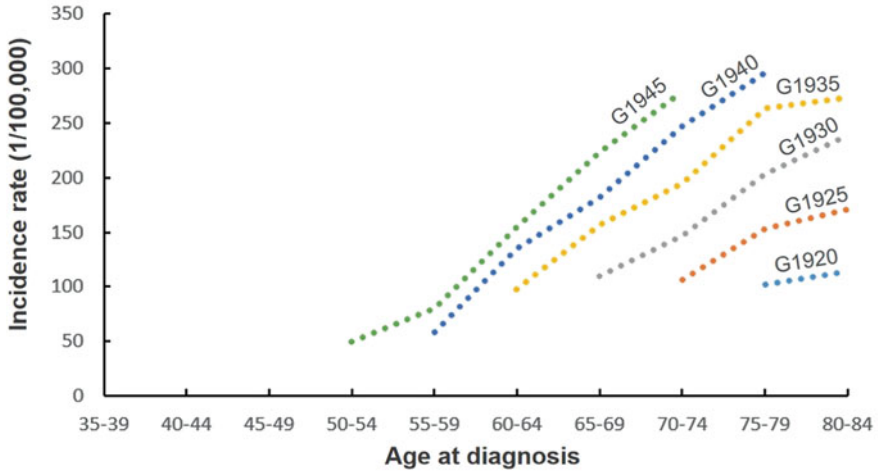
subsequently decline as age advances. Meanwhile, the incidences of lung cancer increased significantly during the 5-year survey periods; the interpretation of this situation may be that the risk for lung cancer continues to increase. The question to be considered is why the lung cancer incidence rates in the city show a decline after ages 65–69 in the cross-sectional age curves. Actually, there seems to be no biological basis for assuming that age-related changes result in a decrease in lung cancer risk in the city based on the present etiology of lung cancer.

However, in Fig. 2.6, when the same lung cancer age-specific incidence rates are presented in six generations (G1920, G1925, G1930, G1935, G1940, and G1945), the rates generally showed a continuous increase with increasing age for each of the six birth cohorts for the given age ranges. The incidence rates among people born in latter generations at the same ages are higher than that in those people born in previous generations and tend to be younger.

### 2.3.3.2 Gender

There are, for certain diseases, substantial differences in the rates between men and women (sex ratio). For example, incidence and mortality rates from most malignant neoplasms are substantially higher in men than in women.

The reason for the substantial differentials in gender may represent the difference in endogenous factors, such as the sex hormones, physiological and anatomical characteristics. The disparity may act as a promoter of disease pathogenesis or as a protective factor. Thus, cervical cancer occurs in women, and prostate cancer occurs in men. Besides endogenous factors, differences in exposure to environmental risk factors may contribute to differences in risk between male and female. Hence, the



**Fig. 2.6** Lung cancer incidence rates in a city based on age-birth cohort curves

high male-female incidence disparity for leptospirosis and schistosomiasis is largely a consequence of exposure to contaminated water in men through certain occupation, while much of the occupational poisoning and accidental fall in men relative to women can be explained by contacting more exposure in some hazardous professions. Additionally, differences in social, cultural, and behavioral habits between men and women are also possible causes of the differences in disease distribution. Lung cancer is largely associated with tobacco smoking in men relative to women.

**2.3.3.3 Ethnic and Racial Group**

Disparities in some diseases among racial and ethnic minorities largely reflect that race and/or ethnicity is another important risk factor affecting the distribution of disease in the population. Interpretation of ethnic differences in disease risk may consider these factors, such as genetic background, geographical environment, religious beliefs, customs, way of life, etc.

According to the data from American Cancer Society [4], Non-Hispanic blacks have the highest incidence and mortality rates of cancer, while Asian and Pacific Islanders have the lowest rates in both genders compared to other ethnic groups (Table 2.2).

**2.3.3.4 Occupation**

There is a close relationship between occupation and diseases. The distribution and severity of diseases are different in many occupation categories due to the different



**Table 2.2** Incidence and mortality rates for cancers by race and ethnicity, US, 2009–2013 (per 100,000)

Race and ethnicity	Male		Female	
	Incidence	Mortality	Incidence	Mortality
Non-Hispanic White	519.3	204.0	436.0	145.5
Non-Hispanic Black	577.3	253.4	408.5	165.9
Asian and Pacific Islander	310.2	122.7	287.1	88.8
American Indian and Alaska Native	426.7	183.6	387.3	129.1
Hispanic/Latino	398.1	142.5	329.6	97.7

Data from the American Cancer Society. Cancer Facts & Figures 2017 [4]

exposure profiles. For example, benzene exposure in the shoemaking workers increases susceptibility to leukemia; coal miners are predisposed to pneumoconiosis, workers with fur processing and animal husbandry are susceptible to anthrax. The cause of these differences in the distribution of disease is the differences in the chance of exposure and/or pathogenic factors. Occupational factors, such as labor conditions, physical labor intensity, mental stress, and social and economic status, also contribute to differences in risk between occupations and diseases.

### 2.3.3.5 Marital Status

A number of epidemiological studies have analyzed the relationship between marital status and disease risk. It is apparent that differences in marital status have an impact on the risk of disease. Research confirms that married people tend to have lower mortality rates than single people, and divorced people have the highest mortality rates. Meanwhile, the incidences of suicide and violence tend to be higher among divorced persons, which indicates that divorce events may cause a lot of adverse effects on mental and psychological aspects, resulting in diseases. Moreover, cousin marriage may contribute to the increased incidence of congenital malformation and hereditary diseases, seriously influencing population quality. The main difference in determining whether persons in marriage get health advantages or if there are certain characteristics of good health or long life might be in favor of an individual's predisposition to marriage.

### 2.3.3.6 Behavior and Lifestyles

Many diseases are associated with bad behavior and lifestyles. Studies have shown that about 60–70% of all chronic diseases, such as malignant tumor, cardiovascular disease, and diabetes, were caused by unhealthy behavior and lifestyles. Common adverse behaviors include smoking, alcoholism, drug addiction, lack of physical activity, unsafe sex, and mental stimulation. Previous studies have confirmed that smoking can cause lung cancer, oral cancer, pharyngeal cancer, laryngeal cancer,

esophageal cancer, and bladder cancer, a significant dose-response relationship between cancer risk and smoking was also identified. Intravenous drug use, homosexual behavior, and unsafe sexual behavior may increase the spread of sexually transmitted diseases. Thus, researches based on the distribution of behavior and lifestyles are helpful to explore the etiology and formulate the strategies for prevention and control.

### **2.3.4 Combinations**

In epidemiologic studies, there is a need to comprehensively study the disease distribution of time, place, and person to know the distribution characteristics and potential risk factors and make relative control measures.

Migrant epidemiology provides useful insights into the relative importance of environmental exposures and genetic factors in the etiology of disease. Migrant refers to a group of people who have moved from their former residence to another place of residence and have permanently changed their place of residence. Migrant epidemiology is a method of etiological research that analyzes the causes of disease by comparing the morbidity of migrants with the population of their country of origin or migrants with the population of their new host country where they have settled. Studies of disease risk in migrants may provide useful information on whether the disease is mainly caused by genetic or environmental factors. The principles of migrant studies are shown in the following two aspects.

- ① If the disease is mainly caused by environmental factors, the incidence or mortality rate of the disease in migrants is different from that of the population from which they originated but is close to those of the new host country in which they have settled.
- ② If the disease is mainly caused by genetic factors, the incidence or mortality rate of the disease in migrants is similar to that of the population from which they originated but is different from those of the new host country in which they have settled.

Generally, the risk of disease among migrants may be influenced by the changing of the living environment and conditions. Meanwhile, the related factors, such as medical and health care, the population characteristics including age, gender, educational level, socio-economic status, religious belief, the length of time for migrants in the new country of residence, and the number of generations, should be considered in migrant epidemiology.

# Chapter 3

## Descriptive Study



Zhenxing Mao and Wenqian Huo

### Key Points

- Descriptive studies are mainly used by observing, collecting, and analyzing relevant data to describe the distribution of disease, health status, and exposure and generate hypotheses for further investigations.
- Descriptive studies mainly include case and case series report, cross-sectional studies, and ecological studies.
- The ability of descriptive studies to prove whether it is a causal association or coincidental phenomenon between exposure and outcome is limited.
- Selection bias, information bias, confounding bias are three major sources of bias in cross-sectional study. Ecological fallacy and confounding factors are the main limitations in ecological study.

Descriptive study, also known as descriptive epidemiology, is the most basic type of epidemiological research method. Descriptive studies are mainly used for describing the distribution of disease, health status, and exposure and generating hypotheses for further investigations but cannot tell causal relations between disease and exposure. Descriptive studies are also mainly used for ascertaining high-risk individuals and evaluating the effects of public health measures, etc. Descriptive studies mainly include case and case series reports, cross-sectional studies, and ecological studies.

---

Z. Mao (✉) · W. Huo (✉)  
College of Public Health, Zhengzhou University, Zhengzhou, China  
e-mail: [huowenqian@zzu.edu.cn](mailto:huowenqian@zzu.edu.cn)

## **3.1 Introduction**

### ***3.1.1 Concept***

Descriptive study is a research method that describes the distribution of diseases or health status and their influencing factors at different times, regions, and populations without changing the current disease status and exposure characteristics of the subjects.

### ***3.1.2 Characteristics of Descriptive Studies***

1. Descriptive studies take observation as the main research method and do not impose any intervention measures on research subjects. Only by observing, collecting, and analyzing relevant data do descriptive studies analyze and summarize the distribution of diseases, health conditions, relevant characteristics, and exposure factors.
2. Descriptive studies generally do not set up a control group. And the ability to prove whether it is a causal association or coincidental phenomenon between exposure and outcome is limited. However, it could provide a preliminary contribution to subsequent studies.
3. Descriptive studies have a shorter duration. The distribution of disease and health status in a population is typically analyzed for transient or temporal characteristics. However, it is easy to implement. The distribution of disease and risk factor distribution can be obtained in a relatively short time.

### ***3.1.3 Application***

1. To describe the prevalence of disease in different regions and different population characteristics. Continuous descriptive studies at different intervals can also provide time trend data of disease.
2. To describe the regional, population, and temporal distribution of risk factors.
3. To provide etiological clues and form a preliminary etiological hypothesis.
4. Through descriptive research, patients at early or different stages can be found and accept early treatment. At the same time, patients with different disease stages and different infection patterns in the population can also be found. So it can be used to study the natural history of diseases.
5. To provide baseline data as the basis for the longitudinal study.
6. Descriptive studies can evaluate the effectiveness of preventive and control measures in the same population before and after the implementation of interventions.

## **3.2 Case and Case Series Report**

### **3.2.1 *Concept***

Case reports usually study a newly discovered or specific disease and its characteristics. A complete case report includes the patient's epidemiological data, such as pre-onset lifestyle characteristics and history of exposure to suspected risk factors. In the case of infectious diseases, attention should also be paid to investigating and reporting the possible exposure to patients, animals, and the environment before and after the onset of illness.

Case series reports are conducted on the basis of case reports used for describing a series of clinical features or cases with similar diagnoses. The content of the case series report is exactly the same as that of the case report mentioned above. However, it should be emphasized that the case series reports should pay more attention to the demographic characteristics of each case, especially the similarity of risk factor exposures and clinical characteristics. Focusing on the chronological sequence of cases and their interconnections is more conducive to forming etiological hypotheses. Generally, case series reports often provide evidence better than case reports.

Hospitals are important places to detect the potential new and special cases; case reports and case series reports are usually carried out by clinicians. Only those with systematic epidemiological training and keen insight can catch abnormal cases in daily diagnosis and treatment activities and report to the local CDC in time. Measures are also taken to prevent further spread of the disease. Both approaches are applicable to infectious and chronic non-communicable diseases.

### **3.2.2 *Application***

#### **3.2.2.1 Identifying New Diseases**

When a new disease occurs, it is necessary to describe the clinical, demographic, and lifestyle characteristics of the patients, behavioral risk factors, the characteristics of working and living environment in detail. Then, we explore the possible causes for diagnosis and make prevention. On the basis of the above research contents, clinicians can also evaluate treatment measures and effects and expand the research contents.

#### **3.2.2.2 Establishing the Diagnosis**

Based on the clinical symptoms, signs, and laboratory examination results of the patients provided by the case reports and case series reports, by combining with the patient's demographic characteristics and epidemiological data (lifestyle

characteristics, targeted risk factor exposure history, time and place of onset, etc.), summarizing the common clinical and epidemiological characteristics of patients, diagnostic criteria can be established for the identification and diagnosis of subsequent similar diseases.

### **3.2.2.3 Forming an Etiological Hypothesis**

From the characteristics of individual cases, it can be preliminarily speculated that some characteristics may be associated with the onset of disease. The characteristics of multiple patients can be obtained from the case series reports. Analyzing the characteristics of these patients can provide more information about the relationship between exposure and disease. On these bases, it can form a preliminary hypothesis that a certain characteristic may be the cause of the disease. However, the power to provide evidence is very weak because of the limitations of this approach. It is a very preliminary etiology suggestion and further research using other epidemiological methods and causal demonstration is needed to validate this etiological hypothesis.

### **3.2.2.4 Identifying Early Disease Outbreaks and Epidemics**

The early manifestations of disease outbreaks and epidemics usually occur in one case and then in several cases, followed by more cases of the same characteristics in susceptible contacts. If the outbreak is not identified and controlled early, the disease can continue spreading through a population, leading to outbreaks and epidemics. Therefore, clinicians should have a keen epidemiological thought and vision. When encountering unusual disease or disease symptom and sign, they should be very vigilant. If this may be a sign of an early outbreak and epidemic of a certain infectious disease, clinicians should report to the local CDC timely and take corresponding preventive and control measures.

## **3.2.3 Case**

### **3.2.3.1 Estrogen Chemical Bisphenol a and Breast Cancer**

A case series report described 15 cases of breast cancer in young women. Nine of the women reported consuming food packaged with estrogenic chemical bisphenol A (BPA) at least once a week, and urine samples of nine patients demonstrated the presence of BPA.

### **3.2.3.2 Occupational Exposure to Vinyl Chloride and Hepatic Hemangioma**

In 1974, Creech and Johnson reported that three of the workers in the vinyl chloride plant were found to have hepatic hemangioma. Three of these patients are clearly unusual in such a small population, and it is easy to form the cause hypothesis that “the occupational exposure to vinyl chloride caused the occurrence of hepatic hemangioma.” In the following year, this hypothesis was confirmed by data from two analytical studies. If there is only one patient, it is not enough to form the cause hypothesis.

### **3.2.3.3 AIDS Discovery Case**

From October 1980 to May 1981, a report of pneumocystis pneumonia was found among young, healthy gay men and women in Los Angeles, United States. This series of reports was unusual because pneumocystis pneumonia previously only occurred in elderly cancer patients with inhibition of the immune system due to chemotherapy. At the beginning of 1981, many cases of Kaposi’s sarcoma were found in young gay men, which is also a noteworthy new discovery. Because this malignant tumor always occurs in the elderly, and the chances of men and women are equal. As a result of these extraordinary discoveries, the US Centers for Disease Control and Prevention immediately implemented monitoring to determine the severity of the problem and developed diagnostic criteria for this new disease. It is quickly noted by monitoring that homosexuals have a high risk of developing the disease. Subsequent case reports and serial case reports indicate that AIDS can also occur through blood transmission in intravenous drug users, blood transfusion patients receiving blood transfusions, and hemophilia patients with blood products. This descriptive data provided clues for the design and implementation of analytical studies and subsequently identified a range of specific risk factors for AIDS. Serum obtained from these cases and comparable controls helped identify the pathogen of AIDS, the human immunodeficiency virus (HIV).

### **3.2.4 Bias**

1. The results are of high promiscuity. The patient is in a natural clinical environment, and the doctor may not be able to control the patient’s ability to seek and receive other treatment or control the patient’s diet and daily life, which may affect the clinical outcome of the disease.
2. The absence of a control group precluded causal inference.
3. The results are less generalizable. Because cases and case series reports are individual. Strictly, it is almost impossible to find other cases of the same

condition in reality. Usually, based on their own knowledge and experience, doctors would choose the case reports which have the most consistent key characteristics for reference.

4. There is a serious publication bias.

### **3.2.5 Limitation**

Although case reports and case series reports are useful in forming etiological hypotheses, their limitations may overrule causal inference.

1. The incidence of disease cannot be obtained from case reports and case series reports. The case report/case series report lacks the population of patients with a disease that is necessary to calculate the disease rate. For example, when calculating the proportion or incidence of breast cancer in women exposed to BPA, the total number of people exposed to BPA or the total number of years must be clear.
2. Case reports and case series reports lack a control group. In the above example, 60% (9/15) of the 15 breast cancer cases were exposed to BPA. The exposure rate appears to be high, but what is the exposure rate in women who do not have breast cancer? This comparison is key to the hypothesis that BPA may be the cause of breast cancer, but it is absent in case reports and case series reports.
3. The cases described in case reports and case series reports are often highly selective subjects, which could not represent the general population well. For example, 15 cases of breast cancer may be from a community hospital with the same severe air pollution or other potential carcinogen concentrations. In this case, a reasonable estimate of the incidence of breast cancer in women in the same community that is not exposed to BPA is needed to infer the relationship between BPA and breast cancer so as to avoid overestimating the link between the two. At the same time, these highly selected cases are highly likely to be reported early, and more cases need to be accumulated, including atypical cases of clinical stages (especially in the middle and late stages), to see the complete history of the disease.
4. There is sampling variability in case reports and case series reports because there might be large natural variations as the disease progresses. The number of cases needs to be increased to estimate the incidence of disease accurately and eliminate the effects caused by chance or sample variation.

## **3.3 Cross-Sectional Study**

### **3.3.1 Concept**

Cross-sectional study is an epidemiological study that describes the distribution of disease or health status among a specific group of population at a specific time and



explores the relationship between variables and disease or health status. Cross-sectional study can get the prevalence of diseases, so it is known as prevalence study. Through cross-sectional study, the occurrence of certain diseases, abnormalities, and vital events in the population can be learned about.

### **3.3.2 Application**

1. To describe the distribution of diseases or health status and provide clues for disease etiology study.
2. Identifying high-risk groups is the first step in early detection, diagnosis, and treatment for chronic diseases.
3. Repeated cross-sectional surveys at different stable stages can not only obtain baseline data of other types of epidemiological studies but also can evaluate the effectiveness of disease monitoring, vaccination, and other prevention and control measures by comparing the prevalence differences at different stages.

Cross-sectional study is the basis and starting point of epidemiological research as well as one of the foothold of public health decision-making. It is a prominent position in epidemiology. Cross-sectional study could not only accurately describe the distribution of disease or health status in a population but also explore the relationship between multiple exposure and disease. But the statistical correlations between disease and exposure revealed by cross-sectional study, which only provides clues to establish causal associations, are derived from analytical studies and cannot be used to make causal inferences.

### **3.3.3 Classification**

Cross-sectional study can be divided into census and sampling survey according to the scope of research objects involved. In the actual work, the use of census or sampling survey mainly depends on the purpose of the research, the characteristics of the research topic, funds, manpower, material resources, and implementation difficulty.

#### **3.3.3.1 Census**

##### **Concept**

Census refers to the survey of all the people in a specific time or period and within a specific range as research objects. A specific time or period means a short time. It can be a certain time or a few days. For too long, disease or health conditions in the

population could change, which may affect census results. A specific range refers to a particular area or population.

### Purpose

- ① Early detection, diagnosis, and treatment can be achieved through census, such as cervical cancer screening in women.
- ② The prevalence of chronic diseases and the distribution of acute infectious diseases, such as the prevalence of hypertension among the elderly and the distribution of measles in children, can be obtained through the census.
- ③ Through a census, the health status of local residents can be obtained, such as residents' diet and nutrition status survey.
- ④ The distribution of disease and its risk factors can be comprehensively understood through census, and the relationship between risk factors and disease can be preliminarily analyzed to provide clues for etiological research.
- ⑤ In a census, all subjects are investigated through a questionnaire or physical examination. In this process, health education could be conducted to popularize medical knowledge.
- ⑥ The normal range of index of all sorts of physiology and biochemistry of the human body can be obtained, just like the measurements of teenage height and weight.

### Conditions for Carrying out the Census

- ① Sufficient manpower, material resources, and equipment are available for case detection and treatment.
- ② The prevalence of diseases should be higher so that more patients can be found and the benefits of census can be improved.
- ③ The disease detection method should be simple, easy to operate, and easy to implement in the field. The experiment should have high sensitivity and specificity.

### Strengths and Limitations

#### *Strengths*

Census surveys all members of a defined population, and there is therefore no sampling error in the census, and it is relatively simple to determine the respondents. Census can provide a comprehensive understanding of the health status and the distribution of diseases or risk factors in a population to establish physiological reference values.

All cases in the population can be found through the census, which provides clues for etiological analysis and research to help with prevention.

Through the census, a comprehensive health education and health promotion activities can be carried out to publicize and popularize the medical knowledge.

### *Limitations*

- ① It is not suitable for the investigation of disease with low prevalence and complex diagnosis methods.
- ② Due to the heavy workload and short survey period, it is difficult to carry out an in-depth and detailed investigation, and there may be missed diagnosis and misdiagnosis. The proportion of no response may be high, affecting the representativeness of the research results.
- ③ Due to the large number of staff participating in the census, the variety of their proficiency in techniques and methods would increase the difficulties to control the quality of the survey.
- ④ Only prevalence or positive rate can be obtained, but not incidence.
- ⑤ Due to the relatively large scope of population involved in census, more manpower, material resources, and time are consumed in research.

### **3.3.3.2 Sampling Survey**

#### Concept

Through random sampling, a representative sample of the population at a specific time point and within a specific range is investigated, and the range of parameters is estimated by the sample statistics, i.e., the overall situation of the population is inferred through the investigation of the research subjects in the sample. In the actual investigation work, there is no need to carry out the census if it is not for the purpose of early detection and early treatment of patients but only to describe the distribution of disease.

The basic requirement of the sampling survey is that the results obtained from the sample can be extrapolated to the entire population. For this reason, the sampling must be randomized, and the sample size must be sufficient (representative).

#### Strengths and Limits

Compared with the census, the sampling survey has the advantages of saving time, manpower, and material resources. At the same time, due to the small scope of the investigation, it is easy to do it in detail. However, the design, organization, and implementation of the sampling survey and data analysis are complex. Duplications and omissions are not easy to find, so they are not suitable for populations with large variations. Sampling is also not appropriate for diseases with low prevalence because

small samples could not provide the information required. In addition, if the sample size is large enough up to 75% of the population, a census may be a better choice.

### **3.3.4 Design and Implementation**

An excellent design scheme is the premise of successful implementation in a research project. It is necessary to pay special attention to the representativeness of the selected subjects in the sampling survey, which is the prerequisite for the overall inference of the research results. Random sampling and avoiding selection bias are important conditions to ensure the representativeness of the research objects.

#### **3.3.4.1 Clarifying the Purpose and Type of Investigation**

According to the research problems expected to be solved, the purpose of the survey should be confirmed, such as to know the distribution characteristics of diseases and the exposure of risk factors or to carry out group health examination. Then, census or sampling survey can be determined to choose based on the specific research purpose according to the specific research purpose.

#### **3.3.4.2 Identifying Subjects**

According to the research objective and the distribution characteristics, geographical scope and time point of investigation, and the feasibility of carrying out the survey in the target population, the research object was determined. In the design, the research object can be defined as all or part of the residents in a certain area. It can also be composed of the floating population at a certain point. It can also be selected as the research object for some special groups. For example, the professional workers exposed to a certain chemical substance can be collected to study skin cancer.

#### **3.3.4.3 Determining Sample Size**

The sample size is the minimum number of observations required to ensure the reliability of the research results. The factors determining the sample size of the present study come from many aspects, but the main influencing factors include the following aspects:

1. Expected incidence ( $P$ ) or standard deviation ( $S$ ). The largest sample size is required when the current incidence rate is 50%.
2. The accuracy of the results of the survey requirements, i.e., the greater the allowable error ( $d$ ), the smaller the sample size required.

3. Significance level ( $\alpha$ ) or the probability of the type I error. The smaller the test level, the more samples are required. For the same test level, the sample content required by the bilateral test is larger than that required by the unilateral test, which is usually taken as 0.05 or 0.01.

Statistical variables are generally divided into two categories: numerical variables and categorical variables. Therefore, the formula used to estimate the corresponding sample size is also different.

#### Estimating Sample Size for Numerical Variables

The following formula is used to estimate the sample size for random sampling:

$$n = \left( \frac{Z_{\alpha} S}{d} \right)^2 \quad (3.1)$$

In the formula,  $n$  is the sample size,  $d$  is the allowable error, i.e., the difference between the sample mean and the overall mean, which is determined by the survey designer according to the actual situation.  $S$  is the standard deviation,  $Z_{\alpha}$  is the normal critical value at the test level and  $\alpha$  is usually taken as 0.05 and  $Z_{\alpha} = 1.96$ .

#### Estimating of Sample Size for Categorical Variables

The following formula is used to estimate the sample

$$n = \frac{t^2 P Q}{d^2} \quad (3.2)$$

In the formula,  $n$  is the sample size,  $P$  is the estimated overall prevalence,  $Q = 1 - P$ ,  $d$  is the allowable error,  $t$  is the statistic of hypothesis testing.

Assuming that  $d$  is a fraction of  $P$ , when the permissible error  $d = 0.1P$ ,  $\alpha = 0.05$ ,  $t = 1.96 \approx 2$ , then formula (3.2) can be written as:

$$n = 400 \times Q/P \quad (3.3)$$

The above sample size estimation formula only applies to the data of Binomial distribution, i.e.,  $np > 5$ ,  $n(1 - p) > 5$ . Otherwise, it is advisable to estimate the sample size by Poisson distribution. The expected value of Poisson distribution and the confidence interval table can be used to estimate the sample size.

The sample size calculation method introduced above is only applicable to simple random sampling. However, in field epidemiologic investigation, cluster sampling is more commonly adopted because it is easy to organize and implement. The sampling error of cluster sampling is large. If the sample size of simple random sampling is

**Table 3.1** The confidence interval of expected value for the Poisson distribution

$1 - 2\alpha$	0.95		0.90	
	Lower	Upper	Lower	Upper
0	0.00	3.69	0.00	3.00
1	0.03	5.57	0.05	4.74
2	0.24	7.22	0.36	6.30
3	0.62	8.77	0.82	7.75
4	1.09	10.24	1.37	9.15
5	1.62	11.67	1.97	10.51
6	2.20	13.06	2.61	11.84
7	2.81	14.42	3.29	13.15
8	3.45	15.76	3.93	14.43
9	4.12	17.08	4.70	15.71
10	4.30	18.29	5.43	16.96
11	5.49	19.68	6.17	18.21
12	6.20	20.96	6.92	19.44
13	6.92	22.23	7.69	20.67
14	7.65	23.49	8.46	21.89
15	8.40	24.74	9.25	23.10
20	12.22	30.89	13.25	29.06
25	16.18	36.90	17.38	34.92
30	20.24	42.83	21.59	40.69
35	24.38	48.68	25.87	46.40
40	28.58	54.47	30.20	54.07
45	32.82	60.21	34.56	57.69
50	37.11	65.92	38.96	63.29

calculated to estimate the sample size of cluster sampling, the sample size will be smaller. Therefore, it is advocated to multiply the sample size for simple random sampling by 1.5 as the sample size for cluster sampling.

**Example** The estimated incidence of colorectal cancer in a city is 30/100,000. How many people should be sampled?

If you take a random sample of 10,000 people, according to the incidence, 30/100,000, the expected number of survey cases is 3. As Table 3.1 shows, when the expected number of cases ( $1 - 2\alpha$ ) is 2, the 95% confidence interval is 0.24–7.22, which means there may be no case. If there was at least 1 case with colorectal cancer, the lower limit of 95% confidence interval is 1.09, and the expected number of cases will be 4. In order to reach at least 4 cases of colorectal cancer patients in the survey results,  $4/X = 30/100,000$ , so  $X = 13,334$ . Finally, 13,334 people should be investigated at least.

In actual work, the sample size can be appropriately increased to avoid errors between the estimated and actual incidence rate.

### 3.3.4.4 Determining the Sampling Method

There is non-random sampling and random sampling. The former includes a typical investigation. A random sampling must follow the randomization principle, which means ensuring that everyone in the population has a known, non-zero probability of being selected as the research object to ensure representativeness. If the sample size is large enough, the data are reliable, and the analysis is correct, the results can then be extrapolated to the population.

The commonly used random sampling methods are simple random sampling, systematic sampling, stratified sampling, cluster sampling, and multi-stage sampling.

#### Simple Random Sampling

Simple random sampling is the simplest and most basic sampling method. The important principle is that each subject is selected with the equal probability ( $n/N$ ). The specific method is as follows: first, all observation objects are numbered to form a sampling frame. Then, some observation objects are randomly selected from the sampling frame by drawing lots or using random number table to form samples.

Simple random sampling is the most basic sampling method and the basis of other sampling methods. However, when the overall number of the survey is large, it is difficult to number each individual in the population. Moreover, the sample is scattered, which is not easy to organize and implement. Therefore, it is rarely used in epidemiological studies.

#### Systematic Sampling

Systematic sampling, also known as mechanical sampling, is to number individuals in a population in order and then randomly select a number as the first survey individual, while the others are selected according to some rules. The most commonly used systematic sampling is isometric sampling, in which all units within the population are sorted and numbered. According to the sample size, the corresponding individual samples are mechanically selected in specific sampling space. The sample numbers taken are:

$$i, i + k, i + 2k, i + 3k, \dots, i + (n - 1)k \quad (3.4)$$

$k$  is sampling space;  $n$  is sample size;  $i$  is the randomly selected starting number.

**Example** If there are 250,000 observation objects in a population, 1000 objects are to be selected for investigation. Sampling can be carried out through systematic sampling. Firstly, the sampling interval is  $k = 250,000/1000 = 250$ , and then one number was randomly selected from the first unit by the simple random sampling

method as the starting point. If  $i$  is 25, the individual numbers were selected successively: 25, 275, 525, 775, 1025, etc.

Compared with simple random sampling, systematic sampling saves time, and the sample distribution is more uniform and representative. However, the disadvantage of systematic sampling is that when the observed individuals in the population have a periodic increasing or decreasing trend, it would produce bias, and the representativeness of the obtained samples will be declined, e.g., the seasonality of diseases, periodic changes of investigation factors.

### Stratified Sampling

Stratified sampling refers to dividing the population into several subpopulations according to certain characteristics, and then conducting simple random sampling from each subpopulation to form a sample. The smaller the intra-group variation and the greater the inter-group variation, the better. Stratified sampling is more accurate than simple random sampling. Moreover, it is more convenient for organization and management.

Stratified sampling is divided into two categories: one is called proportional allocation stratified random sampling, i.e., the sampling proportion within each subpopulation is equal. The other is called optimum allocation stratified random sampling, i.e., the sampling proportion within each subpopulation is unequal. The sampling proportion with small inter-group variation is small, and with large inter-group variation is large.

### Cluster Sampling

Cluster sampling refers to dividing the population into several groups and selecting some groups as observation samples. If all the selected groups are all the respondents, it will be a simple cluster sampling. If some individuals are investigated after sampling again, it is called two-stage sampling. The characteristics of cluster sampling are as follows:

- ① It is easy to organize, convenient to try, and implement;
- ② The smaller the difference between groups, the more groups are extracted, and the higher the accuracy will be;
- ③ The sampling error is large, so it is usually increased by 1/2 on the basis of the simple random sample size estimation.

The above-mentioned four basic sampling methods are introduced. When the sampling method is fixed, the order of sampling error is from large to small: cluster sampling, simple random sampling, systematic sampling, and stratified sampling.



### Multi-Stage Sampling

The sampling process is carried out in multi-stages, with each stage using a different sampling method. Combined with the above sampling methods, multi-stage sampling is commonly used in large epidemiological studies. For example, the InterASIA Study (International Collaborative Study of Cardiovascular Disease in Asia) has adopted the following multi-stage sampling method:

The first stage: the sampling unit is the province and city. Four economically and geographically representative cities were drawn from the south and north, respectively. Beijing and Shanghai were included in the northern and southern samples, respectively, and a total of 10 provinces and municipalities were drawn. It should be noted that in order to fully consider the geographical and economic level of representation, random sampling is not used at this stage, but the random sampling method is used in the next three stages.

The second stage: the sampling units are counties and urban areas. A county and an urban area were randomly selected from the provinces and cities in the first stage, and ten counties and ten urban districts were drawn.

The third stage: the sampling unit is a street, town, or township. Street or town (or township) is randomly selected from each urban area and county, and a total of ten streets and ten towns (or townships) are drawn.

The fourth stage: the sampling unit is an individual. The list of residents of all streets or towns will be used as a sample source (limited to 35–74 years old), and each site will have 400 male and female residents.

The above sampling methods used in four stages are called multi-stage sampling.

Multi-stage sampling can make full use of the advantages of various sampling methods and overcome their shortcomings. The disadvantage of multi-stage sampling is that the demographic data and characteristics of each sub-group should be collected before sampling. Also, the statistical analysis of the data is complicated, such as the sampling weight of the complex sampling design when calculating the sampling error.

#### 3.3.4.5 Data Collection, Collation, and Analysis

In a cross-sectional study, the method of data collection cannot be changed once it is determined. Consistency must be maintained throughout the study to avoid heterogeneity of data. The data collection process should pay attention to the unification definition of exposure and the criteria of disease. All personnel involved in the inspection or testing must be trained to avoid measurement bias with unified investigation and testing standards.

## Determining the Data to be Collected

The most basic principle of the cross-sectional study is whether the subject has a certain disease or characteristics, and the investigator uses grading or quantitative methods as much as possible. In addition, other information, such as social and environmental factors, need to be collected to illustrate the distribution and related factors. The relevant information collected generally includes the following:

- ① Basic information about the individual, including age, gender, ethnicity, education level, marital status, per capita monthly income, etc.
- ② Occupational and exposure status, including nature, type, position, and working years.
- ③ Lifestyle and health conditions, including diet, smoking history, drinking history, physical exercise, depression, anxiety, medical history, disease history, etc.
- ④ Women's reproductive status, including menstrual and obstetrical histories, use of contraceptives, and hormones.
- ⑤ Environmental information, expressed in objective and quantitative indicators.
- ⑥ Prevalence, infection rate, etc.

## Investigator Training

Before the investigation, the investigators should be trained uniformly following a standard protocol. The consistency of the methods and standards for collecting data can be guaranteed. Investigator training is an important part to ensure the accuracy of data.

## Data Collection Methods

In a cross-sectional study, there are three methods for collecting data. The first one is by laboratory measurement or examination, such as blood glucose detection, blood lipid detection, etc. The second way is to investigate the subject through the use of a questionnaire to obtain information on exposure or disease. The third way is to use routine data. For example, get data from the Center for disease control (CDC) and electronic disease records.

## Data Collation and Analysis Methods

Data collation refers to checking the integrity and accuracy of the original data carefully, filling in the missing items, deleting the duplicates, and correcting the errors. Disease or a state of health is verified and classified according to clearly

defined criteria. Then it can be described according to different spaces, time, and the distribution in the crowd.

In data analysis, the population can be further divided into exposed and non-exposed groups or different levels of exposure population. The differences in disease rate or health status between the groups can be compared and analyzed. The subjects can also be divided into disease and non-disease groups to evaluate the relationship between factors (exposure) and disease.

- ① Description of the demographic characteristics. A detailed description of the demographic characteristics (e.g., gender, age, education level, occupation, marital status, and socioeconomic status) can help to easily understand the basic characteristics of the research object and can be used to compare with other studies.
- ② Analysis of the distribution characteristics of the disease: According to the characteristics of the different subjects (gender, age, education level, occupation, marital status, socioeconomic status, etc.), regional characteristics (urban, urban, north, south, mountain, plain, or administrative division, etc.) and time characteristics (season, month, year, etc.) are grouped, the prevalence of a disease or the mean and sampling error of a certain variable are calculated and compared and the correct statistical method is used to test the differences between the different groups.
- ③ Analysis of the relationship between exposure factors and disease: Compare the prevalence of a disease or the mean value of a numerical variable according to the presence or absence of exposure factors or the level of exposure. It is also possible to calculate an odds ratio (OR) to estimate the association and association strength in an epidemiological method (such as a case-control study). Not only univariate analysis but also multivariable adjustments to calculate the ORs are required. What needs to be emphasized here is that cross-sectional study can only provide preliminary clues to the cause.

### **3.3.5 Bias and Control**

Bias is the systematic errors generated in the process from design, implementation, to analysis, as well as the one-sidedness in the interpretation or inference of the results, which leads to the tendency of a difference between the research results and the true value, thus mischaracterizing the relationship between exposure and disease. The common bias in cross-sectional studies includes selection bias, information bias, and confounding bias.

### 3.3.5.1 Selection Bias

Selection bias is the systematic error caused by the differences of the characteristics between the included subjects and those who were not included in the study. It mainly includes the following aspects:

1. Selective bias: In object selection process, due to not strictly sampling, the objects are selected subjectively, which results in the deviation of the research samples from the population. For example, when you want to know the prevalence of hepatitis B in one city last year, if the sample were only information collected from the hepatitis specialized hospital, the prevalence must be higher than the actual rate in the general population.
2. Non-response bias: During the investigation, the subjects did not cooperate or were unable or unwilling to participate for various reasons, resulting in a missed investigation. If the response rate is too low (less than 80% or even 85%), it could produce a non-response bias, and it is more difficult to apply the results to estimate the source population.
3. Survivor bias: In cross-sectional study, survivors of disease are often selected as subjects. Current cases and deaths may have different characteristics, which could not summarize the overall situation. Therefore, the results have some limitations and one-sidedness.

### 3.3.5.2 Information Bias

Information bias is a systematic error that occurs when information is obtained from a research subject during the investigation process. Information bias can come from subjects, investigators, measuring instruments, equipment, and methods.

1. Respondent bias includes recall bias and reporting bias: The subjects were biased by unclear or completely forgotten disease history, drug application history, and risk exposure history.
2. Investigator bias: The bias occurs in the process of collecting, recording, and explaining information from respondents. One reason is, different investigators have different results for the same subject, the other is the same investigator has different results for several surveys of the same subject.
3. Measurement bias is a systematic error caused by inaccurate instrument and incorrect operation procedure. For example, if the sphygmomanometer is not calibrated, all measurement results are higher or lower than true value. The methods of investigation used are not uniform, and bias may occur.

### 3.3.5.3 Confounding Bias

Confounding bias is caused by confounding factors. If the association between exposure and disease is analyzed, then there will be confounding bias.

Bias can be avoided or reduced, so it is necessary to carry out quality control in the research to minimize the occurrence of bias.

1. In the sampling process, keep the randomization principle strictly to ensure the representativeness of the sample.
2. To improve the compliance and test rate of the respondents, each subject should be investigated.
3. To correct the measuring instruments, equipment, and testing methods, including the preparation of questionnaires.
4. To train the investigators, unify survey standards and conduct mutual supervision and spot checks.
5. After the investigation, reviewing and checking the information is needed.
6. In the process of data collation, the correct statistical analysis method should be selected, pay attention and identify confounding and influencing factors.

### ***3.3.6 Strengths and Limitations***

#### **3.3.6.1 Strengths**

1. The implementation time is short, and the results can be obtained quickly. Thus, the research task can be completed in a short time.
2. Compared with other research types, a cross-sectional study is a relatively inexpensive method.
3. It could investigate the association between disease and factors and establish a preliminary etiological hypothesis.
4. A cross-sectional study can provide a basis for making disease prevention and control plans.

#### **3.3.6.2 Limitations**

1. Prevalence, instead of incidence, can only be obtained from a cross-sectional study.
2. Low-prevalence disease and its influencing factors are not suitable for cross-sectional study.
3. The time sequence between exposure and disease cannot be determined, so there is no causal association, and only preliminary etiological clues can be provided.

### ***3.3.7 Cases***

Due to the rapid development of the Chinese economy, the diet and lifestyle have changed greatly. Diabetes, hypertension, hyperlipidemia, and many other diseases

related to diet and lifestyle increased significantly. Li Liming et al. conducted a survey on the situation among Chinese people in 2002.

### **3.3.7.1 Purpose and Type of Study**

The purpose of this study is to investigate the nutrition and health status of Chinese residents and to analyze the main factors affecting the nutrition and health status. Therefore, the cross-sectional study was adopted. Compared with the census, the sampling survey saves more time and cost. Therefore, multi-stage stratified cluster sampling was selected.

### **3.3.7.2 Subjects and Sample Size**

The target population was the national resident population. With 95% accuracy and 90% precision, the minimum sample size was estimated at 225,000. In addition, assuming no response rate of 10%, the final sample size was 250,000.

### **3.3.7.3 Research Content and Data Collection, Collation, and Analysis**

Data collection included questionnaires, medical examinations, laboratory tests, and dietary surveys. Firstly, demographic characteristics, socioeconomic status, disease history, smoking, drinking, and physical activity of the individuals were collected through questionnaires. Secondly, the height, weight, waist circumference, and blood pressure of the individuals were tested. Thirdly, laboratory tests were performed on the serum indexes, including hemoglobin, TC, TG, HDL-C, and LDL-C. Fourthly, the 24-hour retrospective method, food frequency method, and weighing method were used to carry out the dietary survey.

Finally, 243,206 people were enrolled in this study. After adjusting for age and region, the national adjustment rate was calculated by direct standardization method.

The results showed that the consumption of cereals was the maximum, and the dietary structure showed a significant regional difference. The consumption of meat, fruit, and vegetable oil in urban areas is higher than that in rural areas. While the consumption of cereals, tubers, and vegetables in rural areas was higher than that in urban areas. The overweight rate in China is 17.6%, and the obesity rate is 5.6%. Both overweight and obesity rates increase with age. Obesity rates are higher in cities than in rural areas among people over the age of 7. The prevalence of anemia is 15.2%, which is significantly higher in young and middle-aged women than in men. The prevalence of diabetes among Chinese adults is 2.6%, which increases with age. The prevalence in cities rises faster than in rural population.

#### **3.3.7.4 Conclusion**

The nutrition and health status of Chinese population are gradually changing. The prevalence of anemia reflects the lack of trace elements like iron in Chinese population. The prevalence of chronic diseases such as overweight, obesity, and diabetes is increasing rapidly, which has been a threat affecting the health of the Chinese people. In addition, disease prevention should be targeted because of the differences in nutrition and health levels between urban and rural populations.

### **3.4 Ecological Study**

#### **3.4.1 Concept**

Ecological study is also called correlational study. It is used to analyze the relationship between exposures and diseases by describing the exposure and the frequency of diseases in different populations.

#### **3.4.2 Type of Study Design**

##### **3.4.2.1 Ecological Comparison Study**

Ecological comparison study is often used to compare the relationship between the exposure and the disease frequency in different population groups to provide clues for the disease cause. For example, the National Cancer Research Center of the United States has drawn a statistically significant regional difference in the age-adjusted mortality map of oral cancer between 1950 and 1969. The highest mortality is in the urban areas which are dominated by men in the northeast and women in the southeast. Smoking may be a risk factor for oral cancer from this distribution, as smoking in the South is common. Later case-control studies also supported this cause hypothesis. Immigration epidemiological research method can also be applied in ecological studies. It is usually used to analyze the relationship between genetic factors or environmental factors and diseases by comparing the incidence of immigrants and their children with the incidence of residents of the original place of residence and residents of the settlement in different areas.

##### **3.4.2.2 Ecological Trend Study**

The ecological trend study is to continuously observe the changes in exposure levels and diseases in the population and describe their trends. The relationship between

factors and disease is judged by comparing changes in disease before and after exposure changes.

In the implementation of ecological studies, the above two types are often combined. The suspicious etiology of the disease is explored by studying the frequency of occurrence of a disease in multiple regions (multiple groups of comparisons) and at different times (time trends). For example, some researchers analyzed the relationship between water hardness and cardiovascular mortality for gender and age in 63 towns in the UK from 1948 to 1964. It was found that cardiovascular mortality and water hardness were negatively correlated in all genders and ages, especially in men. In urban areas with high water hardness, the increase in cardiovascular mortality was less than in towns with low water hardness.

### ***3.4.3 The Main Application***

1. Etiological assumptions related to the disease distribution can be found through ecological studies. Ecological studies have found that colorectal cancer was more common in developed countries than in developing countries, considering that dietary habits or environmental pollution might be related to the incidence of colorectal cancer.
2. To provide positive or negative evidence for some existing disease causal hypotheses.
3. It can be used to evaluate the effects of intervention experiments or field experiments. For example, promote low sodium intake in the population and then compare the changes in the average sodium intake level before and after the promotion of low sodium salt and the trend of the average blood pressure of the population to evaluate the effect of low sodium salt intervention.
4. To estimate trends in disease changes. Applying ecological trend studies in disease surveillance to estimate trends in a disease can help prevent and control disease. Between 1959 and 1966, the number of deaths from asthma in England and Wales was associated with a simultaneous increase in sales of bronchodilators. After the cessation of bronchodilators in the pharmacy in 1968, the mortality rate of asthma decreased significantly. Therefore, the development of a ban on bronchodilators without prescription was the result of ecological research.

### ***3.4.4 Bias***

Disease information in ecological studies is often derived from historical records (cancer registers, medical records, etc.), while exposure information is often derived from government agency data (tobacco and alcohol sales, etc.). The accuracy of these data directly affects the reliability of research results.



### **3.4.5 *Strengths and Limitations***

#### **3.4.5.1 Strengths**

1. Ecological study can be carried out using existing routine data to save time and money and then quickly yield results. It takes a long time to measure the relationship between a biological indicator and a disease through a prospective study. The preliminary study using an ecological method can narrow the research risk.
2. For unknown disease etiology, etiological clues for further research can be provided by an ecological study. This is the most striking feature of ecological study.
3. In the field of environment or other research, an ecological study is the only alternative research method when the cumulative exposure of an individual is not easy to measure. For example, in the study of the relationship between urban air pollution and lung cancer, it is difficult to estimate the amount of polluted air inhaled by individuals accurately. At this time, multiple methods of ecological comparison can be applied for research.
4. When the range of individual exposure in a population is not large enough, it is difficult to estimate the relationship between exposure and disease. In this case, ecological comparison studies with multiple populations are more appropriate. For example, not only are high-fat diet habits similar, but also the intake is generally high in Western countries. If the relationship between individual fat intake and coronary heart disease were studied only in Western countries, it would be difficult to find a relationship. If a comparative study of the low-fat diet of the Eastern countries was chosen, meaningful results might be found.
5. Ecological studies are appropriate for evaluating population intervention measures. For example, folic acid deficiency in humans can lead to fetal neural tube defect, which was first hypothesized through ecological studies. The addition of folic acid in the pregnant population led to a significant decrease in the incidence of fetal neural tube defects.

#### **3.4.5.2 Limitations**

1. Ecological fallacy: Etiological clues suggested by ecological studies may be either a true or a false association between disease and exposure. The ecological fallacy is a misinterpretation of the association between exposure and outcomes due to an inaccurate assessment.
2. Confounding factors are often difficult to control, especially socio-demographic and environmental variables. Multicollinearity may affect the correct analysis of the relationship between disease and exposure.
3. Because the timing sequence between exposure and disease is not easy to determine, it is difficult to determine the causal relationship between the two variables.

When conducting an ecological study, do not set too many research questions in a study. The differences between population groups should be minimized. The interpretation of the results should be compared with other non-ecological results as far as possible and combined with professional knowledge for comprehensive analysis and judgment.

### **3.4.6 Case**

In order to analyze and evaluate the relationship between life expectancy and fine particulate pollutants in the air, a study from the United States compiled the life expectancy, socioeconomic status, and society of 211 counties in 51 urban areas in 1980 and 2000. Demographic characteristics and concentration of airborne fine particle contaminants were analyzed using regression models to analyze the relationship between airborne fine particle concentration and life expectancy, after adjusting socioeconomic status and demographic variables, as well as the prevalence of smoking. The results of the study showed that a 10  $\mu\text{g}/\text{m}^3$  reduction in the concentration of fine particulate contaminants was associated with an increase in life expectancy of  $0.61 \pm 0.20$  years ( $P = 0.004$ ). And after adjusting the multivariate in the model, the results still remained significant. The results suggested that reduced fine particulate contaminants in the air can increase life expectancy by 15%.

# Chapter 4

## Cohort Study



Li Liu

### Key Points

- A cohort study is an observational study which begins with a group of people who are free of an outcome of interest and classified into subgroups according to the exposure to a potential cause of the outcome. Variables of interest are specified and measured, and the whole cohort is followed up in order to see how the subsequent development of new cases of the disease (or other outcomes) differs between the exposed and unexposed groups.
- There are three types of cohort studies according to the time when information on exposures and outcomes is collected, namely prospective cohort study, retrospective cohort study and ambispective cohort study.
- The measures of associations in cohort studies include relative risk, attributable risk or attributable fraction, population attributable risk or population attributable fraction, and dose-effect relationship.

## 4.1 Introduction

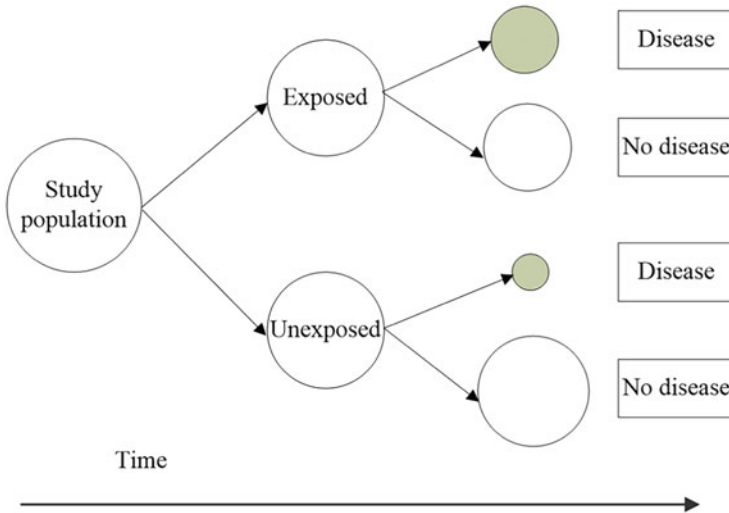
### 4.1.1 Definition

A cohort study is an observational study which begins with a group of people who are free of an outcome of interest and classified into subgroups according to the exposure to a potential cause of the outcome. Variables of interest are specified and measured, and the whole cohort is followed up in order to see how the subsequent development of new cases of the disease (or other outcomes) differs between the exposed and unexposed groups.

---

L. Liu (✉)

School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China



**Fig. 4.1** The design of a cohort study

Exposure means that the subject has been exposed to some substances (e.g., heavy metals) or has some characteristics (e.g., being a carrier of a particular genotype) or behaviors (e.g., alcohol drinking).

The term “cohort” is derived from the Roman army, where it referred to a group of about 480 soldiers, or one-tenth of a legion. Soldiers remained in the same cohort throughout their whole military life, similar to members of epidemiologic cohorts.

According to the time of participants entering the study, the cohorts can be classified into two types: the fixed cohort and dynamic cohort. Fixed cohort means that all participants are enrolled in the cohort at a fixed time or in a short period of time, and followed up until the end of the observation period. The participants have not exited due to other reasons than the outcome, and no new members have joined it. During the whole period, the cohort remains relatively stable. Dynamic cohort, also known as an open cohort, refers to a cohort in which the original members continue to withdraw, and/or new members can join in during the follow-up.

The simplest situation of a cohort study is to recruit one group of population with a specific exposure and one group without that exposure and then follow up for a period of time to see if the participants develop the outcome of interest (Fig. 4.1). The participants must be free of the outcome of interest at the start of the follow-up, which makes it easier to be sure that the exposure precedes the outcome. After a period of time, the investigator compares the incidence rates of the outcome between the exposed and unexposed group. The unexposed group serves as the reference group, providing an estimate of the baseline amount of the outcome occurrence. If the incidence rates are substantively different between the exposed and unexposed groups, the exposure is said to be associated with the outcome. According to the basic principles of cohort studies, there are some basic characteristics:

#### **4.1.1.1 Observational Study**

The exposures in cohort studies are not given artificially, but objectively before the study, which is an important aspect of the difference between cohort studies and experimental studies.

#### **4.1.1.2 Setting up a Comparison Group**

Cohort studies usually set up an unexposed group for comparison during the research design phase. The control group may come from the same population as the exposed group or from different populations.

#### **4.1.1.3 From “Cause” to “Outcome”**

In the course of the cohort study, we usually know the “cause” (exposure factors) first, and then look into the “outcome” (disease or death) through longitudinal observation, which is consistent with experimental research.

### ***4.1.2 Types of Cohort Study***

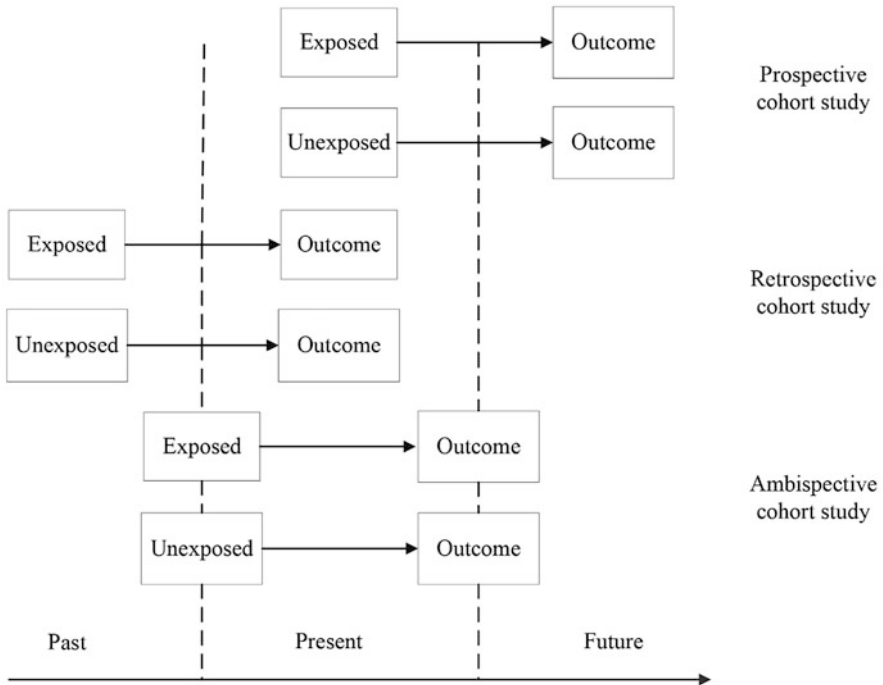
There are three types of cohort studies according to the time when information on exposures and outcomes is collected (Fig. 4.2).

#### **4.1.2.1 Prospective Cohort Study**

In prospective studies, data on exposures are collected at baseline and updated during the follow-up. The outcomes are not available at the beginning of the cohort and should be collected during the follow-up. The investigators could use the most up-to-date measurements to address exposures of interest with minimized bias. However, the investigators need to wait for a relatively long time until a sufficiently large number of events occur. For rare outcomes, the follow-up period may span one, or even several decades.

#### **4.1.2.2 Retrospective Cohort Study (Historical Cohort Study)**

Data on the exposures and outcomes are collected from existing records and can immediately be analyzed. It relies on exposure measurements made before the study set up, which may be available from demographic, employment, medical, or other



**Fig. 4.2** Design of the three types of cohort studies

records. Compared with a prospective cohort study design, it is more useful for rare diseases with a long natural history. By using existing data, the wait time for the exposure to have any impacts on the risk of outcome could be largely reduced. A retrospective cohort study is particularly useful in occupational and environmental epidemiology because if there is a concern that a certain exposure may be a risk factor, it is not reasonable to wait for a long time to confirm in a prospective cohort study. The main disadvantage of a retrospective cohort study is that the exposure data available in records are usually less detailed and accurate than if they were collected prospectively. A retrospective cohort study can be particularly successful when biological specimens were collected in the past so that up-to-date laboratory techniques can be used to detect past exposures. This method could minimize inaccurate exposure measurements in the past, but the stability of the biomarkers during long periods of storage is largely unknown.

#### 4.1.2.3 Ambispective Cohort Study

An ambispective cohort study is a design that combines prospective cohort study with retrospective cohort study. In an ambispective study, a large proportion of participants are still at risk of the outcome when the retrospective cohort are

identified, and the follow-up period can be extended into the future to obtain the maximum amount of information from the cohort. So an ambispective cohort study combines the advantages of both retrospective and prospective cohort studies and to some extent, makes up for their respective deficiencies.

## **4.2 Design of a Cohort Study**

### ***4.2.1 Selection of the Cohort***

The study population includes both exposure and control (unexposed) population. Depending on the purpose and the conditions of the study, the choice of study population varies.

#### **4.2.1.1 Choice of the Exposure Population**

The choice of the exposure population depends on practical considerations and the study hypotheses. There are usually four sources:

##### **General Population**

It refers to a well-defined region of the entire population or its samples. It is composed of individuals with different exposure factors and is suitable for simultaneous observation of multiple exposures and their relationships with multiple diseases. When the exposure group is chosen from the general population, there are two points to be considered: ① Do not intend to pay attention to the incidence of special population, but focus on the general population, so that the research results have universal significance. ② The exposure factors and outcomes of interest are very common in the general population, so there is no need to choose special populations or no special population to choose from. The Framingham Study is a well-known example of a general population cohort.

##### **Occupational Population**

If you want to study a suspicious occupational exposure factor and an outcome, you should select the relevant occupational population as the exposure group. In addition, records on occupational exposures and diseases are often more comprehensive, true, and reliable, so it is a very good source for retrospective cohort study.

### Special Exposed Population

It refers to people with special exposure experiences, which is the only way to study some rare special exposures, such as selecting people who have undergone radiotherapy to study the relationship between radiation and leukemia. If an exposure factor has pathogenic effects, the incidence or mortality of certain disease in special exposed population should be higher than that that in the general population, which could facilitate the identification of the association between the exposure and the disease.

### Organized Population

The organized population can be considered as a special form of the general population, such as school and army members. The selection of such a population mainly relies on the relevant organizational system to facilitate the efficient collection of follow-up information. Given the similar occupational experiences, the occupational people are more comparable.

#### **4.2.1.2 Choice of Control Population**

The unexposed group should be comparable to the exposed group in the distribution of factors that may be related to the outcome of interest except for the exposure. That is, in case of no association between the exposure and outcome, the outcome would have the same incidence in the exposed group and the unexposed group. The control groups mainly include:

#### **An Internal Comparison Group**

An internal comparison group includes unexposed members from the same cohort. Both exposed and unexposed groups are within the selected population. This is usually the best comparison group since the subjects are similar in a lot of aspects. For example, if we want to assess the association between yogurt consumption and the risk of conventional and serrated precursors of colorectal cancer, subjects from the cohort are categorized into groups according to the amount of yogurt consumption, and the group with the lowest intake is used as an internal comparison group, to which the other groups are compared [5]. Cohort studies should try to choose an internal control because it is comparable with the exposed population, easy to conduct, and able to understand the overall incidence of the study.



### An External Comparison Group

When it is impossible to take a well-defined cohort and divide it into the exposed and unexposed groups, the comparison group should be selected outside the cohort, which refers to “external comparison group.” A potential external comparison group is another cohort with similar characteristics but without the exposure of interest. The advantage of this approach is that the follow-up observation can be protected from the exposure group. The disadvantage is the effort to organize another population.

### General Population as a Comparison Group

It can be considered as a kind of external control. The whole population of the geographic area where the exposed cohort is located might be selected as a comparison group (unexposed). Since it is highly impractical to follow up the entire population of a geographic region, incidence rates in the general population are typically derived from routine statistics, which represents an efficient approach compared to studying an additional, unexposed cohort. It takes advantage of existing incidence or mortality statistics across the region, as the morbidity or mortality of the general population is relatively stable and readily available and can save significant time and money, but the disadvantage is that information is often not very accurate. The quality of the information can hardly be checked because it is not collected directly by investigators. Information on potential confounders (other than age, sex, and other basic demographic characteristics) is typically not available in the general population, and confounding by factors such as smoking, cannot be controlled. It should be noted that the control group may contain some exposed subjects, so the total control population applies to a small proportion of the total exposure population. In practice, instead of using a direct comparison of the incidence of the exposed group and the general population, a standardized ratio is used. For example, the standardized mortality rate (SMR) is the ratio of the number of expected morbidity or mortality figures calculated from the incidence or death of exposed groups to the total population.

### Multiple Comparison Groups

Multiple comparison groups refer to that more than one group of people listed above should be selected as control. It can reduce the bias caused by using only one kind of control and enhance the reliability of the results. However, multiple comparison groups undoubtedly increase the workload of the research.

## 4.2.2 Determine the Sample Size

### 4.2.2.1 Matters to Be Considered when Calculating Sample Size

1. In general, the sample size of the unexposed group should not be less than that of the exposed group, usually the same amount. Small sample size may cause increased standard deviation and unstable results.
2. Due to long-term follow-up of cohort studies, the loss of follow-up is inevitable. An estimated rate of loss to follow-up in advance helps to prevent the analysis from being affected by insufficient sample size in the later stage of the study.

### 4.2.2.2 Four Factors Affecting Sample Size

1. The incidence of the outcome of interest in the unexposed population  $p_0$ .
2. The incidence of the outcome of interest in the exposed population  $p_1$ . The greater the difference between the two incidences of the exposed and unexposed populations, the smaller the sample size requires. If the incidence of the exposed group is not easy to obtain, one can try to get the estimate of the relative risk ( $RR$ ) and calculate  $p_1$  by the formula  $p_1 = RR \times p_0$ .
3. Significance level  $\alpha$ : That is the probability of the type I error when making a hypothesis test. The smaller the probability of false positives, the greater the sample size required.  $\alpha$  is usually taken as 0.05 or 0.01.
4. Power ( $1 - \beta$ ):  $\beta$  is the probability of the type II error in the hypothesis test. Power of test refers to the ability to avoid false negatives when testing. The smaller the  $\beta$ , the greater the sample size required. Typically,  $\beta$  is 0.10 and sometimes 0.20.

### 4.2.2.3 Calculation of Sample Size

If the sample size for the exposed and unexposed groups is the same, the sample size required for each group can be calculated using the following formula:

$$n = \frac{(Z_{\alpha} \sqrt{2\bar{p}q} + Z_{\beta} \sqrt{p_0q_0 + p_1q_1})^2}{(p_1 - p_0)^2} \quad (4.1)$$

$p_0$  and  $p_1$  in the formula represent the incidence of the unexposed and exposed groups, respectively;  $\bar{p}$  is the average of the two incidences;  $q = 1 - p$ ;  $Z_{\alpha}$  and  $Z_{\beta}$  are standard normal distribution limits, which can be found from the Standard Normal Distribution Table.

### **4.2.3 Follow-Up**

The follow-up of participants is a very arduous and important work in a cohort study. It should be planned and strictly implemented.

#### **4.2.3.1 Purpose of Follow-Up**

The purpose of follow-up includes three points: identifying whether a subject is still under observation; identifying various outcomes (e.g., disease incidence) in the study population; further collecting data on exposures and confounding factors.

For a variety of reasons, some participants are out of observation during follow-up, a phenomenon known as loss to follow-up, which would have an impact on the findings. When the loss rate is greater than 10%, measures should be taken to further estimate its possible impact. If the loss rate is very high, the authenticity of the study will be seriously questioned. Ensuring follow-up success is therefore one of the keys to successful cohort studies.

#### **4.2.3.2 Follow-up Methods**

Follow-up methods include direct face-to-face interviews, telephone interviews, self-administered questionnaires, periodic physical examinations, environmental and disease monitoring, etc. The follow-up methods should be based on follow-up contents, follow-up objects, and manpower, material, and financial resources. It should be emphasized that the same follow-up method should be used for the exposed and comparison groups, and the follow-up method should remain unchanged throughout the follow-up.

#### **4.2.3.3 Follow-up Contents**

The contents of follow-up are generally consistent with the baseline data, but the focus of follow-up is the outcome of interest. The specific items may be different depending on the purpose and design of the study. In general, one should mainly collect the following information: ① Study outcomes: whether the study population has some kinds of research outcomes. Suspected patients found for the initial examination should be further confirmed. ② Exposure data: what is the exposure of the study subjects? Is there any change? For example, if the study aims to detect the relationship between smoking and lung cancer, one should ask about the amount of cigarette smoking at baseline and during the follow-up. ③ Other relevant information of the study population: the same as the baseline items. ④ Changes in population information: information on lost or retired population, or new arrivals (dynamic cohorts).

#### **4.2.3.4 Endpoint of Observation**

The endpoint of observation means that the subjects develop the desired outcome. For example, when the etiological factor of the disease is studied, often the outcome is the occurrence of the disease or the death caused. When the study subjects develop the outcome of interest, they are no longer observed. In general, the endpoint of the observation is the disease or death, but may also be changes of certain indicators, such as the emergence of serum antibodies and elevated blood lipids, according to the study purpose.

#### **4.2.3.5 Follow-Up Interval**

In theory, follow-up should be carried out after the shortest induction period or incubation period of the disease. The follow-up interval depends on the intensity of exposure and the length of the incubation period of the disease. The weaker the exposure or the longer the incubation period is, the longer the follow-up interval needs. The induction or incubation period of chronic disease is not very clear. In general, the follow-up interval of chronic diseases can be set for several.

#### **4.2.3.6 The Termination Time of Observation**

The termination time of observation refers to the deadline of entire research work, and the expected time to get the result of interest. The termination time of observation is determined according to the length of the observation period, which depends on the incubation period of the disease. In addition, one should also take into account the amount of person-year. One should try to shorten the observation period on the basis of these principles so as to save manpower and material resources and reduce the number of loss to follow-up.

### **4.2.4 *Quality Control***

Cohort studies are by nature time-consuming and expensive. Therefore, the strict implementation process, especially the quality control during data collection, is of particular importance. Generally, the following quality control measures are taken:

#### **4.2.4.1 Selection and Training of the Investigators**

Investigators should maintain strict work ethic and scientific attitude. Honesty and reliability are the basic qualities that investigators should possess. Generally,

investigators should possess the expertise and knowledge required for the investigation. The work ethic, scientific attitude, survey techniques of investigators, and the experience of clinical doctors and laboratory technicians will affect the reliability and authenticity of the survey. Therefore, before data collection, investigators should be trained for better performance during the investigation.

#### **4.2.4.2 Preparation of an Investigator's Handbook**

Due to the large number of investigators involved and the long duration of follow-up in cohort studies, an Investigator's Handbook, including operating procedures, precautions, and a complete description of the questionnaire is necessary.

#### **4.2.4.3 Supervision during the Follow-Up**

Common supervision measures include: repeating the survey among some participants by another investigator, checking numerical or logical errors, comparing the distribution of variables collected by different investigators, analyzing temporal trends of variables, and recording the interviews by using tape recorders, etc.

### **4.3 Data Collection and Analysis**

#### **4.3.1 Data Collection**

The investigators should first collect the baseline information of every participant, mainly including information on exposure status (e.g., the type, duration, and dose of the exposure), personal characteristics (e.g., health status, age, gender, occupation, educational level, marital status), and other circumstances (e.g., home environment, lifestyle and family history of disease). Participants are followed over time, and baseline information is compared with later follow-ups. It also works as a basis to characterize baseline exposures (e.g., classify individuals into exposed or unexposed group, ascertain degrees of exposure and potential confounders), and to obtain tracking materials for follow-up and key information for inclusion or exclusion. The major methods to collect baseline information include data records (e.g., employment, medical examinations, insurance), questionnaires or interviews, physical examinations and tests of biological samples, as well as environmental measurements. Besides baseline information, data collection throughout the process of follow-up is also important (e.g., changes of exposures and measurements of outcomes over time). For more detailed information, please see the second section on follow-up of this chapter.

### 4.3.2 Measures of Outcome Frequency

#### 4.3.2.1 The Basic $2 \times 2$ Tables Summarizing the Results of a Cohort Study

Disease incidence could be described by the cumulative incidence or incidence density. Cumulative incidence is generated by dividing the number of incident cases by the number of persons at risk in the cohort, as shown in Table 4.1:

The cumulative incidence in the exposed group =  $d_1/n_1$ .

The cumulative incidence in the unexposed group =  $d_0/n_0$ .

The cumulative incidence represents the individual risk of developing the disease of interest with no unit. It is a proportion, not accounting for possible different periods of follow-up time, thus mainly used in fixed cohorts. When studying acute outcomes within a short period of follow-up, such as outbreaks, cumulative incidence could be used to estimate the risk of the disease, given a fixed period of follow-up. However, in most circumstances, such as chronic disease research, the periods of follow-up are relatively long; thus, the cumulative incidence is no longer appropriate since the follow-up time usually differ across cohort members. In this situation, the outcome of interest is preferably described by rate, which is incidence density, the other index to reflect disease incidence, and it is widely utilized in dynamic cohorts. Incidence density is calculated by dividing the number of outcome events by the person-time at risk, as shown in Table 4.2:

Incidence density in the exposed group =  $d_1/T_1$ .

Incidence density in the unexposed group =  $d_0/T_0$ .

One should note that a person “at risk” refers to the fact that the outcome of interest can occur within the given time frame. Thus if subjects are immune, they are no longer at risk of getting this disease. If on the other hand, the event of interest is uterine cancer, a hysterectomized woman would not be “at risk.” Measurements of risk and incidence of disease could provide valuable information related to the public health burden of the outcome of interest, which is important for disease prevention and public health management.

**Table 4.1** Measures of cumulative incidence

Exposure status	Cases	Non-cases	Total	Cumulative incidence
Exposed	$d_1$	$n_1 - d_1$	$n_1$	$d_1/n_1$
Unexposed	$d_0$	$n_0 - d_0$	$n_0$	$d_0/n_0$

**Table 4.2** Measures of incidence density

Exposure status	Cases	Person-time at risk	Incidence density
Exposed	$d_1$	$T_1$	$d_1/T_1$
Unexposed	$d_0$	$T_0$	$d_0/T_0$

### 4.3.2.2 Person-Time

In a dynamic cohort, study subjects have unequal periods of time from entry into the cohort to disease occurrence or end of follow-up, and this must be taken into account. Person-time is introduced to reflect the exposure experience of a subject in this circumstance. Total person-time is the summation of the time at risk of individual cohort members to develop the disease, which is often the denominator of the incidence density. The common unit of person-time is person-year. As shown in Table 4.3, people entered the cohort at different ages and experienced separate lengths of time. Before the end of the follow-up, four subjects were diagnosed with disease of interest. The person-years of each person are presented in the last column, and the total person-time in this example is 91 person-years.

This exact computation method is based on the duration of participation of each individual; however, for large cohorts, one may not obtain detailed information for each participant, then approximation method is an alternative though with less precision. The approximate person-years are considered as the average number of the population multiplied by the number of years of observation. The average number of the population refers to the average number of the population at the beginning of two contiguous years or the number of the population in the middle of a specific year. In a hypothetical cohort study which started on September 1, 2014, and finished on September 1, 2017, the numbers of subjects were 15,262 in the beginning, and 15,276 at the end, and more details are shown in Table 4.4. The average population in the 20–29 age group are 26,203 persons:  $(8724 + 8736) /$

**Table 4.3** Data from a fictitious cohort

Person ID	Age at entry	Years of follow-up	Age at end of follow-up	Age at diagnosis	Person-years at risk
1	34	14	48		14
2	37	20	57	52	15
3	30	12	42		12
4	33	17	50	41	8
5	37	9	46		9
6	38	16	54	49	11
7	43	11	54		11
8	32	20	52	43	11
Total		120			91

**Table 4.4** Numbers of subjects in a hypothetical cohort study at different times stratified by age groups

Age groups	2014-09-01	2015-09-01	2016-09-01	2017-09-01
20–29	8724	8736	8740	8730
30–40	6538	6570	6554	6546
Total	15,262	15,306	15,294	15,276

**Table 4.5** Data from a fictitious cohort study to calculate person-years with simple life table

Observing time ( $x$ )	No. of objects				Pearson-years ( $T_x$ )
	At the beginning ( $N_x$ )	Entering the cohort ( $E_x$ )	Occurring outcome events ( $D_x$ )	Lost to follow-up ( $L_x$ )	
2011	1898	76	4	22	1923
2012	1948	70	6	18	1971
2013	1994	52	7	15	2009
2014	2024	30	5	19	2027
Total					7930

$2 + (8736 + 8740)/2 + (8740 + 8730)/2 = 26,203$ . The average population is then multiplied by the number of follow-up years to get the person-time.

Another method to calculate the person-time is to utilize simple life table. The basic equations are as follows:

$$T_x = N_x + \frac{1}{2} (E_x - D_x - L_x) \quad (4.2)$$

$$N_{x+1} = N_x + E_x - D_x - L_x \quad (4.3)$$

$x$  refers to a certain period of time, usually representing 1 year;  $T_x$  is the person-time during  $x$  time;  $N_x$  is the number of population at the beginning of  $x$  time;  $E_x$  is the number of subjects entering the cohort during  $x$  time;  $D_x$  is the number of occurring outcome events during  $x$  time; and  $L_x$  is the number of subjects who are lost to follow-up.

According to the equations above, one can get a simple life table, and the total person-years are the sum of every  $T_x$ .

For example, according to Table 4.5, the person-years in 2011 are

$$\begin{aligned} T_{2011} &= N_{2011} + \frac{1}{2}(E_{2011} - D_{2011} - L_{2011}) \\ &= 1898 + (76 - 4 - 22) / 2 = 1923 \end{aligned}$$

The number of population at the beginning of 2012 is

$$\begin{aligned} N_{2012} &= N_{2011} + E_{2011} - D_{2011} - L_{2011} \\ &= 1898 + 76 - 4 - 22 = 1948 \end{aligned}$$

So the person-years in 2012 are



$$T_{2012} = 1948 + (70 - 6 - 18) / 2 = 1971$$

By that analogy, person-years in 2011, 2012, 2013, and 2014 are 1923, 1971, 2009, and 2027, respectively, and the total person-years are 7930.

### 4.3.2.3 Standardized Mortality Ratio (SMR)

For cohorts with a general population comparison group, one usually estimates the association between an exposure and an outcome by calculating standardized mortality (or incidence) ratios (SMRs). The SMR is the ratio of the observed number of deaths in the cohort and the expected number of deaths in the cohort, given the age-specific mortality rates of a reference population and the age structure of the cohort.

$$\text{SMR} = \frac{\sum_{i=1}^I n_i}{\sum_{i=1}^I t_i \times a_i} \quad (4.4)$$

Where  $I$  stands for the age group,  $n_i$  denotes the number of observed deaths of the age group,  $t_i$  denotes the number of person-years in the age group, and  $a_i$  represents the age-specific mortality rate of the age group from the reference population.

The SMR is commonly adjusted for age, calendar period, and other characteristics like race. Example: There were 1000 workers aged between 40 and 50 in a factory, and four of them died of lung cancer in 2000. Assuming that the mortality of lung cancer among the total population aged between 40 and 50 is 2‰ in 2000, then the expected number of death is 2, and we have known that the practical number of deaths is 4; thus the SMR is 2 ( $4/2 = 2$ ).

### 4.3.2.4 Statistical Tests

To test the statistical difference of incidence rate between the exposed and unexposed groups, U test is commonly used in practice. However, there are some noteworthy conditions to abide by relatively large sample size, not too small  $p$  (incidence rate) and  $1 - p$  (e.g.,  $n \times p$  and  $n \times (1 - p)$  are both over five), and approximately normal distribution of incidence rates.

$$u = \frac{p_1 - p_0}{\sqrt{p_c (1 - p_c) (1/n_1 + 1/n_0)}} \quad (4.5)$$

$p_1$  and  $p_0$  are incidence rates in the exposed group and the unexposed group, respectively;  $n_1$  and  $n_0$  are numbers of subjects in the exposed and unexposed groups, respectively; and  $p_c$  is incorporative sampling rate ( $p_c = \frac{X_1+X_0}{n_1+n_0}$ ,  $X_1$  and  $X_0$  are the numbers of outcome events in the exposed and unexposed groups, respectively). One should subsequently compare the  $U$  value with the standard  $U$  table, then seek out the corresponding  $P$  value and make inference based on the significant level.

Other statistical tests include probabilistic methods based on binomial or Poisson distribution, Chi-Square test, or score test. Similarly, it is notable that each test has its conditions.

### 4.3.3 Measures of Association

#### 4.3.3.1 Relative Risk (RR)

RR refers to the ratio of the probabilities of an outcome in the exposed and unexposed groups. Its value is a positive real number with a range from 0 to  $+\infty$ , and could take the following form:

$$RR = \frac{I_1}{I_0} \quad (4.6)$$

$I_1$  and  $I_0$  refer to risk or rate of outcome in the exposed and unexposed groups, respectively.

There are two alternative and equivalent expressions: the risk ratio and the rate ratio.

Risk ratio is based on the cumulative incidence, with not accounting for person-time. In Table 4.1, risk ratio could be expressed as:

$$RR = \frac{d_1/n_1}{d_0/n_0} \quad (4.7)$$

Rate ratio is the most natural way to express relative risk. It uses incidence density, which takes person-time into account. In Table 4.2, the rate ratio would then be:

$$RR = \frac{d_1/T_1}{d_0/T_0} \quad (4.8)$$

One can also estimate the 95% confidence interval (CI) of the RR using the Woolf method based on the variance of RR. According to Table 4.1, the variance of  $\ln$  RR is computed as follows:

**Table 4.6** General criteria to estimate the strength of association of relative risk

Relative risk		Strength of association
1.0–1.1	0.9–1.0	None
1.2–1.4	0.7–0.8	Weak
1.5–2.9	0.4–0.6	Moderate
3.0–9.9	0.1–0.3	Strong
10-	<0.1	Infinite

Monson [6]

$$Var(\ln RR) = \frac{1}{d_0} + \frac{1}{d_1} + \frac{1}{n_0 - d_0} + \frac{1}{n_1 - d_1} \tag{4.9}$$

and

$$95\%CI \text{ of } \ln RR = \ln RR \pm 1.96\sqrt{Var(\ln RR)} \tag{4.10}$$

One could obtain the 95% CI of RR by taking the antinatural logarithm of 95% CI of  $\ln RR$ .

Risk ratio and rate ratio have the same epidemiological implication, but their values are usually different in the same study. The interpretation of the relative risk is as follows:

If  $RR > 1$ , the risk of disease for the exposure is increased compared with the unexposed group;

If  $RR < 1$ , the risk of disease for the exposure is decreased compared with the unexposed group;

If  $RR = 1$ , there is no association.

The risk in the reference group multiplied by the corresponding RR approximates the risk in the exposed group. The value of RR reflects the level of association. Here are the general criteria to estimate the correlation intensity (Table 4.6):

### 4.3.3.2 Attributable Risk (AR) and Attributable Fraction (AF)

The RR mainly measures the level of risk associated with the exposure to a risk factor. It cannot reflect the impact of the factor in a population. To address this issue, AR and AF are introduced. RR mainly provides clues for etiology, while AR and AF are important for disease prevention and public health. AR, also known as the risk difference or excess risk, is the measure of the rate of disease related to the exposure to a risk factor. Attributable risk is applied to quantify risk in the exposed group which could be attributable to the exposure.

$$AR = I_1 - I_0 = \frac{d_1}{n_1} - \frac{d_0}{n_0} \quad (4.11)$$

or

$$AR = I_1 - I_0 = RR \times I_0 - I_0 = I_0(RR - 1) \quad (4.12)$$

AF is the proportion of the total number of cases related to the exposure to a risk factor. It allows to calculate the proportion of disease attributable to the exposure in the exposed group. This can also be viewed as the proportion of disease in the exposed group that can be avoided through the elimination of the risk factor. It is calculated by dividing the risk difference by the incidence of disease in the exposed group and then multiplying it by 100 to convert it into a percentage

$$AF = \frac{I_1 - I_0}{I_1} \times 100\% \quad (4.13)$$

or

$$AF = \frac{RR - 1}{RR} \times 100\% \quad (4.14)$$

AR and AF are both calculated from incidence rates. One should note that they only make sense for a causal association of a risk factor with an outcome occurrence. The underlying assumption is that no other potential confounders are involved in the occurrence of the outcome.

#### 4.3.3.3 Population Attributable Risk (PAR) and Population Attributable Fraction (PAF)

PAR estimates the proportion of disease attributed to the exposure in the study population. PAR can be looked at as the proportion of a disease that could be prevented by eliminating a causal risk factor from the population. PAR tends to be a function of time because both the prevalence of a risk factor and its effect on the exposed population may change over time, as may the underlying risk of disease. Definitions for PAR and PAF are given by

$$PAR = I_t - I_0 \quad (4.15)$$

$$PAF = \frac{I_t - I_0}{I_t} \times 100\% \quad (4.16)$$

Where  $I_t$  represents the incidence of disease in the total population, and  $I_0$  indicates the incidence of disease in the absence of exposure.

PAF is also given as:

$$\text{PAF} = \frac{P_e(\text{RR} - 1)}{P_e(\text{RR} - 1) + 1} \times 100\% \quad (4.17)$$

Where the prevalence of exposure " $P_e$ " is the proportion of individuals exposed to the risk factor.

#### 4.3.3.4 Dose-Effect Relationship

In some circumstances, there may exist a dose-effect relationship between the exposure and the outcome. To address this, one could stratify the exposure into several levels, with defining the lowest level as a reference, and then calculate RRs of other groups compared to the referent group. Taking Table 4.7 as an example, along with the increase of serum cholesterol level, the relative risk of developing coronary heart disease also increases, which indicates that there may exist a dose-effect relationship between serum cholesterol levels and incidence of coronary heart disease. If necessary, one can further make a trend test.

## 4.4 Common Bias and Controlling

### 4.4.1 Selection Bias

Selection bias occurs when the selection of the exposed and unexposed individuals is related to the occurrence of the outcomes of interest. This is a major potential problem in retrospective cohort studies, since knowledge about the exposure and outcome is likely to differentially influence participants. However, it is generally not a problem in prospective cohort studies, since the outcome of interest has not occurred. A serious potential concern is loss to follow-up in prospective cohort studies [7], which arises when study subjects refuse to participate in or cannot be

**Table 4.7** The occurrence of coronary heart disease stratified by serum cholesterol levels in a fictitious cohort study

Cholesterol level	No. of participants	No. of cases	Risk	Relative risk
Very low	200	2	0.01	1(reference)
Low	300	15	0.05	5
Intermediate	400	40	0.1	10
High	300	60	0.2	20
Very high	100	30	0.3	30

found for the data collection during follow-up. Retention of subjects might be differentially related to both exposure and outcome, and this brings a similar effect that can prejudice the results, causing either an underestimate or an overestimate of an association. For example, if an exposed individual will develop the outcome in the future, but she/he is more likely to be lost to follow-up, then the exposed incidence will be underestimated, along with the RR tending towards the null. Loss to follow-up can result in bias and reduce the statistical power. The primary way to reduce this bias is to improve compliance and response rate of participants.

#### **4.4.2 Information Bias**

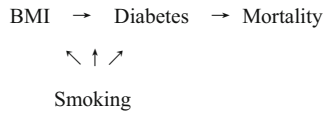
Similar to selection bias, information bias occurs in different ways under different study designs. Reporting bias is one of the potential information biases in cohort studies since the exposure status may influence the reporting of the outcome. For example, in an investigation about occupational hazard, workers are more likely to report having experienced various harmful exposures when this refers to labor guarantee or benefits; thus, some associations may be overestimated. If possible, it would be better to utilize some objective methods and sources of data, such as medical records and laboratory tests, to ascertain the exposure and outcome status. Another important form of information bias is detection bias. Detection bias occurs when knowledge of exposure status differentially increases the likelihood of detecting the outcome of interest among the exposed in cohort studies. A typical example is that a medically relevant exposure could bring about more medical visits and an increased possibility of a diagnostic evaluation, which increases the probability of detecting the outcome in the exposed group. An effective way to address this issue is to apply blinding method to collect information.

Besides, other factors may also contribute to information bias. For example, in the collection of laboratory data, the quality of instruments and reagents, selected measurement standard, measuring conditions and technical competence of the operator are all potential factors influencing the results. Additionally, scientific questionnaires and complete records are also imperative.

#### **4.4.3 Confounding**

Except for selection bias and information bias, confounding is also an important factor that can cause systematic bias in epidemiology, thus the investigators must consider it from study design to data analysis. Confounding distorts the underlying correlation of the exposure with the outcome of interest. The factors causing confounding are called confounders. The criteria for a factor to become a confounder are as follows: the factor must be related with both the exposure and the disease of interest, and at the same time it must not be an intermediate variable in the causal

chain between the exposure and the disease of interest. Directed Acyclic Graph (DAG) is an effective method to distinguish a confounder and a collider. In the following example:



Smoking is a confounder when exploring the association between BMI and the prevalence of diabetes, or the association between the prevalence of diabetes and mortality. However, when exploring the association between smoking and BMI, diabetes acts as a collider (a variable directly affected by two or more other variables with arrows pointing to itself in the DAG, but not the other way around).

In cohort studies, confounding occurs when risk factors are unevenly distributed between the exposed group and the unexposed group. The major methods to control confounding are restriction on inclusion criteria, randomization, and matching. Besides, statistical procedures such as standardization, stratification analysis, and multivariate analysis are also available.

## 4.5 Advantages and Disadvantages of Cohort Studies

### 4.5.1 Advantages of Cohort Studies

1. Strong ability to identify cause-effect association because of the temporal relationship between the exposure and the outcome, reliable data personally observed by researchers and computable indicators reflecting relevance intensity such as RR, AR, etc.
2. Helpful in understanding the natural history of disease in the population.
3. Unexpected outcome data are obtained to analyze the relationship between multiple outcomes and a cause.
4. Able to study the effects of rare exposures.
5. Avoiding recall bias at enrollment.

### 4.5.2 Disadvantages of Cohort Studies

1. It is not suitable for disease with low morbidity because large sample size is needed.
2. In a long follow-up period, lost to follow-up of subjects would cause bias.
3. A large amount of manpower, material resources, and financial resources are required.

4. During the follow-up, the entry of unknown variables and the changes of known variables could influence the outcome, making the analysis complicated.

## 4.6 Example of a Cohort Study

To facilitate the understanding of cohort studies, the design, implementation and main results of a cohort study “**Fresh Fruit Consumption and Major Cardiovascular Disease in China** [8]” is cited. This study is from The China Kadoorie Biobank Study a nationwide, prospective cohort study involving 10 diverse localities (regions) in China. For more details, please see *Du H, Li L, Bennett D, Guo Y, et al. Fresh Fruit Consumption and Major Cardiovascular Disease in China [J]. N Engl J Med. 2016;374(14):1332-1343.*



# Chapter 5

## Case-Control Studies



Qian Wu

### Key Points

- The case-control study population consisted of a case group selected from those with the disease of interest and a control group selected from those who did not have the disease.
- Case-control studies belong to observational studies. It set up a control group.
- In case-control studies, Odds Ratio was used to estimate the strength of the association between disease and exposure factors.
- Selection bias, information bias, and confounding bias are major sources of bias in case-control studies.

### 5.1 Overview of Case-Control Studies

The purpose of the case-control study is to evaluate the relationship between the disease and the exposure factors suspected of causing the disease. Both cohort and case-control studies are analytical studies, their main difference lies in the selection of the study population. In a cohort study, the subjects do not have the disease when entering the study and are classified according to their exposure to putative risk factors, in contrast, subjects in case-control studies are grouped according to the presence or absence of the disease of interest. Case-control studies are relatively easy to conduct and are increasingly being applied to explore the causes of disease, especially rare diseases. Case-control studies are used to estimate the relative risk of disease caused by a specific factor. When the disease is rare, case control study may be the only research method.

---

Q. Wu (✉)

School of Public Health, Xi'an Jiaotong University, Xi'an, China

e-mail: [wuqian@xjtu.edu.cn](mailto:wuqian@xjtu.edu.cn)

© Zhengzhou University Press 2023

C. Wang, F. Liu (eds.), *Textbook of Clinical Epidemiology*,

[https://doi.org/10.1007/978-981-99-3622-9\\_5](https://doi.org/10.1007/978-981-99-3622-9_5)

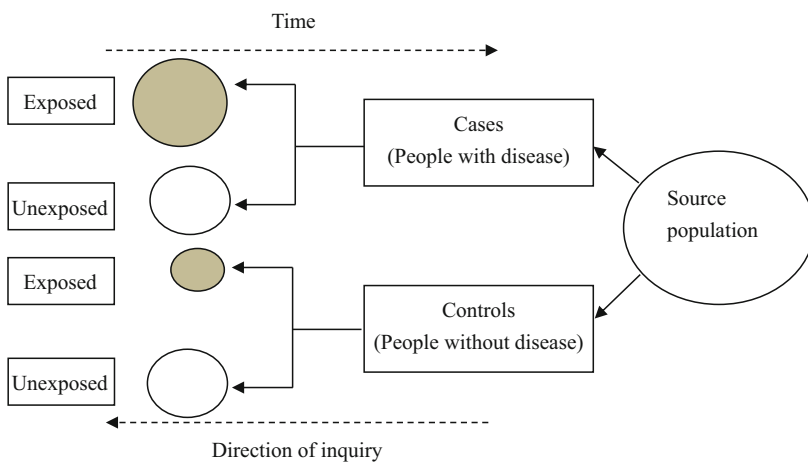
### 5.1.1 History

Case control study has a long history. In 1843, Guy compared male occupations with lung diseases with those with other diseases. But it was not until 1926 that Janet Lane Claypon first proposed a case-control study in a breast cancer research. Richard Doll's research on smoking and lung cancer in the 1950s gave a great impetus to the applications of case-control study. Since then, case-control studies have become more prominent in biomedical literature, and their design, implementation, and analysis have become more standardized in methodology.

### 5.1.2 Definition

A case-control study involves cases from those individuals with disease of interest and controls from those who are without the disease. Previous exposure histories of case and control subjects were examined to evaluate the relationship between exposure and diseases. If the exposure history of the case group and the control group is different, it is possible to infer that the exposure may be related to the disease. The difference in exposure between the case and control group helps to identify potential risk factors. The purpose is to explore whether there are factors related to the disease. The basic principle of a case-control study is shown in Fig. 5.1.

A case-control study is called a retrospective study because researchers need to investigate the exposure factors of the subjects before the occurrence of the disease. Sometimes retrospective studies are used to represent case-control studies. It may be confusing because the terms retrospective and perspicacity are also used to describe the time of data collection related to the current date. In this sense, case-control



**Fig. 5.1** Design of case-control study

studies can be retrospective, when all data are related to the past; it can also be forward-looking, in which data collection continues over time. Therefore, retrospective study is not the essential characteristics of case-control study. The essence of a case-control study is to divide the subjects into case and control groups according to the presence or absence of the disease of interest.

**Example** Some researchers surveyed the relationship between plasma metal concentration and the incidence rate of coronary heart disease (CHD) [Yu Yuan, Yang Xiao, Wei Feng, et al. Plasma Metal Concentrations and Incident Coronary Heart Disease in Chinese Adults: The Dongfeng-Tongji Cohort. *Environ Health Perspect.* 2017,125(10): 107007.]. The researchers compared 1621 CHD cases with 1621 controls free of CHD in Shiyan City, Hubei Province, China, in 2013. All of the participants were retired. Concentrations of aluminum, arsenic and barium, were significantly higher in cases (57.41, 2.32, 40.53  $\mu\text{g/L}$ ) than controls (48.95, 1.96, 35.47  $\mu\text{g/L}$ ). The study presented the concentrations of aluminum; arsenic and barium were higher in the cases than in the controls, indicating that circulating metals were associated with an increased incidence of CHD.

For example, information of participants' disease and their plasma metal was extracted from previous studies. In 2013, investigators according to the interest disease divided retirement employees into two groups. The case group is retirement employees with CHD, while control group is free of major cardiovascular disease. The researchers explored metal concentrations in plasma of participants from 2008 to 2013.

Firstly, the case-control study recruited patients according to their current disease status. Exposure history was inquired for in each case and control. Data were mostly collected after disease occurred, thus case-control study was considered retrospective, which was a limitation. Compared with cohort design, case-control study design has weak support for causal hypothesis. However, it provides more powerful evidence than cross-sectional studies in analyzing and interpreting the results. Case-control study is one of the commonly used research designs. The reason is that the implementation of case-control study is relatively simple and convenient compared with other study designs.

### 5.1.3 Type of Design Case-Control Studies

There are three kinds of case-control studies. First is the *traditional case-control design*. In this type, cases and controls are recruited from population. The case group is assumed to include all cases that occurred in that hypothetical cohort up to the time when the study is conducted. Control group is selected from those without the disease of interest throughout the study period. There are three subgroups of traditional case-control, which are unmatched, frequency matching, and individual matching case-control studies. Next is the *nest case-control design* which is conducted in a cohort population. At the beginning of nest case-control study ( $t_0$ ),

members of the cohort are collected exposure factors. Cases and controls are identified subsequently at time  $t_1$ . The control group is selected from the cohort members who do not meet the case definition at  $t_1$ . Third is a *case-cohort design*. in the first step, a population was identified as the cohort for the study, and a sample within that cohort was selected as the control group using a randomized method. In the whole cohort, all cases of the disease to be studied were collected as a case group. Finally, the two groups were compared and analyzed to explore the factors affecting disease onset, disease survival time, and prognosis.

### 5.1.4 Characteristics of Case-Control Study

Case-control studies belong to *observational study*. Case-control study draws inferences from a sample to a population where the independent variable is not under the intervention of the investigator because of ethical concerns or logistical constraints.

Case-control study set up a *control group*. The differences of exposure were compared between case and control group.

Case-control studies are a special type of *retrospective study*. Investigators look back in time and access prior exposure status between two groups.

The relative risk (RR) cannot be calculated directly in case-control studies, and the Odds Ratio (OR) can be used to estimate the RR.

### 5.1.5 Application

Case-control studies are suitable for investigating rare diseases or diseases with a long latency period, as subjects are selected from the outset based on their outcome status. Therefore, compared to cohort studies, case-control studies are faster and relatively less expensive to implement, require relatively fewer subjects, and allow for multiple exposures or risk factors to be assessed for a single outcome.

#### 5.1.5.1 Example of a Case-Control Study

In October 1989, physicians in the United States reported three patients with a newly recognized disease characterized by marked peripheral eosinophilia with features of scleroderma. After reporting this obvious association, more cases were found in the United States and Europe. To illustrate a possible link between EMS and the tryptophan manufacturing process, they conducted case-control studies to assess potential risk factors, including the use of tryptophan from different manufacturers. In early November 1989, they carried out a case-control study that demonstrated an epidemiologic association between the consumption of tryptophan products and the eosinophilia-myalgia syndrome (EMS). The case-control studies were used to

evaluate potential risk factors, including the use of tryptophan from different manufacturers. The investigators analyzed the tryptophan samples using high-performance liquid chromatography to determine the other chemical component. The results found that 29 of 30 case patients (97%) and 21 of 35 controls (60%) of the subjects using tryptophan had consumed tryptophan produced by one company. The EMS outbreak in 1989 was due to the ingestion of a chemical ingredient that was associated with a specific tryptophan manufacturing condition in one company.

This study suggests several important characteristics of case-control studies. Firstly, the design provides a suitable research method for studying this rare disease of EMS. Case-control study is applicable to the etiology of rare diseases. Secondly, case-control studies allow researchers to investigate several risk factors at the same time. In this research, researchers explored the effect of tryptophan and other factors on EMS. Finally, a case-control study usually does not “prove” causality, but it can suggest a hypothesis. The researchers believe that more research is necessary to identify the composition of the chemicals that trigger EMS and to clarify the pathogenesis of the syndromes. Follow-up revealed that the removal of tryptophan-containing products from the market resulted in the near elimination of reported cases of EMS.

## 5.2 Design of Case-Control Studies

Case-control study is the most commonly used method of analytical epidemiology. In its implementation, the selection of research objects is crucial. Especially the selection of control group is difficult to master. It is usually required that the control should represent the source population that generated the case.

The case-control study determines whether the subjects are case group or control group according to the status at the beginning of the investigation. This status is considered as the outcome variable of the study. The outcome may be whether the subject has been diagnosed with a certain disease or has experienced a complication. Once outcome status is identified and subjects are categorized as cases or controls. Then, information on exposure to one or several risk factors is then collected retrospectively, usually through interviews or surveys.

### 5.2.1 *Basic Principles*

There are three principles of case-control study design. First, it is the study population, also called a source population. The source population may produce the cases and controls. The selection of the control group should not be influenced by exposure factors. Overall, the key issue is for the control group to be representative of the population that generated the cases. The second is de-confounding principle. De-confounding address issues that arise when the exposure of concern is associated

with other possible risk factors. Confounding factors can be eliminated by getting rid of the variability of that factor. For example, if gender is a possible confounding factor, selecting only males would eliminate gender variability altogether. Finally, the principle of comparability was introduced in the two investigation processes. The precision of the exposure measurements was consistent between the control group and the case group. For example, in studies on the effects of smoking on lung cancer, researchers have used nicotine levels in urine to measure smoking in the case group, while questionnaires to measure the controls group, which is inappropriate. Bias due to different measurement methods between cases and controls should be eliminated.

The selection of controls and cases was determined based on the presence or absence of interested disease and could not be influenced by exposure status. Cases and controls do not have to be representative of everyone; in fact, they can be restricted to any specific subgroup, such as elderly, male, or female.

### ***5.2.2 Selection of Cases***

Case groups for case-control studies should be representative of all cases in a population. Case selection is based on interested disease and does not have to consider exposure. Cases were available at the beginning of the study. Cases may include new cases, existing cases, and deaths.

New cases are preferred when selecting cases to avoid the influence of survival factors related to the etiology of the disease. Cases found in one clinic or treated by a physician are alternative cases for case-control studies. The source population of cases treated at a clinic is all those who may be seen at that clinic. Reviewing previous studies, many case-control studies were conducted using one or a small group of hospitals or clinics. This will help to obtain cases in a timely manner and increase the possibility of cooperation, thus limiting selection bias. At the same time, however, there may be problems in the definition of the population from who the case originated.

Community-based population disease registries, particularly for cancer and birth defects, are generally considered to be the best source of cases. This is because the population at risk may be clearly defined by geographic or administrative boundaries.

### ***5.2.3 Selection of Controls***

The most difficult task in case-control studies is the selection of the control group. The control group should be selected from the population that generated the cases with interested disease. Controls are persons without the disease. A key and difficult aspect of population-based case-control studies is to identify a control group in a more efficient way. Otherwise, it would be necessary to demonstrate

that the population providing the control group had the same exposure distribution as the population that was the source of the cases, a very stringent requirement that can rarely be demonstrated. The control group should be selected independent of their exposure status. There are four types of controls in case-control studies.

### **5.2.3.1 Population Controls**

The best control group ensure that controls are random sample of all noncases in the same population that produced the cases. Another way to ensure that cases and controls are comparable is to draw from the same cohort which is called a nested case-control study. The approach, relative to simply analyzing the data as a cohort study, is that analyses are more efficient.

A control group is selected from the same institution or community. Neighbors or friends were controls, and if these individuals showed results of interest, they would be classified as cases. Selecting a control from a neighbor or friend of the case is also a more feasible method. All households in the area surrounding the case were censored and approached in random order until a suitable control was found. It is important to note that the control was present while the case was being diagnosed. The same difficulty is faced with the use of friend control, i.e., random selection from the census of friends provided in each case. The main advantage of friend control is the low level of non-response.

### **5.2.3.2 Hospital or Disease Registry Controls**

The method of selecting controls from hospitals or clinics is more feasible, but it is hardly representative of the source population. For example, a case-control study investigates the relationship between depression and social and economic factors. A particular clinic may be known to have the best depression specialists in a particular area. If both cases and controls are selected from that clinic, then the depression cases may represent the entire region, while the controls represent only the local neighborhood. Cases and controls may then have different social and economic characteristics. Therefore, cases and controls should be selected from multiple diagnosis and treatment institutions to improve their representativeness.

Controls from a medical practice may be more appropriate than controls from hospitals in an urban health center study. The control may have the same high response level as the case. In the medical practice, they may be interviewed in the hospital, which is an advantage from the perspective of the principle of comparable accuracy. The likelihood of patients going to different hospitals varies. If a patient has the disease being studied, the likelihood of going to a specific hospital will be different from the likelihood of going to that hospital for patients with other diseases. In addition, the exposure may be related to the diseases of some controls. Hospital-based case-control studies generally believe that the disease of the control has not associated with exposure. It is hoped that controls for these diseases will effectively

form the basis of the study in a randomized sample. Because there is little certainty about the independence of exposure and disease diagnosis, the standard recommendation is to select controls with multiple diagnoses to ensure that failure of any of them to meet the criteria will not affect the study. If a diagnosis is found to be related to exposure, these controls can be excluded.

## **5.2.4 Matching**

In case-control studies, matching is a common method to control confounding factors. Matching means that the control group is similar to the case group in some characteristics (such as age and sex).

The goal of matching is to control confounders and increase the efficiency of study. If the factors used for matching are related to exposure, the matched control sample usually has a more case-like exposure distribution than the unmatched control sample. Matching eliminates differences in the distribution of certain confounding factors between cases and controls, thus improving the efficiency of the study. In this way, studies can achieve a strong statistical power with a smaller sample size.

Matching begins with the identification of the case group. The investigator then selects a control group from the source population. Matching is divided into two types, depending on whether it is performed at the individual or group level.

### **5.2.4.1 Matching Type**

Matching can be performed on a group of subjects, which is called group matching, or on a subject-by-subject basis, which is called individual matching.

#### **Group Matching**

Group matching means that the matching factors are in the same proportion in the case and control groups and is also referred to as frequency matching. For example, the percentage of women in the case group was 45%, so we chose the control group with 45% women as well. Keeping the control group and case group have the same characteristics (e.g., proportion of male participants). Such that, a group of controls is matched to a group of cases on a particular characteristic (e.g., gender).

#### **Individual Matching**

Investigators select a specific control for each case by matching variables. For instance, if the first case enrolled in a study is a 40-year-old black woman, we will



seek a 40-year-old back female control. Each case can be matched with more than one control group. However, the ratio of controls to cases rarely exceeds 4:1, as the higher the ratio the increasing difficulty of implementation.

#### **5.2.4.2 Overmatching**

If more variables are matched, it may be difficult to find appropriate controls. And we were unable to explore possible associations of the disease with any of the variables already matched in the cases and controls. In this way, overmatching may happen.

An overmatch is a match that causes a loss of information in the study. There are two types of overmatching. The first type is a match that impairs statistical efficiency, such as a variable related to exposure but not to disease being matched. The second type is a match that impairs validity, such as an intermediate variable between exposure and disease being matched. If the investigator happens to match on a factor that is itself related to the exposure, overmatching will appear. For example, in a particular study of NSAIDs and renal failure, if arthritis symptoms were matched in cases and controls, and arthritis symptoms were usually treated with NSAIDs. Matching for arthritis may then affect NSAIDs. This overmatching can decrease the association between exposure and disease.

#### **5.2.5 Exposure**

An important element of case-control studies is to determine the difference in past exposure to a factor between cases and controls. The validity of case-control studies also depends on measuring exposure. In the case-control design, the exposure status of the case is usually investigated after the occurrence of the disease, usually by asking the patient or relatives or friends. The purpose of measuring exposure is to assess the extent of the subject's exposure over a period of time prior to the onset of the disease. The method of collecting exposure data should be the same for cases and controls.

Most case-control studies use questionnaires or interviews to determine the exposure of subjects. The validity of this information will depend in part on the attitude of the subject. People are able to remember well some constant information, such as where they lived in the past and what they did for a living. However, the long-term memory of subjects for specific dietary information may be less reliable. Exposure is sometimes measured by biochemical tests (e.g., calcium in the blood) and may not accurately reflect relevant past exposures if not designed in advance. For instance, lead in the blood of children at age 6 years is not a good indicator of exposure at age 1–2 years. This problem can be avoided if exposure is estimated from established record systems (routine blood tests or stored results from

employment records) or if information is collected prospectively for case-control studies so that exposure data can be collected before disease occurs.

Exposure information can sometimes be determined from historical records. For example, a case-control study on the relationship between sinusitis and multiple sclerosis determined their contact history by searching the general practitioner records of patients and control groups. As long as the records are reasonable and complete, this method is usually more accurate than the method relying on memory.

### 5.2.6 *Sample Size*

The sample size was calculated to ensure confidence in the findings and conclusions of the study. Every researcher wants to complete a meaningful scientific study. The estimation of the sample size is a necessary consideration in the study design. Should an applicant receive funding from a funding agency if a sufficient number of subjects are not enrolled in the study, resulting in no chance of finding a statistically significant difference? Most funding agencies are concerned about sample size and power in the studies they support and do not fund studies that would waste limited resources.

There is also a problem with too large sample size. If the number of samples recruited exceeds the required amount, the duration of the study will be extended. Excessive sample size will also affect the quality of the investigation work and increase the burden and cost of research.

Recognize that sample size is essential to ensure scientifically meaningful results and proper management of financial, organizational, material, and human resources. Let's review how to determine statistical capacity and sampling size. Statistical power is calculated with regard to a particular set of hypotheses.

Statistical power is calculated based on a set of assumptions. Epidemiological hypothesis usually compares the observed proportion or ratio with the assumed value. Statistical power refers to the probability that the null hypothesis will be rejected if the specific alternative hypothesis is true.  $\beta$  denotes the Type II error, i.e., the probability of not rejecting the null hypothesis when the alternative is true. A study should be at least 80% power, and typically studies are designed to have 90–95% power to detect an outcome. What factors affect the power of a study? There are  $\alpha$ ,  $\beta$ , effect size, variability, and  $n$ .

$\alpha$  is the probability of type I error, also known as the significance level of the test hypothesis. This is often determined to be 5% or 1%, implying that the researcher is willing to accept the risk of making a mistake in the alternative hypothesis.

Statistical power is related to effect size, sample size, and significance level. All other factors being equal, an increase in effect size, sample size, or significance level will yield more statistical power.

The sample size of case-control study is calculated according to Formula 5.1.

$$n = \left( \frac{[Z_{\alpha} \sqrt{(1+m)\bar{p}'(1-\bar{p}')} + Z_{\beta} \sqrt{p_1(1-p_1) + mp_0(1-p_0)}]}{(p_1 - p_0)^2} \right)^2 \quad (5.1)$$

$$\bar{p}' = \frac{p_1 + p_0/m}{1 + 1/m} p_1 = \frac{p_0 \text{OR}}{1 + p_0(\text{OR} - 1)}$$

Here  $n$  is that needed individuals in each group,  $\alpha = \text{alpha}$ ,  $\beta = 1 - \text{power}$ . OR is the odds ratio which is the ratio of the exposure ratio between cases and controls. “ $m$ ” is ratio of the sample size of the control group to the sample size of the case group. “ $p_1$ ”—probability of exposure in case,  $p_0$  can be estimated as prevalence of exposure in the control group.

The formula gives the minimum number of cases needed to detect true odds ratio or case exposure with power and bilateral type  $I$  error probability  $\alpha$ .

Calculation of sample size for individual matched case-control studies.

The estimated case sample size for paired matched case-control studies was calculated according to Eq. 5.2, and the control sample size was  $r \times n$ .

$$n = [Z_{1-\alpha/2} \sqrt{(1+1/r)\bar{p}(1-\bar{p})} + Z_{\beta} \sqrt{p_1(1-p_1)/r + p_0(1-p_0)}]^2 / (p_1 - p_0)^2 \quad (5.2)$$

$$p_1 = (\text{OR} \times P_0) / (1 - P_0 + \text{OR} \times P_0)$$

$$\bar{P} = (P_1 + rP_0) / (1 + r)$$

Where  $\alpha = \text{alpha}$ ,  $\beta = 1 - \text{power}$ ,  $P_1$ ,  $P_0$  denote the estimated exposure rates of the case and control groups in the target population, respectively.

### 5.3 Data Collection and Analysis

When researchers have determined the outcomes (disease or health status) of interest in the case-control study and the factors to be studied, they can develop methods for collecting information. The data should include information about research outcomes and factors. Data analysis involves two parts job. First is descriptive data. Next is statistical inference and measure of association. The odds ratio represents an indicator of the association between the disease and each factor of interest.

Researchers often consider data analysis to be the most enjoyable part of epidemiological research. Because after all the hard work and waiting, they have a chance to gain answers. The basic method of analysis in case-control studies is to compare the proportion of exposure in the case and control groups and to calculate the OR.

### 5.3.1 *Main Analysis Objectives*

Assess and refine data quality. Describe the study population and its relationship to the target population. Assess potential bias. Estimate the frequency of exposure. Estimate the strength of the association between exposure factors and disease.

A quality data analysis consists of three phases. In the first stage, the analyst should review the recorded data for accuracy and completeness. Next, the analyst should summarize the data in a concise form and perform descriptive analyses, such as classifying observations according to key factors, using a contingency table. Finally, the summarized data are used to estimate epidemiologic measures of interest, usually expressed in terms of strength of association with appropriate confidence intervals.

### 5.3.2 *Descriptive Analysis*

The number of study subjects and the composition of the various characteristics are described. The exploration of the data reports the frequencies. These measures will provide the basis for important subgroups. Standardization or other adjustment procedures may be required to account for differences in age and other risk factor distributions, duration of follow-up, etc. Compare whether certain basic characteristics are similar between case and control groups.

### 5.3.3 *Statistical Inference*

The indicator that indicates the strength of the association between disease and exposure in case-control studies is the odds ratio (OR). Data analysis included calculating odds ratios as a measure of the association between the disease and the interested factors. When analyzing data on the relationship between exposure and disease variables, we usually have to make statistical inferences about relationship. Several means were employed to avoid random errors, such as p-value and confidence interval (CI) tests. But we should understand that the role of statistically significant is limited. Statistical significance is usually based on the *P*-value: depending on whether the *P*-value is less than or greater than the critical value, usually 0.05. The critical value is then referred to as the alpha level of the test, and the result is considered “significant” or “insignificant.”

The type of analysis used in case-control studies depends on whether controls are sampled in an unmatched or matched manner. Different analysis methods are used for different matching methods.

### 5.3.3.1 Unmatched (Frequency Matching) Design

In case-control studies, researchers attempt to assess the strength of the association between disease and study factors. The investigators analyzed the proportion of exposure in the case and control groups. Data from unmatched or frequency matching case-control studies are summarized in Table 5.1. For better understanding, only two levels of exposure are discussed here. Each object can be divided into four basic cells, which are defined by disease and prior exposure status.

A simple unmatched case-control study, such as that in Table 5.1, can be analyzed by using OR (odds ratio) for association. In case-control studies, groupings are made according to the presence or absence of disease. Therefore, we can't measure health outcomes or disease incidence rate. The proportion of persons in the study who have the disease is no longer determined by risk of developing the disease, but rather by the choice of investigator. So, investigators could not calculate RR (relative risk). Investigators can obtain valid estimates of risk ratios by using OR. When the disease interested is a rare disease, the odds ratio approximates the risk ratio or RR. However, this is not always the case, researcher should be careful taken to interpret the odds ratio appropriately.

$\chi^2$  test and statistical inference (formula 5.3)

$$\chi^2 = \frac{(|ad - bc| - \frac{N}{2})^2 N}{n_1 n_0 m_1 m_0} \tag{5.3}$$

#### Odds Ratio

The odds ratio (OR) is an index of the association between exposure and disease or outcome. The odds ratio is the ratio of exposure in the case group divided by the ratio of exposure in the control group. With the notation in Table 5.1, the odds of exposure for case represent the probability that a case was exposed divided by the probability that a case was not exposed. The odds are estimated by the following formula.

$$\text{Odds of case exposure} = \frac{\text{Exposed cases}}{\text{All cases}} / \frac{\text{Unexposed cases}}{\text{All cases}} = \frac{a}{a + b} / \frac{b}{a + b} = \frac{a}{b}$$

Similarly, the odds of exposure among controls are estimated by the following formula:

**Table 5.1** The result of case-control study

	Case	Control	Total
Exposed	<i>a</i>	<i>b</i>	<i>a + b (m<sub>1</sub>)</i>
Unexposed	<i>c</i>	<i>d</i>	<i>c + d (m<sub>0</sub>)</i>
Total	<i>a + c (n<sub>1</sub>)</i>	<i>b + d (n<sub>0</sub>)</i>	<i>a + b + c + d (n)</i>

$$\text{Odds of control exposure} = \frac{c}{d}$$

The odds of exposure for cases divided by the odds of exposure for the controls are expressed as the OR. Substituting from the preceding equations, the OR is estimated by formula 5.4

$$\text{OR} = \frac{\text{odds of case exposure}}{\text{odds of control exposure}} = \frac{a/c}{b/d} = \frac{a \times d}{c \times b} \tag{5.4}$$

OR indicated “How many times more exposed are cases than no-case exposed?” Since OR have a different scale of measurement than RR, the answer to this question can sometimes differ from the answer to the corresponding question about RR. However, case-control studies are concerned with rare diseases, for which RR and OR are very similar.

### Interpreting the Odds Ratio

A case-control study comparing the smoking habits of 58 lung cancer cases with 93 controls showed the following results (Table 5.2).

$$\text{OR} = \frac{a \times d}{b \times c} = \frac{22 \times 86}{7 \times 36} = 7.5$$

The proportion of lung cancer cases exposed to smoking was 7.5 times greater than the proportion of controls who smoked. It is suggested that there is a strong association between lung cancer and smoking. Smoking could thus be a factor that increases the probability of having lung cancer.

As can be seen, we can determine the risk factors by calculating the OR. It is important to recognize that case-control studies are comparing the odds of exposure [(a/c)/(b/d)] between cases and controls. Conceptually, this is very different from comparing the odds of illness [(a/b)/(c/d)] between exposed and unexposed individuals, which is the result we are really interested in.

Fortunately, in rare disease studies, the ratio [(a/c)/(b/d)] of the ratio of cases and controls with exposure is equal to *ad/bc*. It can also be seen that the odds ratio [(a/b)/(c/d)] in favor of disease in exposed and unexposed populations is also equal to *ad/bc*.

**Table 5.2** Results of a case-control study of lung cancer and smoking

	Individuals with lung cancer (cases)	Individuals without lung cancer (controls)
Smokers	22 ( <i>a</i> )	7 ( <i>b</i> )
Nonsmokers	36 ( <i>c</i> )	86 ( <i>d</i> )
Total	58	93

**Table 5.3** Study on the association between obesity and eating vegetables

	Obese individuals (Cases)	Non-obese individuals (controls)
Eat vegetables	121	171
Do not eat vegetables	129	79
Total	250	250

**Table 5.4** Results of a study on depression and eating vegetables

	Individuals with depression (cases)	Individuals without depression (controls)
Eat vegetables	80	80
Do not eat vegetables	120	120
Total	200	200

Sometimes, the factors studied would reduce the probability of developing the disease. Such factors are known as protective factors of the disease. For instance, 250 obese individuals (cases) in a case-control study were compared to 250 non-obese individuals (controls) in terms of vegetable consumption in their diet. The results are shown below (Table 5.3).

$$OR = \frac{a \times d}{b \times c} = \frac{121 \times 79}{129 \times 171} = 0.43$$

The proportion of cases eating vegetables was 0.43 times greater than the proportion exposed in the control group. Therefore, the proportion of eating vegetables in the case group was 48% lower than the exposure proportion in the control group was 68%. The results of the case-control study showed that compared with the control group, the case group were less likely to eat vegetables. Eating vegetables may be a protective factor in reducing obesity.

Sometimes case-control studies did not find an association between study factors and outcomes. In this case, the OR for the strength of the association between factors and disease in the case-control study was 1.0. For example, in a case-control study, 200 people with depression were compared with 200 people without depression regarding their vegetable consumption (Table 5.4).

$$OR = \frac{a \times d}{b \times c} = \frac{80 \times 120}{80 \times 120} = 1.00$$

The odds of eating vegetables among depressed patients were the same as the odds in the control group. An OR of 1.00 was calculated, indicating a lack of association between depression and eating vegetables. The results of the study did not show an association between eating vegetables and suffering from depression.

In summary,  $OR > 1$  indicates that the factor may increase the risk of disease,  $OR < 1$  indicates that the factor may attenuate the risk of disease, and  $OR = 1$  indicates no association.

### Confidence Interval Estimation of Odds Ratio

An OR is a point value estimate, which may have a random error. The OR confidence interval gives the range of estimates of the OR. The range of estimates is calculated based on a given set of sample data. The OR confidence interval reduces the random error generated by a single study. Ninety-five percent confidence interval (CI) means a 95% probability which the interval includes the true OR. If 95% CI range includes “1,” it is not statistically significant since it could be either a risk factor ( $OR \geq 1$ ) or a protective factor ( $OR \leq 1$ ). If 95% CI range is greater than 1, the exposure is a significant risk factor ( $OR \geq 1$ ) with a probability of higher than 95%.

An approximate 95% CI around the point estimate of OR for an unmatched case-control study can be calculated using the formula (5.5).

$$OR_{95\%CI} = (OR) \exp \left[ \pm 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right] \quad (5.5)$$

Where  $\exp.$  is the natural logarithm, and  $a$ ,  $b$ ,  $c$ , and  $d$  represent the numerical entries into the summary format in Table 5.1.

$$95\%CI = (7.5) \exp \left[ \pm 1.96 \sqrt{\frac{1}{22} + \frac{1}{36} + \frac{1}{7} + \frac{1}{86}} \right] = (7.5) \exp(\pm 1.96 \times 0.477)$$

$$\text{Lower bound} = (7.5) \exp(-1.96 \times 0.477) = (7.5) \exp(-0.94) = 2.9$$

$$\text{Up bound} = (7.5) \exp(+1.96 \times 0.477) = (7.5) \exp(+0.94) = 19.1$$

The CI provides two values, low ( $L$ ) and high ( $U$ ), with a specific confidence level between these two values for the population parameter. A 95% confidence interval means that if we conduct a study, there is a 95% probability that the results will fall within the confidence interval. The above example illustrates that the interval between 2.9 and 19.1 includes a probability of 0.95 for the true OR value.

#### 5.3.3.2 Matched Design

In individually matched case-control studies, the analysis must take into account the matched sampling scheme. When a control is matched to one case, summary data in the format shown in Table 5.5 can appear. This table is different from the one that we



**Table 5.5** A 1:1 matched case-control study

	Control exposed	Control unexposed	Total
Case exposed	$a$	$c$	$a + c$
Case unexposed	$b$	$d$	$b + d$
Total	$a + b$	$c + d$	$a + b + c + d$

introduced in our previous group matching analysis. Each cell in Table 5.5 represents not one subject but a pair (one case and one control). Each case-control pair can be classified as one of the exposure states. Just as Table 5.5, “ $a$ ” means numbers of pairs that both case and control exposed while “ $c$ ” means numbers of pairs that case exposed but control unexposed. “ $b$ ” means numbers of pairs that case unexposed but control exposed. “ $d$ ” means number of pairs that both case and control unexposed. In the analysis of individual matching studies, only pairs with inconsistent exposure were used. Inconsistent pairs of exposures occur when the exposure status of the case differs from that of the control group.

## 2 × 2 Table

### $\chi^2$ Test and Statistical Inference

$$\chi^2 = \frac{(|b - c| - 0.5)^2}{b + c}$$

### OR and 95%CI

The OR of individual matched case-control study is calculated by simple ratio.

$$OR = \frac{c}{b} \tag{5.6}$$

$$OR_{95\%CI} = (OR) \exp \left[ \pm 1.96 \sqrt{\frac{1}{b} + \frac{1}{c}} \right] \tag{5.7}$$

The significance of individual matching OR is the same as that of group matching case-control study. Endometrial cancer and estrogen are used as examples to illustrate the procedure for calculating OR in individual matched case-control studies. The 390 pairs consisted of 390 patients with endometrial cancer and 390 controls,

**Table 5.6** Hypothetical matched case-control study

	Control exposed	Control unexposed	Total
Case exposed	84	96	180
Case unexposed	48	162	210
Total	132	258	390

Table 5.6. Exposure is defined as women who have ever taken any estrogen. The OR from the study is as below.

$$OR = \frac{c}{b} = \frac{96}{48} = 2.00 \quad OR_{95\%CI} = (1.40 \text{ to } 2.89)$$

The calculation method of OR 95% confidence interval (CI) of individual matched case-control study is the same as that of group matched case-control study. Formula 5.7 gives the formula for calculating the OR 95% confidence interval of individual matched case-control study. The approximate 95% CI for the OR is 1.40 to 2.89. This individually matched case-control study showed a moderate association between endometrial cancer and estrogen use.

## 5.4 Common Bias and Controlling

A case-control study is an observational study in which subjects are enrolled based on the presence or absence of the disease of interest. The exposure history of both groups is then evaluated to determine the strength of the association between disease and exposure. Case-control studies are susceptible to observational epidemiological study bias. These biases include selection, information, or confounding biases.

### 5.4.1 Selection Bias

Selection bias is the most common bias in case-control studies. Selection bias may exist if the control group is not from the source population that generated the cases. For example, to study asthma, cases of asthma are drawn from high school students, while people without asthma are drawn from the elderly population to form a control group. The fact that the control and case groups are not a source population has the potential to introduce serious bias. The factors that cause asthma may be different in younger and older people. Thus, based on studies of such mismatched cases and controls, many of the factors that may be found to be associated with asthma may simply be due to the different ages of the two populations.

Sampling of controls and cases can sometimes be stratified, e.g., by sex and age group. In addition to this, there should be randomization in subgroups of subjects

with and without disease. However, researchers are often not randomly sampled, and selection bias arises. This bias poses a significant impact on the validity of case-control studies.

Bias does occur when the sampling fractions depend jointly on exposure and disease, usually because exposed controls are more or less likely to be sampled than non-exposed controls. When hospital patients are utilized as cases and controls, the control is not a random sample of the target population because the control is a subset of hospital patients. Cases in case group are only part patients in the hospital. Patients and hospitals are mutually selective. The systematic differences in some characteristics between the case group and the control group are unavoidable, resulting in an admission rate bias. This is also known as Berkson bias.

The following factors contribute to selection bias.

#### **5.4.1.1 Prevalence-Incidence Bias**

More information might have been obtained if the survey respondents had chosen existing cases, but much of this information was only relevant to survival and may have overestimated the etiologic role of certain exposure factors. In addition, survivors of a disease change their habits so as to reduce the level of a risk factor or distort their pre-morbid habits when they are investigated, resulting in the association of a factor with the disease being incorrectly estimated. This type of bias is usually referred to as prevalence-incidence bias. Therefore, new cases should be included in the investigation as much as possible to avoid the effect of prevalence-incidence bias.

#### **5.4.1.2 Unmasking Bias**

Patients often seek medical attention for certain symptoms unrelated to the causative agent, thereby increasing the detection rate of early cases and leading to an overestimation of exposure. This systematic error is then referred to as unmasking bias.

#### **5.4.1.3 Subject Refuses Participation**

In case-control studies, the most common reason is that subjects refuse to participate, either by actively refusing to sign a consent form or by passively not returning questionnaires or failing to attend laboratory tests at the specified time. Cases tended to be highly motivated to participate, while controls selected from the population were not willing to participate. Participation rates in the control group tended to depend on a number of factors related. For example, rejection rates for telephone surveys are higher for people who are older, less socially connected, less educated, and have lower incomes.

### 5.4.2 *Information Bias*

Information bias is a systematic bias in the process of collecting and organizing information due to flaws in the methods used to measure exposure and outcome. Even if the classification of subjects' exposure and outcome is completely accurate, bias may be introduced due to different choices in case-control studies. More commonly, subjects are incorrectly classified in terms of exposure status or outcome, and estimates of association can be biased. These errors are often referred to as *misclassification*. Misclassification can be classified as differential misclassification or non-differential misclassification.

Differential misclassification is also referred to as "recall bias." Recall bias may arise when cases remember past exposures more completely than controls. This often happens because cases tend to try to find out the cause of their disease. As a result, when they are interviewed, they tend to report more information about the past. Control do not deliberately report information about past exposures.

The second type of information bias is non-differentiated misclassification. Non-differential error classification means that the frequency of errors is similar in the case and control groups. Misclassification of exposure status is more serious than misclassification of outcome. However, both misclassifications can bias a study. For example, a case-control study was conducted to explore the relationship between a high-fat diet and coronary artery disease. Subjects with heart disease and controls without heart disease were recruited and asked to fill out a questionnaire about their dietary habits. Then they were determined whether to consume a high fat diet. It is difficult to accurately assess the amount of fat in the diet from questionnaires. Therefore, it would not be surprising if there were errors in the classification of exposure. In such cases, misclassification may occur regardless of the final disease status. When exposed is qualitative variables, non-differential misclassification always favors the null. Or, if there is an association, whether positive or negative, it tends to minimize it. For example, the OR between a high-fat diet and coronary heart disease is 5.0, but a biased estimate might give an OR is 2.4 if about 20% of exposed subjects are misclassified as "non-exposed" in both disease and control. This implies that the bias tends towards the null.

If there are multiple exposure levels, non-differential misclassification may bias the estimate toward or away from the null, which rely on the category to which the subject was misclassified.

### 5.4.3 *Confounding Bias*

Confounding is that the relationship between exposure factors and outcomes is distorted by external variables. The systematic error generated by this distortion is the confounding bias. Confounding factors usually have three characteristics. One is a variable associated with the exposure and independent of that exposure, and the

third is a risk factor for the disease. The distortion introduced by confounding factors can be significant, and it can even change the direction of the effect. However, confounding bias can be adjusted for in the analysis, which is different from selection and information bias. For example, the crude death rate in city A may be higher than the crude death rate in city B, but after adjusting for age, there is no difference in the adjusted death rate between cities A and B. The age-induced deviation in crude death rates in two cities is known as the confounding bias.

There are two strategies for controlling confounding. Prevent confounding bias from occurring in the first place, which can be done by limiting or matching during the study design phase. Next is to deal with it when it occurs by using analytic techniques such as stratification and statistical model. The effectiveness of all of these strategies except randomization depends on the ability to identify and measure any confounders accurately.

## 5.5 Strengths and Weaknesses of Case-Control Studies

### 5.5.1 Advantage of the Case-Control Study

Case-control studies save time, cost less, and are the most effective design. Case-control studies are the preferred choice for rare disease research. This is because in a cohort design, studies of rare diseases must follow many people to identify those with outcomes. Case-control studies, on the other hand, do not have to worry about no outcomes occurring. Case-control studies are also advantageous in studying diseases with longer latency periods.

In addition, case-control studies have several other advantages. First, occurrence of exposure in subjects retrospectively investigated in case-control studies. Investigators do not have to follow study subjects over time as in cohort studies. Investigators do not have to follow study subjects over time to collect exposure and disease information as they do in cohort studies. Finally, the sample size of the case-control study was small. Compared to cohort studies and experimental studies, case-control studies are easier to implement. (Table 5.7).

**Table 5.7** Advantages and disadvantage of case-control studies

Advantages	Disadvantages
Suitable for research on rare diseases	The relative risk of disease cannot be directly estimated
Suitable for long latency chronic disease studies	Not suitable for studying rare exposures
Smaller sample size required compared to other types of studies	More susceptible to selection bias than alternative designs
Less expensive than alternative designs	Information on exposure may be less accurate than that available in alternative designs.
Save time over other types of study designs	

### 5.5.2 *Disadvantage of the Case-Control Study*

Case-control studies are divided into case and control groups according to the presence or absence of the disease of interest. Therefore, incidence rates could not be calculated for either group. Without knowing the incidence, it is not possible to calculate the relative risk in case-control studies. One can calculate the OR in a case-control study, which is a measure of association that approximates relative risk under certain condition.

The temporal sequence of exposure and disease may be difficult to determine in a case-control study, so it may not be possible to know whether the exposure occurred before the disease. For example, A case-control study of asthma in high school students suggests an association between asthma and cat ownership. However, it may be difficult to know whether high school students had cats first or whether they had asthma attacks first. People usually choose newly diagnosed cases to overcome this drawback.

Although case-control studies have advantages in studying rare diseases, they are not suitable for studying rare exposures (Table 5.7). For example, we would like to study the risk of asthma associated with working in a nuclear submarine shipyard and would probably not prefer a case-control study because only a small percentage of people with asthma would be exposed to this environmental factor.

Case-control studies are grouped by study disease, so they can only be used to study one disease. However, it is possible to study the association between a disease and multiple factors. If want to study more than one disease, you can consider a cohort study design.

In conclusion, case-control studies are a more efficient research method, but the results are susceptible to the influence of known and unknown confounding variables. Case-control studies are suitable for investigating the association between diseases and factors, and the etiology of diseases. When there is limited evidence on a topic, there are cost-effective ways to raise and investigate hypotheses before conducting larger and more expensive studies. Sometime, they are often the only choice of research method, especially when cohort studies or randomized controlled trials are impractical. Case-control studies investigated information about each subject's exposure up to a certain time period. Case-control studies require first defining the case, then identifying the source population that generated the case, and finally identifying the case group and control group. The studies have some strong characteristics such as being cheap, efficient.

# Chapter 6

## Experimental Epidemiology



Xing Liu

### Key Points

- Follow the ethical principles! Read the Declaration of Helsinki. Always remember: “The health of my patient will be my first consideration.” and “A physician shall act in the patient’s best interest when providing medical care.”
- Experimental study serves as the “gold standard” in medical studies for causal inference.
- Different from observational studies, researchers determine the status of exposure of the participants in experimental studies.
- A successful randomization with a large sample size is powerful in eliminating confounding due to known and unknown confounders at baseline. However, if adherence to treatment is poor, or loss to follow-up is serious, new confounding will arise. Loss to follow-up may also introduce selection bias.

An experimental study is the most powerful design in examining causal relationships. The three major types of experimental study in humans include clinical trial, field trial, and community trial, which differ by objectives, principles, implementations, and target populations. Clinical trial aims to evaluate the treatment effects of new drugs or therapies among patients to improve the prognosis. Field trial aims to examine the potential preventive effect of the intervention in reducing morbidity or mortality among healthy individuals. Community trial implements the intervention among healthy people at the group level instead of at the individual level. By performing experimental studies, researchers make causal inferences, confirm the risk factors and protective factors for diseases, and evaluate the effects of interventions in disease prevention and control.

---

X. Liu (✉)  
School of Public Health, Fudan University, Shanghai, China  
e-mail: [liuxing@fudan.edu.cn](mailto:liuxing@fudan.edu.cn)

## 6.1 Basic Ideas of Experimental Study

An experimental study is a prospective study comparing the effect of an intervention against a control in humans. In experimental studies, participants are assigned to groups with different treatments or agents and followed up for a period of time to see if the outcomes vary across groups.

An experimental study is generally expensive and ethically restricted, usually focusing on a narrow question in a highly selected population with a well-defined protocol. Therefore, the experimental studies are reserved for relatively mature research questions suggested by previous observational studies appealing for further confirmative evidence. The key difference between an experimental study and an observational study is that the status of exposure is decided by researchers in an experiment. Not all research question is amenable to the experimental design. It is not ethical to expose people to harmful substances, and it is not always feasible to study the long-term effect of an intervention.

In a classic two-group experiment, one group receives the treatment of interest, and the other group does not. The experimental groups are expected to be identical with respect to extraneous factors affecting the outcome. Thus, if the treatment of interest does not have any effect on the outcome, an identical outcome frequency with random variation would be observed between the two groups. In other words, if the frequency of the outcome varies across the groups, the difference is attributable to the treatment effect plus random variation. This objective can be achieved if all extraneous factors and conditions which may have an effect on the outcome have been controlled. However, in human studies, it is not possible to create completely identical groups with respect to all extraneous factors. Instead, researchers expect the groups to be comparable and exchangeable, with the net effect of extraneous factors to be minimized and much less than the effect of the treatment. In a classic experimental study, a control group is always needed to provide a basis for comparison, randomization is often employed to minimize the influence of confounding, and blinding is used when possible, to eliminate the biases that arise from knowing the treatment assignment.

### 6.1.1 Study Question

Each experimental study should have a specific question stated clearly and in advance. This encourages proper design and enhances the credibility of the findings. The primary question should be the one the researchers are most interested in and the one that could be adequately answered. Generally, the primary question is based on comparing outcomes across treatment groups. The outcome could be a beneficial event including improved prognosis, prolonged survival, increased rate of cure, released symptoms, reduced complications, or improved quality of life. There also may be a series of secondary questions in an experimental study, which can be



elucidated by the data collected. The secondary questions may comprise different response variables and subgroup hypotheses. Both primary and secondary questions should be relevant scientific questions, with important implications in medicine or public health. Adverse events or side effects should also be collected through the experimental study. Unlike the primary and secondary questions, adverse events and side effects may not always be specified in advance. Investigators usually monitor a variety of clinical and laboratory measurements and record the reports from participants. The safety and well-being of participants are the most crucial concerns in performing an experiment. Investigators should always monitor the balance of benefits and risks, and be guided by the independent ethical review committees.

### ***6.1.2 Choice of Intervention***

The intervention techniques employed by an experimental study may be single or a combination of diagnostic, preventive, or therapeutic biologics, drugs, regimens, devices, or procedures. In an experimental study, the intervention a participant receives is assigned by the investigator for the purpose of a study instead of the subject's need. Ethical constraints severely limit the types of interventions and circumstances for an experiment to be performed on human subjects. Adherence to the scientific protocol should not conflict with the subject's best interests. Any exposure given to participants should cause no known harm and should be limited to potential prevention or cure of disease.

### ***6.1.3 Choice of Control***

The choice of the control group is an important design issue in experimental studies, for it provides the basis to make a valid comparison. The methodological principle of choice of control is that the distribution of extraneous factors is the same between the intervention group and the control group to make the two groups comparable. The ethical principle of choice of control is that if there is an optimal, known best therapy or standard, usual care, the new intervention should be compared against it, or added to it. The commonly used types of control groups include standard control, placebo control, self-control, and cross-over control.

#### **6.1.3.1 Standard Control**

Standard control is the most commonly used control in clinical trials. The optimal or standard treatment is assigned to the control group or to both groups, while a new treatment or new therapy is assigned or added to the intervention group. The effect of the new treatment should be compared against the standard care when the latter is

available, instead of against a placebo. Although comparing a new treatment to a placebo or blank control might provide a larger effect size, the goal of the study is to determine whether the new treatment is better than the one currently used, but not if the new treatment has any effect.

### **6.1.3.2 Placebo Control**

When a new intervention is added to standard care or usual care, it is compared against that care plus a placebo. Or when there is no standard care available, the new intervention is compared against a placebo. Placebo control is also commonly used in field trials including the trial of vaccines. Using a placebo control has two major benefits: helps in keeping the blinding, and helps in controlling the “placebo effect.”

For keeping the blinding successful, the formulation, size, and appearance of the placebo should be identical to the new drug. Thus, the participants would not know which study group they are in, improving their adherence to taking the treatment, and preventing them from dropping out once they know they are not in the intervention group. The researchers would also have no idea about which groups of participants are taking the intervention therapy, preventing them from making differential observations and data collection.

The placebo itself has no treatment or preventive effect at all. However, using any form of the drug may induce certain “effect” in both the intervention group and the control group. This kind of psychological benefit is called the placebo effect, even if it occurs among participants in the intervention group. By using a placebo in the trial, the psychological effects in both groups cancel out, and the real effect of the intervention can be observed. However, if the drug or the treatment of intervention has a certain side effect, the subjects might gradually realize the assigned group, and blinding might be broken, thus the compliance and the control of the placebo effect might be weakened.

### **6.1.3.3 Self-Control**

The subjects themselves may serve as the control group before the intervention is given. Or the contralateral body or organ may serve as the control when intervention is assigned to one side. But researchers still have to pay attention if the extraneous factors change before and after the intervention is given. If so, the estimate of the effect might still be confounded.

### **6.1.3.4 Cross-Over Control**

The cross-over design also allows the subject to serve as his or her own control, while in this case, the study has more than one period. In the first period, each subject receives either intervention or control treatment, and in the second period the

alternative. The order in which intervention or control is given to the subject is usually randomized. Depending on the characteristics of the intervention, a wash-out period is required between the two periods. The use of the cross-over design is thus limited to those interventions that the effects during the first period can be washed out. A cross-over control may have more than two periods and may have more than two arms.

### **6.1.4 Randomization**

Observational studies are often used to compare the effects of different treatments given to patients in clinical settings. However, when some of the manifestations affect both the outcome and the treatment allocation, the effect estimates can be biased. The patients in different treatment groups differ in many ways, and the groups might be incomparable. For example, the general condition of a patient has a definite impact on the disease progression and prognosis, and the general condition also determines whether the doctor would choose surgery or a more conservative treatment. In this kind of situation, the differences observed in the outcome between groups may contribute to not only the potential treatment effect but also the confounding brought by the severity of the disease. And this type of confounding is called “confounding by indication.” Thus, the effect estimates gained from observational studies are faced with uncontrolled confounding when the different treatment groups are not comparable, since not all confounders can be realized, identified, measured, and controlled.

The observed association in an observational study comprises the treatment effect, systematic bias, and random error. An experimental study aims to eliminate the part of systematic bias. First of all, it is crucial to reduce to the best extent of incomparability in different treatment groups by balancing the extraneous factors affecting the outcome. Ronald Aylmer Fisher and others developed the practice of randomization to account accurately for extraneous variability in experimental studies. A random assignment mechanism is used to assign treatments to subjects, and the mechanism is unrelated to those extraneous factors that affect the outcome. Thus, the difference in the outcomes across groups that is not attributable to treatment effects could be attributed to chance. A study with random assignment of exposure allows computing the probability of the observed association under the hypothesis and making a statistical inference based on the compatibility between the observation and the hypothesis. Randomization guarantees that statistical tests have valid rates of false positive error.

Successful randomization with a sufficient sample size generates comparable groups at baseline. Not only known confounding factors but also those unknown confounders are balanced across groups. However, compliance with the follow-up and adherence to treatment is critical during the study period to make the effect estimate valid. If adherence to treatment is influenced by extraneous factors affecting the outcome, confounding will arise and affect the effect estimate between exposure

received and the outcome. If the loss to follow-up is severe, it would not only affect the study efficiency but also introduce selection bias and confounding if the loss to follow-up is differential with regard to exposure, outcome, and extraneous factors. Therefore, it is important to maintain a low rate of loss to follow-up and high adherence to assigned treatments during the study period.

#### **6.1.4.1 The Randomization Process**

Randomization is a mechanism or process by which each subject has the same chance of being assigned to either the intervention group or the control group. Several methods of randomly assigning subjects are introduced here. The commonly used methods for randomization include simple randomization, blocked randomization, and stratified randomization.

#### **6.1.4.2 Simple Randomization**

Simple randomization or complete randomization is the most elementary form of randomization. To toss an unbiased coin when a participant is eligible for randomization is an example. One might also use a random digit table on which the equally likely digits from zero to nine are arranged by rows and columns. For larger studies, one may use a random number-producing algorithm provided by most statistical software to generate random numbers in the interval from 0.0 to 1.0 for each subject. The procedure might assign subjects to group A with probability  $p$  and subjects to B with probability  $1-p$ . If the random number is between 0 and  $p$ , the subject would be assigned to group A, otherwise to group B. And this procedure can be adapted to more than two groups. The advantage of simple randomization is that it is easy to implement. Simple randomization generates an anticipated proportion for the number of subjects in each arm in the long run with a large sample size. However, at any point in the process of randomization, or when the sample size is small, there could be a substantial imbalance. Although this kind of imbalance would not cause the statistical tests to be invalid, it harms the statistical efficiency.

#### **6.1.4.3 Blocked Randomization**

Blocked randomization is used to avoid serious imbalance in the number of subjects assigned to each group when the sample size is small. It also helps to have balanced numbers at any point in the randomization procedure during enrollment. If participants are randomly assigned with equal probability to groups A or B, then for blocks with even size, one-half of the participants would be assigned to each group. The order in which the treatments are assigned in each block is randomized. For example, if a block of size 4 is used, there are six possible combinations of treatment assignments: AABB, ABAB, ABBA, BAAB, BABA, and BBAA. Select one

from these arrangements randomly and apply it accordingly to the four participants entering the study. Repeat for every consecutive group of four participants until all are randomized. Advantage of the block randomization is that the number of participants in each group is always balanced during the process of randomization, at any time point, and with any sample size. The disadvantage is that strictly speaking, data analysis is more complicated for blocked randomization than for simple randomization. And the use of blocked randomization should be taken into consideration during data analysis.

#### **6.1.4.4 Stratified Randomization**

Randomization balances extraneous factors that affect the outcome in studies with large sample sizes and for small studies on average. However, for one single study especially with a small sample size, it is possible that not all baseline characteristics distribute evenly across groups. When there is the concern of imbalance for major prognostic factors, one might employ stratified randomization within the strata of those factors considered. If several factors are considered, the number of strata is the product of the number of subgroups for each factor. Within each stratum, the randomization process could be a simple randomization or a blocked randomization.

#### **6.1.5 Blinding**

Blinding is one of the solutions to reduce systematic biases in experimental studies. Not knowing which group the participant is in, the adherence to the exposure, the compliance to the follow-up, and the measurement of outcomes can be improved. Thus, blinding is often employed when it is possible. Some kinds of trials can only be conducted without blinding, including those that have surgical procedures, changes in lifestyle, or behavioral interventions. The main disadvantage of an unblinded experiment is that participants may report symptoms and side effects differentially between intervention and control groups. Also, researchers may measure and collect these data differentially when knowing which group the subjects are from. Moreover, the participants in the control group may have a higher possibility of leaving the study, when knowing that there would not be any extra benefits.

There are at least four levels of blinding in a trial. First, the participant does not know which treatment group he or she is in. The adherence and compliance to the study would not be influenced, and the accuracy of the report of symptoms would not be affected. Second, the staff assigning participants to different treatment groups do not know which group the participants would be assigned to. This avoids participant assignment based on the staff's willingness. Third, the physicians taking care of participants during the whole process do not know which treatment group the participants are in. This ensures the health care provided and symptoms and clinical manifestations recorded would not be affected. Fourth, the researchers including

statisticians do not know which treatment group the participants are in. This makes sure that the measurements of treatment effects, the record of side effects, and data analyses would not be affected.

There are three common methods for performing blinding in practice: single-blind, double-blind, and triple-blind.

### **6.1.5.1 Single-Blind**

In a single-blind study, the participants do not know which treatment group they are in. Thus, the biased report of symptoms and side effects by subjects can be reduced. However, the researchers can still influence the administration of treatment, data collection, and analysis in a single-blind study.

### **6.1.5.2 Double-Blind**

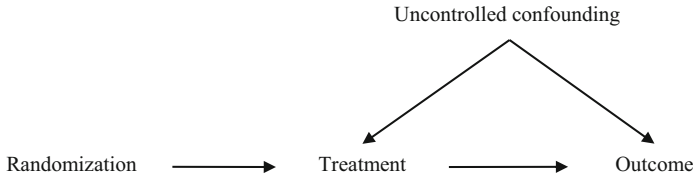
In a double-blind study, neither the participants nor the researchers know the treatment assignment. The risk of bias is greatly reduced in a double-blind study. The actions of investigators would occur equally to participants from both groups. Double-blind studies are usually more complex to carry out than a single-blind or unblinded study. An effective data monitoring protocol should be set up. And the emergency unblinding procedures must be established.

### **6.1.5.3 Triple-Blind**

A triple-blind study is an extension of a double-blind design. And it may have different definitions under different circumstances. In some designs, it is the committee monitoring response variables that is not aware of the treatment assignment. While in some designs, it is the group performing statistical analyses that has no idea of the treatment assignment. Thus, the bias introduced during statistical analyses can be avoided.

## **6.1.6 Data Analysis**

Data analysis in experimental studies has special strategies. Noncompliance with the assigned treatment results in a discrepancy between the treatment assigned and the treatment actually received. The standard practice of data analysis in the experimental study is making comparisons based on the treatment assigned instead of received. Such a practice is called the intent-to-treat analysis (ITT). Comparisons based on the treatment received are called according-to-protocol (ATP) or per-protocol analysis (PP). If the compliance to treatment is poor, or there is a considerable loss to



**Fig. 6.1** A causal diagram with valid instrument randomization, for the treatment—outcome effect. If 1. Randomization affects outcome; 2. Randomization affects outcome only through treatment; 3. Randomization and outcome share no common causes; and then randomization can be taken as a valid instrumental variable in examining the association between treatment and the outcome. The association between treatment and outcome might have been affected by uncontrolled confounding, however, the association between randomization and outcome has not been confounded

follow-up during the study, the association between the exposure received and the outcome might be biased. ITT analysis preserves the validity of the test for the null hypothesis of treatment effects.

In an intent-to-treat analysis, no matter how the compliance to the assigned treatment is, the analysis takes the assigned treatment as the exposure to test the null hypothesis between the exposure and the outcome. As mentioned earlier, successful randomization is not affected by the extraneous factors affecting the outcome. And randomization has an effect on the outcome through and only through the actual treatment received. Although the association between the treatment actually received and the outcome can be confounded, the association between the treatment assigned (randomization) and the outcome will not be confounded. This makes randomization a valid instrumental variable in examining the association between the treatment and the outcome. The use of the instrumental variable protects the validity of the test of the null hypothesis between treatment and outcome, although the effect estimate might have been biased (Fig. 6.1).

### 6.1.7 Sample Size

An experimental study should have sufficient statistical power to detect the differences across treatment groups. The sample size of a study is decided by the following aspects:

- (1) The significance level, denoted as  $\alpha$ . It is the probability of a false positive finding, or Type I error.
- (2) The probability of a false negative result, or Type II error, denoted as  $\beta$ .  $1 - \beta$  is the statistical power of the test.
- (3) The difference between the measurements of the outcome across the groups.

### 6.1.7.1 Sample Size Calculation for Dichotomous Response Variables

$$N = \frac{[Z_\alpha \sqrt{2P(1-P)} + Z_\beta \sqrt{P_c(1-P_c) + P_e(1-P_e)}]^2}{(P_c - P_e)^2} \quad (6.1)$$

Where  $N$  = the sample size for each group,  $P_c$  is the event rate for the control group,  $P_e$  is the event rate for the treatment group,  $P = (P_c + P_e)/2$ ,  $Z_\alpha$  is the critical value which corresponds to the significance level  $\alpha$ , and  $Z_\beta$  corresponds to the power  $1 - \beta$ .

### 6.1.7.2 Sample Size Calculation for Continuous Response Variables

$$N = \frac{2(Z_\alpha + Z_\beta)^2 \sigma^2}{d^2} \quad (6.2)$$

Where  $N$  = the sample size for each group,  $\sigma$  is the estimated standard deviation,  $d$  is the estimated difference of the means,  $Z_\alpha$  is the critical value which corresponds to the significance level  $\alpha$ , and  $Z_\beta$  corresponds to the power  $1 - \beta$ .

### 6.1.7.3 Sample Size Calculation for “Time to Failure”

$$N = \frac{2(Z_\alpha + Z_\beta)^2}{[\ln(\frac{\lambda_c}{\lambda_e})]^2} \quad (6.3)$$

Where  $N$  = the sample size for each group,  $\lambda$  is called the hazard rate or force of mortality,  $Z_\alpha$  is the critical value which corresponds to the significance level  $\alpha$ , and  $Z_\beta$  corresponds to the power  $1 - \beta$ .

## 6.2 Clinical Trial

### 6.2.1 Basic Ideas of Clinical Trial

Clinical trial is an experimental study with patients as subjects. The goal of a clinical trial is to evaluate a new drug or therapy for a disease to improve prognosis, reduce mortality or improve the quality of life among patients. It also collects information



on the adverse effects of a new treatment and provides evidence on the effectiveness and safety of the treatment to enter clinical use.

Subjects in a clinical trial are patients with the disease in question. Participants in a clinical trial should meet the criteria of eligibility well-defined in advance. Patients who do not meet those criteria should not be enrolled. And subjects whose illness is too severe or too mild usually will not be considered eligible since they are less likely to permit the form of treatment or to complete the follow-up. Patients with complicated conditions are usually excluded especially at earlier stages of the trial because of the need to minimize differences in the extraneous factors affecting the outcome between treatment groups. Therefore, at earlier stages of the trial, the participants are usually a highly selected population with restricted criteria for inclusion, affecting the generalization of the conclusion.

## **6.2.2 *Phases of Clinical Trial***

When comparing the effectiveness of a new drug, several phases of clinical research must be performed. Classically the trials of pharmaceutical agents involve phases I to IV.

### **6.2.2.1 Phase I Studies**

Phase I studies collect early data in humans after preclinical information is obtained from in vitro or animal studies. Participants in phase I studies are generally healthy volunteers with sample sizes ranging from 20 to 100. Phase I studies characterize pharmacokinetics and pharmacodynamics and estimate the tolerability in humans. The questions including bioavailability, body compartment distribution, and drug activity are answered by phase I studies. The maximally tolerated dose, the safety range of the dose, and the recommended dose is explored at this stage. Phase I also collects data on side effects.

### **6.2.2.2 Phase II Studies**

Phase II studies evaluate whether the drug has any biological activity or effect once the dose or range of dose is determined with sample sizes ranging from 100 to 300. A phase II study usually employs a randomized control design, compares the effect of the new drug against the standard drug or a placebo, and evaluates the effectiveness and safety of the new treatment. Phase II studies continue to collect side effects data, evaluate the safety, and recommend the dose for clinical use.

### **6.2.2.3 Phase III Studies**

Phase III studies are generally multi-center randomized controlled trials conducted in different countries with sample sizes ranging from 300 to 3000 or more. Phase III studies further evaluate the effectiveness and safety of the new drug or therapy against the standard care, confirming the value in clinical use. Phase III collects data on the adverse effects and the interaction of the drug with other drugs. The treatment approved after phase III can be used in clinical settings.

### **6.2.2.4 Phase IV Studies**

Phase IV studies are conducted after the new treatment is approved and used clinically. All patients who received the new drug can be considered participants. The participants enrolled before phase IV are generally highly selected with restricted criteria for eligibility, which limits the generalizability of the study conclusions. Phase IV studies observe the drug efficacy in the real world, and those patients with complex conditions may also be enrolled. Thus, the limitations of earlier studies can be improved. Phase IV studies are generally open cohort studies, monitoring drug efficacy, side effects, and interaction with other drugs at a large scale in the long run. Phase IV studies can collect data on side effects especially the ones that occur rarely or late.

## **6.2.3 Case Study of Clinical Trial**

A randomized, phase II study examined the efficacy of carboplatin and pemetrexed with or without pembrolizumab for advanced, non-squamous non-small-cell lung cancer (NSCLC). The investigators enrolled 123 non-squamous NSCLC stage IIIB or IV patients without former chemotherapy and targetable EGFR or ALK genetic aberrations from 26 medical centers in the USA and Taiwan, China. A 1:1 ratio in blocks of four randomization assigned 60 to the group of pembrolizumab plus chemotherapy, and 63 to the group of chemotherapy alone. The primary endpoint was the proportion of patients who had radiologically confirmed complete or partial responses. Fifty-five percent (95% CI 42–68%) of patients in the pembrolizumab plus chemotherapy group achieved this objective response compared with 29% (18–41%) of patients in the chemotherapy alone group (treatment difference 26% [95% CI 9–42%];  $P = 0.0016$ ). The incidence of grade 3 or worse treatment-related adverse events was similar between groups (39% in the pembrolizumab plus chemotherapy group and 26% in the chemotherapy alone group). The most common grade 3 or worse adverse events in the pembrolizumab plus chemotherapy group were anemia (12%) and decreased neutrophil count (5%). In the chemotherapy-alone group, the most common were anemia (15%) and decreased neutrophil count,

pancytopenia, and thrombocytopenia (3% each). 2% of patients in the pembrolizumab plus chemotherapy group experienced treatment-related death compared with 3% in the chemotherapy group. These results suggested that the combination of pembrolizumab plus chemotherapy may be an effective treatment for patients with early, advanced non-squamous NSCLC.

## **6.3 Field Trial**

### ***6.3.1 Basic Ideas of Field Trial***

Field trial differs from a clinical trial in the subjects. The participants in a clinical trial are those patients diagnosed with the disease of interest in clinical settings; while the participants in a field trial are those healthy people from the community. Field trial often requires a larger sample size and recruit participants who are not under clinical management. Therefore, they are often more expensive and difficult to conduct. A field trial is limited to studying the prevention of common or severe diseases. The interventions for field trials include health supplements, vaccines, and changes in lifestyle. The principles of study design, control selection, randomization, and blinding for experimental studies apply to field trials. Field trials are used to confirm the causal relationship, risk factors, and preventive factors for diseases and to reduce morbidity.

### ***6.3.2 Design and Implementation***

Participants in the field trial are free-living healthy people recruited from the community. The management and conduct of a field trial would be more difficult than a clinical trial. A well-designed feasible protocol on a solid scientific question is crucial for a successful field trial.

#### **6.3.2.1 A Specified Question**

A clear scientific question should be stated in advance including the specific intervention and anticipated outcome. The objective of the study should be based on a clear research hypothesis. The intervention should be derived from a risk factor for disease with relatively sufficient evidence from observational studies. And the conclusions gained from the study should have benefits for individuals or public health.

### **6.3.2.2 Inclusion and Exclusion Criteria**

The participants for the field trial are healthy people from the community and are at risk for disease of interest. The inclusion and exclusion criteria should be defined in advance based on the study objective and should be implemented strictly. Participants can be enrolled from those communities with low mobility to avoid a substantial loss to follow-up, otherwise, selection bias may arise and harm the validity of the conclusion. Also, if the disease of interest is of low incidence rate in the population, it is suggested to conduct a field trial in the population at higher risk for the disease to save resources for long-term follow-up. Restricted inclusion criteria and highly selected participants may have an influence on the generalizability of the research conclusion.

### **6.3.2.3 Choice of Intervention**

A clear definition and description of the intervention are necessary. The dose, contents, method, frequency of application, etc. of the intervention should be introduced clearly. Adherence to intervention is critical for field trial participants.

### **6.3.2.4 Time and Interval of Follow-up**

The time of each visit and interval during follow-up are decided by the effect of an intervention. Investigators balance the need for collecting necessary data, maintaining participants in the follow-up, and the cost. During the study, it is important to improve the compliance and adherence of the participants to avoid a loss of follow-up and selection bias.

## ***6.3.3 Case Study of Field Trial***

Efficacy of a bivalent L1 virus-like particle vaccine in prevention of infection with human papillomavirus types 16 and 18 in young women: a randomized controlled trial

Genital human papillomavirus (HPV) infection leads to cervical cancer, a major cause of cancer deaths in women worldwide. 230,000 die and 470,000 are diagnosed due to cervical cancer annually. The most prevalent oncogenic HPV strains, HPV-16 and HPV-18, can be vaccinated to prevent up to 70% of cervical cancers from developing. A bivalent HPV-16/18 L1 virus-like particle vaccine was tested for efficacy, safety, and immunogenicity in a randomized, double-blind, controlled trial. Between July and December 2000, 1,113 North American and Brazilian women aged 15–25 were enrolled with an average age of 20. HPV infection was tested using

self-obtained cervicovaginal samples and cervical cytology. After randomization, 560 of 1,113 women received the vaccine and 553 received the placebo. 958 women completed the first phase through month 18, with similar rates of vaccination and placebo dropouts. According-to-protocol HPV-16/18 vaccine effectiveness against the incident and persistent infection was 91.6% and 100%, respectively. Intention-to-treat analysis showed 95.1% efficacy against persistent infection. Neither the vaccine nor the placebo groups experienced any vaccine-related adverse effects. In this trial, the bivalent HPV vaccine proved efficacious, safe, well-tolerated, and highly immunogenic.

## **6.4 Community Trial**

### ***6.4.1 Basic Ideas of Community Trial***

Community trial conducts intervention among healthy people, and the interventions are given at the population level instead of at the individual level. Community trial is used to evaluate the effect of interventions that are not suitable to be given at the individual level. For example, some interventions on dietary factors are easier to be performed at the family level; changing the source of drinking water from the river to tap water is easier to be conducted at the community level. These kinds of interventions are not given individually.

Community trial often uses cluster randomization. The success of cluster randomization depends on the relative sample size within each group compared to the total sample size. If the number of clusters is large, randomization has a higher possibility to be successful. If there are only two communities randomized, the meaning of randomization is limited and the comparability of baseline characteristics of the two communities has a great impact on the results. During the study, investigators need to pay attention to the changes in extraneous factors including mobility, economic changes, medical care conditions, and implementation of other programs in the community.

### ***6.4.2 Case Study***

Research on prevention and control strategies of liver cancer in Qidong and the effect of the community trial

Liver cancer is one of the most common malignant tumors in China, which has a serious impact on people's health. According to a survey from 1990 to 1992, the standardized mortality rate of liver cancer in China was 17.83/100,000 person-years, accounting for about 18.8% of cancer deaths. Nationwide, the mortality rate of liver cancer in the 90s was higher than in the 70s. The incidence of liver cancer increased

after the age of 40 and increases with age, and the age of onset was earlier in high-incidence areas. The male-to-female sex ratio was close to 3:1.

The increase in the incidence of liver cancer may come from the improvement of liver cancer diagnosis, the increase in the proportion of middle-aged and elderly people, and the increase in the incidence of liver cancer caused by the increase of environmental carcinogens. In the early 1970s of the twentieth century, the risk factors for liver cancer were not yet clear, and health workers carried out a large number of investigations and studies in Qidong. The earliest case-control study carried out in 1973 included 100 cases of primary liver cancer, 100 cases of other malignant tumors, and 100 cases of healthy people, and explored the association between liver history, tumor history, pesticide exposure and poisoning history, drinking water source and water quality, tobacco, alcohol and eating habits, family history, and other factors and liver cancer. Patients with hepatitis, liver cirrhosis, and respiratory diseases in Qidong People's Hospital since 1964 were followed up to confirm that patients with liver disease had a high risk of liver cancer. Since 1976, a prospective cohort study has been carried out in Qidong, and long-term follow-up of nearly 15,000 people has been carried out, and the incidence of liver cancer among hepatitis B surface antigen carriers was 361.55/100,000, the incidence rate of non-carriers was 30.90/100,000, and the relative risk was 11.70, confirming the association between hepatitis B virus and liver cancer. The evidence accumulated by years of long-term research suggested and basically clarified that hepatitis B virus, aflatoxin, drinking water pollution, and gene susceptibility were risk factors for liver cancer in this population.

Therefore, the prevention and control strategy of liver cancer in Qidong area was as follows: to carry out intervention research on the suspected causes of liver cancer. By observing changes in the incidence and mortality of liver cancer, the effect of the intervention was evaluated and the etiology was further verified. A range of intervention strategies and specific interventions were identified and implemented. In the early 1970s of the twentieth century, measures of "prevention and control of hepatitis, improvement of drinking water, and prevention of mildew in food" were proposed. Put forward the requirement of "hydration of drinking water wells" to reduce residents' drinking of ditches and river water, and later formed a "deep well tap water supply network" to improve the quality of drinking water; Corn harvesting adopted "fast harvest and quick drying into the warehouse to remove mold", and then changed the staple food to rice, changing the eating habits of residents and reducing the intake of aflatoxin from corn. Various measures have been taken to cut off the transmission of hepatitis B virus and protect susceptible people, and since 1983, large-scale neonatal hepatitis B vaccination has been carried out in Qidong to reduce the epidemic of hepatitis B virus. The academic views and research decisions based on etiology research have been responded to and supported by the government, forming a comprehensive prevention and control strategy on the spot.

Interventions and on-site implementation of major risk factors for liver cancer include:

1. Anti-mildew and Detoxification

As the main chemo-preventive measure, it was important to reduce the intake of food contaminated with aflatoxin by the population. A number of case-control studies and food testing have found a significant association between mildew in corn and the occurrence of liver cancer. Prevention interventions were implemented at two levels: changing the structure of staple foods at the community level to promote the use of rice, with 96.4% of the population switching to rice by 1986; At the individual level, it was promoted to prevent the intake of mildew corn, and preventive measures are taken in the "harvest, storage, and eating" process. This greatly reduced the aflatoxin exposure of Qidong residents.

2. Improve Drinking Water

Based on Qidong's research, Professor Su Delong proposed that the high incidence of liver cancer was related to drinking water pollution. The incidence and mortality of liver cancer among residents with different types of drinking water differed significantly: the incidence of liver cancer in drinking ditch water could be as high as 141.40/100,000, and the incidence of drinking deep well water was 0.23/100,000. Algal toxins, microcystins, and other substances in ditch water are cancer-promoting factors of liver cancer and may interact with aflatoxin. Although there was no direct evidence of carcinogenesis, the drinking water improvement project has solved the problem of drinking water pollution for Qidong residents, and by 2010, 99% of residents were drinking pipe water.

3. Prevention and Treatment of Hepatitis B

HBsAg was screened in blood donors in Qidong, and positive people were not allowed to donate blood, cutting off the transmission route of the virus and reducing the epidemic. A randomized controlled intervention trial of hepatitis B vaccine immunization for the prevention of liver cancer in nearly 80,000 infants between 1984 and 1990 reported a decrease in HBsAg positivity, reporting a 75.9% immune protection rate and a decrease in HBV carrier rate among vaccinated people. After more than 20 years of follow-up in the second phase, vaccination was found to have sustained immunity against chronic HBV infection.

4. Carry Out Research on Early Diagnosis and Early Treatment

For the secondary prevention of liver cancer prevention and treatment, strategies and research on early diagnosis and treatment were carried out in the area. In the first stage, a large-scale screening of alpha-fetoprotein—a biomarker of liver cancer was carried out in 1.8 million people in the 1970s, and a large number of early cases were detected and treated; The second stage was in the 1980s: the high-risk group of liver cancer in Qidong was defined as HBsAg-positive men aged 30–59 years; In the third stage, in the 1990s, periodic screening of high-risk groups was carried out and the screening effect was evaluated. The screening results showed that the early case detection rate of the screening team was high, and the survival rate was higher than that of the control group; In the fourth stage,

Qidong was established as a national sample for early diagnosis and treatment of liver cancer in 2006. Most of the long-term survivors of liver cancer in Qidong were beneficiaries who were found through screening and resected surgically, which shows that screening can detect early cases, and after receiving appropriate treatment, survival can be extended or even cured.

The decrease in morbidity and mortality is the goal of tumor prevention and treatment and an important indicator to test the effect of intervention strategies and measures. After more than 40 years of efforts, the age-standardized incidence and mortality of liver cancer in Qidong have decreased. Although the crude incidence and mortality rate of liver cancer in Qidong have increased in the past 40 years, after controlling for the factors of population growth and the increase in the proportion of the elderly population, the incidence of age-standardized liver cancer decreased from 49.95/100,000 in 1972 to 38.22/100,000 in 1990 and 25.75/100,000 in 2011.

The decline in the incidence and mortality of liver cancer in Qidong was accompanied by significant evidence of changes in risk factors for liver cancer. From 1989 to 2012, the level of aflatoxin adducts representing aflatoxin exposure decreased significantly, from 19.2 pg/mg in 1989 to 2.3 pg/mg in 1999 and undetectable in 2009. The drinking water of residents was changed from the easily polluted house ditch water and the water from the Yangtze River to tap water from deep wells and the Yangtze River, and the quality of drinking water was significantly improved. After 2002, the vaccination rate of hepatitis B among newborns reached 100%, the short-term and medium-term efficacy of the vaccine was confirmed, and the long-term effect and association with the decline in the incidence of liver cancer have yet to be confirmed by long-term follow-up. The above facts and evidence from changes in biomarkers, ecological changes, and changes in population immunoprevention show that even if the mechanism of action of some risk factors for liver cancer has yet to be elucidated, after controlling these risk factors, the incidence and mortality of liver cancer in the population have indeed decreased significantly, which is enough to prove that these preventive measures are effective.



# Chapter 7

## Screening and Diagnostic Tests



Fen Liu

### Key Points

- Screening is the process of using quick and simple tests to identify and separate persons who have an illness from apparently healthy people.
- The validity of a screening test is defined by its ability to correctly categorize subjects who do or do not have a disease into corresponding groups. The components of validity include sensitivity, specificity, Youden's index, and likelihood ratio. Reliability is an index that reflects the stability of the testing results. That includes agreement rate and Kappa statistic. The PPV is defined as the probability of the persons having the disease when the test is positive. The NPV is the percentage of the persons not having the disease when the test is negative. The position of the cutoff point for a screening test will determine the number of true positives, false positives, false negatives, and true negatives. For continuous measurement data of a screening test, the cutoff point is determined mostly by the ROC curve.
- Screening the high-risk population or performing multiple tests increased the validity of a screening test.
- Volunteer bias, lead-time bias, and length-time bias are three major sources of bias in screening test.

Screening is an effective strategy for early detection of diseases and is considered a secondary prevention program in public health; diagnostic tests are helpful in confirming diagnoses of diseases and can help the doctors determine the therapeutic plans for patients. Along with the progress in science and technology, novel

---

F. Liu (✉)

School of Public Health, Capital Medical University, Beijing, China

e-mail: [liufen05@ccmu.edu.cn](mailto:liufen05@ccmu.edu.cn)

© Zhengzhou University Press 2023

C. Wang, F. Liu (eds.), *Textbook of Clinical Epidemiology*,

[https://doi.org/10.1007/978-981-99-3622-9\\_7](https://doi.org/10.1007/978-981-99-3622-9_7)

screening and diagnostic tests are continuously put forward. Thus, the quality of screening and diagnostic tests is a critical issue. In this chapter, we will address the questions on how to assess the quality of various screening and diagnostic methods, in particular, the newly available ones, and how to make reasonable decisions on their application.

## **7.1 Design a Screening or Diagnostic Test**

Although the purpose, observational subjects, and requirement of screening and diagnostic tests are different, the principle for the evaluation of these two types of tests is similar. Therefore, we take a screening test assessment as an example to discuss.

Screening is the process of using quick and simple tests to identify and separate persons who have an illness from apparently healthy people. To evaluate a new screening test, we need to compare the results of the test to that of a standard test, which is called the “gold standard” via using the blinding method.

### ***7.1.1 Gold Standard (Reference Standard)***

A “gold standard” method refers to the most reliable method to diagnose a disease, which is also referred to as standard diagnosis. Application of gold standard can distinguish whether the disease is truly present or not. The gold standard can be biopsy followed by pathological examination, surgical discovery, bacteria cultivation, autopsy, special examination, and imaging diagnosis; it also can be an integrated combination of several diagnostic criteria (such as Jones diagnosis standard, etc.). The outcomes of long-term clinical follow-up obtained by applying the affirming diagnostic methods were also used for the gold standard.

### ***7.1.2 Study Subjects***

The subjects of a screening test include the case group who has a specific disease and controls who do not have the disease. They should be representative of the target population. Therefore, the case group should include various types of the studied disease: mild, moderate, or severe; early, middle, or late stage; typical or atypical; with or without complication; treated or untreated, in order to make the result of the study more representative and applicable to the general population. In contrast, the control group should include individuals without the studied disease, but with other illnesses, particularly those that are not easily distinguishable from the studied disease. The testing of study subjects should be kept within the same research period

through either continuous sampling or proportional sampling, rather than by the researchers' choice. Otherwise, a selection bias may be present, which influences the validity and reproducibility of the test.

### 7.1.3 Sample Size

The sample size is determined based on the following factors: sensitivity, specificity, permissible error, and alpha level. The formula for sample size calculation is as follows:

$$n = \frac{Z_{\alpha}^2 p(1-p)}{\delta^2} \quad (7.1)$$

$n$  is the sample size of abnormal or normal subjects in the study.

$Z_{\alpha}$  is the  $Z$  value for normal distribution of cumulative probability, which is equal to  $\alpha/2$ .

$\delta$  is admissible error, usually, it is set at a 0.05 ~ 0.10 level.

$p$  is the estimation of sensitivity or specificity of the test. Sensitivity is used to calculate the sample size of the case group, while specificity is used for the control group. This formula requires the sensitivity or specificity approaching 50%. When the sensitivity or specificity  $\leq 20\%$  or  $\geq 80\%$ , the corrected formula is needed:

$$n = \left[ \frac{57.3Z_{\alpha}}{\sin^{-1}(\delta/\sqrt{p(1-p)})} \right]^2 \quad (7.2)$$

## 7.2 Evaluation of a Screening Test

When evaluating a new screening test for a disease, the gold standard for the disease should be used simultaneously. The subjects will be divided into two groups based on the test results: case group and control group (non-disease group). The results of the gold standard and the screening test are then compared. The first step of this comparison is to generate a two-by-two table and calculate several indexes.

As shown in Table 7.1, in cell  $a$ , the disease of interest is present, and the screening test result is positive, a true-positive result. In cell  $d$ , the disease is absent, and the screening test result is negative, a true-negative result. In both  $a$  and  $d$  cells, the screening test result agrees with the actual status of the disease. Cell  $b$  represents individuals without the disease who have a positive screening test result. Since these test results incorrectly suggest that the disease is present, they are considered to be false positives. Subjects in cell  $c$  have the disease but have negative screening test

**Table 7.1** Comparison of the results of a screening test with the gold standard

Screening test	Gold standard		Total
	Patients	Controls	
Positive	True positive ( $a$ )	False positive ( $b$ )	$a + b$
Negative	False negative ( $c$ )	True negative ( $d$ )	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

results. These results are designated false negatives because they incorrectly suggest that the disease is absent.

### 7.2.1 Validity of a Screening Test

The validity of a screening test is defined by its ability to correctly categorize subjects who do or do not have a disease into corresponding groups. The components of validity include sensitivity, specificity, Youden's index, and likelihood ratio.

#### 7.2.1.1 Sensitivity

The sensitivity of a screening test is defined as the proportion of persons with the disease in the screened population who are identified as ill by the test. Sensitivity is calculated as follows:

$$\text{Sensitivity}(\text{Sen}) = \frac{a}{a + c} \times 100\% \quad (7.3)$$

If someone with the disease is incorrectly called "negative," it is a false-negative result. The false-negative rate is complementary to sensitivity.

#### 7.2.1.2 Specificity

Specificity of a test is defined as the proportion of disease-free people who are so identified by the screening test. Specificity is calculated as follows:

$$\text{Specificity}(\text{Spe}) = \frac{d}{d + b} \times 100\% \quad (7.4)$$

If some people without a disease are incorrectly called "positive," it is a false-positive result. The rate of false-positive is complementary to the specificity.

### 7.2.1.3 Youden's Index

Youden's index ( $YI$ ) is also called the accuracy index, which is frequently used to evaluate the overall performance of a test. The formula of the Youden's index is:

$$YI = Sen + Spe - 1 \quad (7.5)$$

It ranges from 0 to 1. The greater the index is, the better the validity.

### 7.2.1.4 Likelihood Ratio

The likelihood ratio ( $LR$ ) reflects the validity of screening test; it is an integrative index that can reflect the sensitivity and specificity altogether, i.e., the ratio of true-positive or false-negative rates in disease group to the false-positive or true-negative rates in the group without the disease. Using the results of the screening tests, we can calculate all the  $LR$  of the tests, which thus reflect the overall validity of a screening test.

The positive likelihood ratio of a screening test is the ratio of true-positive rate to false-positive rate, and negative likelihood ratio is a ratio of false-negative rate to true-negative rate. The computation formulas for positive likelihood and negative likelihood ratios are as follows:

$$LR^+ = \frac{a/(a+c)}{b/(b+d)} = \frac{Sen}{1-Spe} \quad (7.6)$$

$$LR^- = \frac{c/(a+c)}{d/(b+d)} = \frac{1-Sen}{Spe} \quad (7.7)$$

The likelihood ratio is more stable than sensitivity and specificity, and it is less influenced by prevalence.

There is an example that would be helpful in understanding the calculation of these indices.

**Example** Suppose, we perform a diabetes screening test in a cohort of 1000 people, of whom 20 are diabetic patients and 980 are not. A test is available that can yield either positive or negative results. We want to use this test to distinguish subjects who have diabetes from those who do not. The results are shown in Table 7.2. How do we evaluate the validity of the screening test?

These results showed that of the study population, 90% were positive in the screening test, but the remaining 10% were not diagnosed. Among the individuals without diabetes, 95% tested negative with the screening, and 5% were misdiagnosed in the screening.

**Table 7.2** The results of a screening test and the gold standard test for diabetes

Results of screening	Gold standard		Total
	Have the disease	Don't have the disease	
Positive	18	49	67
Negative	2	931	933
Total	20	980	1000

$$\text{Sensitivity} = (18/20) \times 100\% = 90\%$$

$$\text{Specificity} = (931/980) \times 100\% = 95\%$$

$$\text{False-negative rate} = (2/20) \times 100\% = 10\%, \text{ or } 1 - 90\% = 10\%$$

$$\text{False-positive rate} = (49/980) \times 100\% = 5\%, \text{ or } 1 - 95\% = 5\%$$

$$\text{Youden's index} = 0.90 + 0.95 - 1 = 0.85$$

$$LR^+ = 0.90/0.05 = 18.00$$

$$LR^- = 0.10/0.95 = 0.11$$

## 7.2.2 Evaluation of the Reliability of a Test

Reliability or repeatability is an index that reflects the stability of the testing results, i.e., if the results are replicable when the test is repeated. In a study, almost all variations of measured data stem from the observer's variation (intra-observer and inter-observer variation), measuring instruments, reagents variation, and research object's biological variation (intra-subject variations), etc.

### 7.2.2.1 Coefficient of Variation

For a continuous variable, the variations of data are commonly measured with standard deviation (*SD*) and coefficient of variation. The coefficient of variation (*CV*) is obtained by dividing the *SD* by mean (percentage).

$$CV = \left( \frac{SD}{\bar{X}} \right) \times 100\% \quad (7.8)$$

### 7.2.2.2 Agreement Rate and Kappa Statistic

Agreement (consistency) rate is also called accuracy rate, which is defined as the proportion of the combined true positive and true negative number of the total population evaluated by a screening test, i.e., the percentage of the results of a screening test that is in accordance with those of the gold standard method. Below is the formula for calculating accuracy rate:

$$\text{Agreement rate} = [(a + d)/(a + b + c + d)] \times 100\% \quad (7.9)$$

**Table 7.3** Kappa value judgment standard

Kappa value	Consistency strength
<0	Poor
0 ~ 0.2	Weak
0.21 ~ 0.40	Light
0.41~0.60	Moderate
0.61 ~ 0.80	High
0.81 ~ 1.00	Strong

For counted variable, the observation coincidence rate or kappa statistic is used to determine data reliability (repeatability or precision).

This is the calculation of kappa:

$$Kappa = \frac{\left( \begin{array}{c} \text{Percent} \\ \text{agreement} \\ \text{observed} \end{array} \right) - \left( \begin{array}{c} \text{Percent} \\ \text{agreement} \\ \text{expected} \\ \text{by} \\ \text{chance} \\ \text{alone} \end{array} \right)}{100\% - \left( \begin{array}{c} \text{Percent} \\ \text{agreement} \\ \text{expected} \\ \text{by} \\ \text{chance} \\ \text{alone} \end{array} \right)} \quad (7.10)$$

Kappa is an index that judges consistency in levels between different observers. Landis and Koch suggested that kappa greater than 0.75 represents an excellent agreement beyond chance, while a kappa less than 0.40 shows poor agreement, and a kappa of 0.40 to 0.75 represents intermediate to good agreement (Table 7.3). Testing for the statistical significance of kappa, please refer to the relevant book.

### 7.2.3 Predictive Value

Sensitivity and specificity are indicators of the accuracy of a test, which can be considered the characteristics of a screening or diagnostic test itself. However, the predictive value is affected by both the sensitivity and specificity of the test and the prevalence of the disease in the population to be tested. There are positive predictive value (*PPV* or *PV+*) and negative predictive value (*NPV* or *PV-*).

The *PPV* is defined as the probability of the persons having the disease when the test is positive. The *PPV* is calculated as follows:

$$PPV = \frac{a}{a + b} \times 100\% \quad (7.11)$$

The *NPV* is the percentage of the persons not having the disease when the test is negative.

$$NPV = \frac{d}{c + d} \times 100\% \quad (7.12)$$

Take the data in Table 7.2 as an example again for the calculation of predictive values:

$$PPV = \frac{18}{18 + 49} \times 100\% = 26.87\%$$

$$NPV = \frac{931}{2 + 931} \times 100\% = 99.79\%$$

The *PPV* of 26.87% means that 67 individuals are positive in screening, but among them, the number of real patients is 18, accounting for 26.87% of the total positive results. The *NPV* of 99.79% indicates that 933 persons have negative test results, and among them, the number of individuals “not having the disease” is 931, accounting for 99.79% of the total negative results.

Predictive value is affected by the prevalence of a disease in a specific population, or by the pretest probability of the presence of a disease in an individual. We can use the formula derived from Bayesian theorem of conditional probability to show the relationships of predictive value, sensitivity, specificity, and prevalence.

$$PPV = \frac{\text{Sensitivity} \times \text{Prevalence}}{\text{Sensitivity} \times \text{Prevalence} + (1 - \text{Specificity}) \times (1 - \text{Prevalence})} \quad (7.13)$$

$$NPV = \frac{\text{Specificity} \times (1 - \text{Prevalence})}{(1 - \text{Sensitivity}) \times \text{Prevalence} + \text{Specificity} \times (1 - \text{Prevalence})} \quad (7.14)$$

The more sensitive a test is, the higher will be its negative predictive value (the more confident clinicians can be that a negative test result rules out the disease being sought). Conversely, the more specific the test is, the better will be its positive predictive value (the more confident clinicians can be that a positive test confirms or rules in the diagnosis being sought). Because predictive value is also influenced by prevalence, it is not independent of the setting in which the test is used.

As the numbers in Table 7.4 show, positive results even for a very specific test, when applied to patients with a low likelihood of having the disease, will be largely false positives. Similarly, negative results, even for a very sensitive test, when applied to patients with a high chance of having the disease, are likely to be false negatives. In summary, the interpretation of a positive or negative result of a



**Table 7.4** The screening results of diabetes in populations with different values of sensitivity, specificity, and prevalence

Prevalence (%)	Sensitivity (%)	Specificity (%)	Screening results	Gold standard		Total	PPV (%)	NPV (%)
				Patients	Non-patients			
50	50	50	+	250	250	500	50	50
			-	250	250	500		
			Total	500	500	1000		
20	50	50	+	100	400	500	20	80
			-	100	400	500		
			Total	200	800	1000		
20	90	50	+	180	400	580	31	95
			-	20	400	420		
			Total	200	800	1000		
20	50	90	+	100	80	180	56	88
			-	100	720	820		
			Total	200	800	1000		

screening or diagnostic test is dependent on the setting in which the test is carried out, in particular, the estimated prevalence of the disease in the target population.

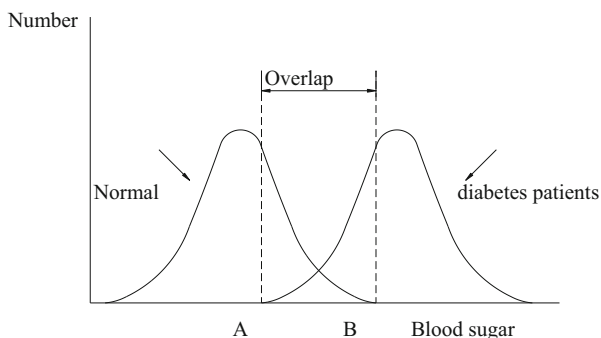
### 7.2.4 Determination of Cutoff Point for a Screening Test

Ideally, the sensitivity and specificity of a screening test both should be 100%. In practice, when we plot the value of a screening test for a disease group and non-disease group on the same graph, the distribution often overlaps, the test does not separate normal from diseased with 100% accuracy. Figure 7.1 is the schematic graph showing the distributions of test results for patients with and without the disease. The area of overlap indicates where the test cannot distinguish normal and abnormal. We need to determine a balance by an arbitrary cutoff point (indicated by A and B) between normal and disease. The position of the cutoff point will determine the number of true positives, false positives, false negatives, and true negatives. If we want to increase sensitivity and include all true positives, we can use A as a cutoff point, but by doing this, we increase the number of false positives, which means decreased specificity. Likewise, if we want to increase specificity by using B as a cutoff point, it will lead to decreased sensitivity.

We can also use the blood sugar data in Table 7.5 as an example to illustrate how changes in the cutoff point will affect the sensitivity and specificity of a screening test.

To make decisions on the appropriate cutoff point for a screening test, the following principles need to be taken into consideration. For a proven serious disease that can be cured if diagnosed early, a high sensitivity may be suggested. If a false-positive result would detrimentally affect a patient both mentally and physically, such as cancers, which may put a patient at risk of surgery and chemotherapy, a test with high specificity would be required. If both the sensitivity and specificity are important, the junction point of curves might be used as the cutoff point.

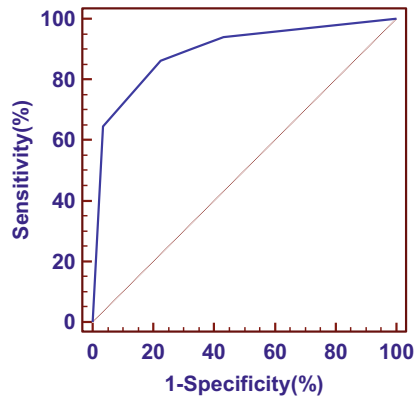
**Fig. 7.1** Blood sugar level distribution in normal people and diabetes patients



**Table 7.5** The effects of cut-off points of 2 h after-meal blood sugar on sensitivity and specificity of the screening test

Blood sugar (mg dL <sup>-1</sup> )	Sensitivity (%)	Specificity (%)
80	100.0	1.2
90	98.6	7.3
100	97.1	25.3
110	92.9	48.4
120	88.6	68.2
130	81.4	82.4
140	74.3	91.2
150	64.3	96.1
160	55.7	98.6
170	52.9	99.6
180	50.0	99.8
190	44.3	99.8

**Fig. 7.2** The ROC curve of blood sugar in the diabetes diagnosis



**7.2.4.1 ROC Curve**

For continuous measurement data of a screening test, the cutoff point is determined mostly by the receiver operator characteristic (ROC) curve. ROC curve is a graphical plot of true positive rate (sensitivity, Y-axis) against the false negative rate (1 – specificity, X-axis) for different cutoff point. A ROC curve could reflect the relationship between the sensitivity and the specificity of a test (Fig. 7.2). By convention, the point nearest to the top-left corner of the ROC curve is set for optimal cutoff point.

As shown in Fig. 7.2 and Table 7.5, when sensitivity is 88% and specificity is 68%, the sum of the false positive and false negative rates is the minimum. Accordingly, the blood sugar level of 120 mg dL<sup>-1</sup> can be set as the optimal cutoff point for diabetes screening in this population.

### **7.2.4.2 The Area under ROC Curve**

ROC curves can also be used to compare clinical values of two or more screening tests, thus helping clinicians choose the best screening test. The area under the ROC curve is a measure of the test's accuracy. The larger the area under the ROC curve, the better the diagnostic test. The maximum value for the area under the ROC curve is 1, which indicates a perfect test; an area of 0.5, on the other hand, represents a worthless test.

We can use statistical software, such as MedCalc, SPSS, and SAS, to compute the area under the ROC curve and compare the areas under ROC curve between two or more screening tests (for details, please refer to related statistics books).

## **7.3 Improving the Efficiency of Screening and Diagnostic Tests**

In order to increase the sensitivity and specificity of a screening test, several methods can be used, such as screening high-risk population or performing multiple tests.

### ***7.3.1 Selecting Population with a High Prevalence***

The predictive value of a test is influenced by the sensitivity, specificity, and prevalence of a disease. When sensitivity and specificity are constant, it is influenced mainly by the prevalence rate. Since morbidity has larger influence on the positive predictive value, the latter would have very low value if a screening test is carried out in a population with a low prevalence rate of the disease to be tested. However, if a high-risk population is screened, the positive predictive value can be significantly increased.

### ***7.3.2 Use of Multiple Tests***

A method combining two or more tests is called multiple tests. In general, multiple tests can be carried out in two ways, simultaneous testing and sequential testing.

#### **7.3.2.1 Simultaneous Testing**

In simultaneous testing (parallel tests), the sample is evaluated with more than one screening test simultaneously; a positive result of any test is considered evidence for

**Table 7.6** The results of simultaneous and sequential testing

Multiple tests	Test results		Diagnosis
	Test A	Test B	
Simultaneous testing	+	–	+
	–	+	+
	+	+	+
	–	–	–
Sequential testing	+	+	+
	+	–	–
	–	+	–
	–	–	–

**Table 7.7** An example of screening results using multiple tests (%)

Screening methods	Sensitivity	Specificity	PPV	NPV
Test A	80	60	33	92
Test B	90	90	69	97
Simultaneous testing (A and B)	98	54	35	99
Sequential testing (A and B)	72	96	82	93

the target disease. Simultaneous testing can improve sensitivity and negative predictive value, but lower the specificity and positive predictive value (Table 7.6).

### 7.3.2.2 Sequential Testing

Sequential testing (serial testing) means multiple screening tests are used in series, the individual is considered to be positive if all the test results are positive but is stopped when the previous test result is negative. Sequential testing increases specificity and positive predictive value but decreases sensitivity and negative predictive value.

Take the hypothetical example in Table 7.7 as an example, in which a population is screened for hepatocellular carcinoma using ultrasonography and serum alpha-fetoprotein (AFP) level. If two tests with 80% and 90% sensitivity, respectively, were used simultaneously, the sensitivity of the simultaneous testing will be increased up to 98%. However, there is a loss of specificity (decreased to 70%) compared to each test alone. In sequential testing, there is a gain in specificity (increased up to 96%), but a loss in sensitivity (down to 72%).

From the results above, we can summarize the regular pattern of sensitivity and specificity in different multiple tests. How to make the decision to choose either simultaneous or sequential testing is based on the actual situation.

## 7.4 Potential Bias in Screening Tests

There are three major sources of bias, which are specified to each screening test.

### 7.4.1 *Volunteer Bias*

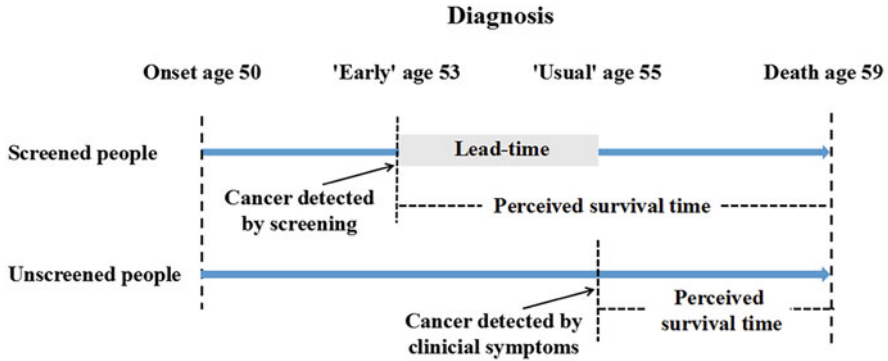
The characteristics may be different between people who attend a screening and those who do not, especially when those factors are directly related to the survival of patients. Individuals with a higher risk of a disease are more likely to voluntarily join a screening program, as they might be more health-conscious and with higher compliance tend to have a better prognosis. For example, women with a significant family history of breast cancer are more likely to join a mammography program than those without it. This tendency is reflected by a higher rate of diagnosis in a series of screening tests than what is truly reflective of the population. Likewise, the screened people tend to have a larger percentage of adverse clinical outcomes than it would be in the general population.

The most effective way to avoid volunteer bias is to recruit a pool of volunteers and then assign them randomly to receive screening or not to receive it.

### 7.4.2 *Lead-Time Bias*

Lead time refers to the duration from early detection of disease (usually by screening) to the presentation of clinical symptoms and thus being diagnosed in the standard way. Especially for chronic diseases, the cases of which progress slowly, therefore patients with those diseases are more likely to be detected by screening and likely to have increased survival time than unscreened cases. In fact, the screening has no effect on the outcome of the disease; it only resulted in an earlier diagnosis of the disease when compared to traditional diagnostic methods. To illustrate the lead-time bias, we take a cancer screening test (shown below) as an example. As shown in the illustration, the tumor is detected at different ages with or without the screening test, but the patients die at the same age (Fig. 7.3), indicating that the overall survival of patients is not altered by the screening test.

So, unless we have some idea of the actual lead-time, perhaps from previous studies, we should not use survival time after diagnosis to evaluate a screening program. Instead, we should consider the effects on longer-term age-specific morbidity or mortality rates of the disease. The survival rates are therefore less likely to reflect the true benefits of early treatment better.



**Fig. 7.3** An illustration of lead-time associated with screening and cancer development process

### 7.4.3 Length-Time Bias

Many screening programs are implemented to detect cancers. Doctors and researchers hypothesize that tumors with low growth rates have better outcomes than more aggressive types. However, it is found that screening is more likely to detect slower-growing, less deadly tumors due to their longer preclinical stages. In other words, patients with concealed, less fatal cancers may not know the fact before their death from other diseases, if without screening. This example results in the “length-time” bias associated with screening tests, which gives the appearance that screening can benefit patients and prolong their life span, when, in fact, the test selectively detects those diseases that progress slowly, thus allowing the patient to live longer.

# Chapter 8

## Bias



Lu Long

### Key Points

- Bias refers to various influencing factors in epidemiological research, including design, implementation, analysis, and inference, also known as systematic error. Three major threats to validity are selection bias, information bias, and confounding bias.
- Selection bias occurs when the characteristics of the subjects are different from the source population, which leads to the deviation of the research results from the real situation.
- Information bias, known as observational bias, refers to the inaccuracy or incompleteness of the exposure or outcome information obtained during the implementation of the research, which results in the misclassification of the exposure or disease of the research subjects and affects the validity of the results.
- Confounding bias is due to the existence of one or more external factors that mask or exaggerate the link between research factors and diseases, thus partially or wholly distorting the actual association.

### 8.1 Introduction of Bias

Bias, also known as systematic error, refers to various influencing factors in epidemiological research, including design, conduct, analysis, and inference.

The existence of these influencing factors, including design errors, data acquisition distortion, incorrect analysis, or not logical inference, leading to the association between exposure and outcome is misestimated, and this actual relationship is systematically distorted, which leads to the wrong conclusion. Bias is an important issue that affects the authenticity of the results. Thus, we must fully understand the

---

L. Long (✉)

West China School of Public Health, Sichuan University, Chengdu, China



source of bias and its causes and minimize the occurrence of bias in our studies to ensure the authenticity of the study. There are two directions of bias, i.e., positive bias and negative bias. Positive bias means that the measured value of the study overestimates the true value, on the contrary, it is negative bias.

We generally classify bias as selection bias, information bias, and confounding bias.

## 8.2 Selection Bias

### 8.2.1 Definition

Selection bias means that the characteristics of the selected subjects are different from those of the unselected, which leads to a deviation of the research results from the real situation.

### 8.2.2 Classification

#### 8.2.2.1 Self-Selection Bias

Self-selection bias, or volunteer bias, is one source of selection bias. Self-selection bias is one type of bias that results from individuals disproportionately selecting themselves to join a group. For example, researchers selected soldiers from the Smoky Atomic Test in Nevada to investigate the leukemia incidence. In this study, 76% of the soldiers are members of the cohort with known outcomes, and the remaining 24% were identified as the cohort without known outcomes. Those who knew the outcomes, 82% were traced by the investigators, while others reached out to surveyors. We ordinarily consider self-reported subjects as a threat to validity because self-reporting may be related to the study results.

In the Smoky Atomic Test study, among the 62% ( $2\% \times 76\%$ ) of cohort members, investigators traced four target cases and the 14% ( $18\% \times 76\%$ ) of cohort members also traced four target cases who reported themselves. We assume the leukemia incidence without known outcomes (24%) is similar to that of the subjects traced by the investigators. In that case, we should expect that only  $(24\%/62\%) \times 4 = 1.5$ , meaning that about one or two cases occurred among this 24% of the members without known outcomes, only a total of nine or ten cases in the entire cohort. If instead, we suppose that the 24% without known outcomes had the same incidence of leukemia as subjects with known outcomes. We would calculate that  $8(24\%/76\%) = 2.5$ , meaning that about two or three cases occurred among this 24%, in the entire cohort we will observe 10 or 11 cases. However, among the 24% + 14% of the cohort, all cases were untraced among the self-reported, leaving no case among those without known outcome. T. The total number of cases will be only

8 in the entire cohort. This example indicates that self-selection bias is a small but a real problem in research.

**8.2.2.2 Berksonian Bias**

Berkson’s bias or Berksonian bias is also known as admission rate bias. It usually occurs in a hospital-based case-control study because the selected case or controls represent only a subset of patients with a disease rather than an unbiased sample of the corresponding target population. Affected by medical conditions, residence, socio-economy, education, and other factors, patients have specific selectivity to hospitals, and hospitals also have a specific selectivity to patients, which results in problems in sample representativeness and bias in a hospital-based case-control study.

For example, hospital-based case-control study was used to explore the relationship between birth control pills and thrombophlebitis. Cases were recruited from people with thrombophlebitis in a hospital. And randomly selected as patients without thrombophlebitis in a certain ward of the same hospital as a control group. Suppose there are 5000 patients with thrombophlebitis and 5000 patients without thrombophlebitis. Oral contraceptive accounts for 15% in each of them. It is assumed that admission rates for these three conditions are relatively independent (Table 8.1).

It can be calculated from Table 8.1 that the correlation of thrombophlebitis and oral contraceptive,  $OR = (750 \times 4250)/(4250 \times 750) = 1.0$ , which indicates that there is no correlation among oral contraceptive and thrombophlebitis.

Now assume the admission rate of case group was 25% while control group was 60%, and the admission rate of oral contraceptive was 40%. The composition of the comparative study samples is shown in Table 8.2.

The admission rate of the 750 patients with thrombophlebitis and exposure to contraceptive was 25%, So the number of thrombophlebitis hospitalizations were  $750 \times 25\% = 187.5 \approx 188$ ; and 40% of the remaining were hospitalized due to exposure to contraceptive, and the number of hospitalized patients was  $(750 - 750 \times 25\%) \times 40\% = 225$ , and the total hospitalizations was 413.

**Table 8.1** Exposure and disease distribution in the total population

Group	Oral contraceptive		Total
	Yes	No	
Case	750	4250	5000
Control	750	4250	5000

**Table 8.2** Distribution of exposure and disease in the hospital

Group	Oral contraceptive		Total
	Yes	No	
Case	413	1063	1476
Control	570	2550	3120

The admission rate of the 4250 patients with thrombophlebitis rather than exposure to contraceptive was 25%, So the number of cases group was  $4250 \times 25\% = 1062.5 \approx 1063$ .

The admission rate of the 750 patients without thrombophlebitis who were exposed to contraceptives was 60%, so the hospitalizations were  $750 \times 60\% = 450$ , and 40% of the remaining patients were hospitalized because of exposure to contraceptives, the hospitalizations was  $(750 - 750 \times 60\%) \times 40\% = 120$ , with a total hospitalization of 570.

The admission rate of the 4250 patients without thrombophlebitis and the oral contraceptives was 60%, So the number of total hospitalizations was  $4250 \times 60\% = 2550$ .

According to the above data,  $OR = (2250 \times 413)/(570 \times 1063) = 1.53$ , oral contraceptive was positively correlated with thrombophlebitis.

There was no association between oral contraceptives and thrombophlebitis in the total population, but a case-control study using hospital samples found a positive correlation. The degree of the association was influenced by the admission rate, which deviated from the true association in the population. This is Berksonian Bias.

### 8.2.2.3 Detection Signal Bias

Detection signal bias, known as unmasking bias, is also a common selection bias. If the exposure factor to be studied has no tural causal relationship to the disease, however, its presence may cause the subject to develop symptoms or sighs related to the disease to be studied, leading to earlier or more frequent visits to the doctor, which increases the detection rate of the disease and makes it more likely to be included as a case in the study. Suppose these patients are taken as case groups in case-control studies. In those cases, there will be systematic differences in certain characteristics (such as exposure factors) between admitted patients and non-admitted patients, leading to misestimating the true associations between exposure factors and outcomes. For example, several studies found that oral estrogen was associated with endometrial cancer and believed that oral estrogen was a risk factor for endometrial cancer. However, many scholars later proposed that estrogens do not cause cancer to occur, but only allow cancer to be diagnosed. Because estrogen can stimulate the growth of the endometrium, making the uterus prone to bleeding. The women who take estrogen are more likely to seek medical attention, this made early-stage endometrial cancer patients easier to be identified. In contrast, while case-control studies with such patients as case group led to an increased proportion of oral androgens in endometrial cancer patients, thereby overestimating the association between estrogen and endometrial cancer.

#### **8.2.2.4 Neyman Bias**

Neyman bias, called prevalence-incidence bias, was first described by Neyman in 1955 and occurred in the case-control study design. When carrying out case-control studies, we can select three cases: cases-incident cases, prevalent cases, and death cases. If all the admitted cases are survived cases, especially the cases with a long disease course, may be related to the survival, but not to the onset of the disease. It may, thus misestimate the etiological effect of these factors. On the other hand, survivors of disease may change some of their existing exposure. When they are investigated, they may mistake these changed exposure characteristics as their disease conditions, resulting in errors in the correlation between these factors and the disease.

#### **8.2.2.5 Loss of Follow-Up**

Cohort studies, clinical trials, and clinical prognosis studies generally require follow-up of subjects. For the long observation period, the follow-up process cannot avoid the absence of outcome events due to relocation of subjects, death due to other reasons (competitive risk), or withdrawal from the study due to poor treatment effects, adverse reactions, and other reasons. Loss of follow-up will affect the representativeness of the research objects, thus affecting the authenticity of the results. Therefore, this bias is called loss of follow-up bias.

### **8.2.3 Control**

It is difficult to eliminate or correct its effects on the results once select bias occurs. Therefore, scientific research design should be performed to reduce and avoid such bias.

#### **8.2.3.1 Scientific Research Design**

In the research process, we(researchers) should clear the global and the sample population and predict the various bias that may be generated in the sample selection process based on the nature of the study. In the case-control study, we should avoid selecting cases in a single hospital, and we can set up community control and hospital control at the same time. Even if the cases can only be selected from the hospital, they should also be randomly sampled in the different areas and different levels of hospitals. In the cohort study, we can establish various controls, including comparing incidence in exposed populations and all populations or compare incidence in exposed populations and other unexposed populations, to reduce the effects of selective bias.

### **8.2.3.2 Develop Strict Inclusion and Exclusion Standards**

In both observational research and experimental research, we must have developed a strict, clear unified standard about inclusion and exclusion, including disease diagnostic criteria and exposure criteria, enabling the selected research object to better represent the overall. After the exclusion standard determines the selection, it strictly complies with the study's implementation phase and cannot be changed casually.

### **8.2.3.3 Maximize Response Rates**

Various measures should be taken to obtain the cooperation of the subjects as far as possible, improve the response rate, reduce or prevent the occurrence of loss of follow-up, and control the selection bias. During the study, we should increase the subjects' understanding of the significance of the study through various ways. When the non-response rate or loss of follow-up rate is more than 10%, we should be cautious in analyzing the research results. A random sampling survey should be conducted on the non-responders or lost respondents if possible, and the results of the sampling survey should be compared with those responders. If there is no significant difference, it shows that the non-response or loss of follow-up has little effect on the results; oppositely, we should explain appropriately. Strategies to reduce loss to follow-up include: screening of willingness prior to registration, detailed collection of participants' contact information, using effective incentives, and maintaining regular contact with participants. In addition, the sample size can be appropriately increased to reduce the impact of the loss of follow-up or non-response on the results after the corresponding sample size is calculated in the design stage.

### **8.2.3.4 Randomization Principle**

Randomization can be divided into two different forms of random sampling and random allocation. Random sampling means the opportunity of each target object extracted into the study queue is equal, making the research sample representative, avoiding bias due to the subjective, arbitrary choice of research objects; random allocation is the equivalent opportunity for participants to be assigned to the experimental group or control group without the effect of researchers and participants' subjective wishes or unconscious objective reasons. The purpose of random distribution is to make the non-research factors evenly distributed in each group and to increase the transferability among groups.

## 8.3 Information Bias

### 8.3.1 *Definition*

Information bias, known as observational bias, refers to the inaccuracy or incompleteness of the exposure or disease information obtained during the implementation of the research, which results in the wrong classification of the exposure or outcome of the research subjects and affects the authenticity of the results. Information bias generally occurs when there are errors in the measurement, which is also known as classification error or misclassification for discrete variables. Misclassification can divide into differential misclassification and nondifferential misclassification. Compared with nondifferential misclassification, differential misclassification has a greater impact on study results. Due to the directions of differences in the misclassification among groups, the effect value may be overestimated or underestimated.

### 8.3.2 *Classification*

#### 8.3.2.1 Differential Misclassification

Differential misclassification refers to classification errors that rely on the actual values of other variables. The most common differential misclassification is recall bias. Suppose an interview of congenital malformations in a case-control study, we generally obtain the etiological information from the mother. We selected mothers who have recently given birth to a deformed baby as a case, whereas mothers who had recently given birth to an apparently healthy baby as a control. The mothers of deformed infants are better able to recall exposures than mothers of healthy infants, leading to a kind of differential misclassification, referred to as recall bias. Because the birth of a deformed infant can stimulate the mother to recall all events that may have played some role in the unfortunate outcome. The difference produced by this recall bias is an apparent effect unrelated to any biological effect. Recall bias is likely to arise in any case-control study that requires recall of past experiences. Klemetti and Saxen [9] considered time as a critical indicator of recall accuracy.

When establishing or verifying a research hypothesis, if personal biased views are reflected in the process of data collection, it will lead to interviewer bias. The resulting inducement bias is also classified as interviewer bias if the researcher intentionally induces the subject to provide the required information. In cohort studies or experimental epidemiological studies, more detailed examination of exposure or intervention group may be performed if the investigator has previously assumed that the exposure or intervention is associated with the occurrence of outcome. It leads to a misjudgment of the study results.

Not all misclassification will exaggerate the association under study, but examples of the opposite can also be found. When investigating sensitive issues with the subjects, they will deliberately minimize the information. For example, patients with sexually transmitted diseases such as syphilis and gonorrhea may be reluctant to let investigators know about their history of exposure to unprotected sex because of stigma, and the resulting bias may underestimate the association between unprotected sex and sexually transmitted diseases.

### 8.3.2.2 Nondifferential Misclassification

Classification error that is independent of other variables is called nondifferential misclassification.

Presumably, bias due to independent nondifferential misclassification of exposure or disease is predictable in the direction, i.e., toward the null. Some researchers have used complex procedures to demonstrate that misclassification is nondifferential. Unfortunately, decomposing continuous or categorical data into fewer categories can transform non-differential errors into differential misclassifications even under blinding is accomplished or in cohort studies where disease outcomes have not yet emerged. Non-differentially alone does not guarantee a bias toward the null. Even if nondifferential misclassification is implemented, it may come at the cost of increasing the total bias.

Both disease and exposure can occur nondifference misclassification. When the proportion of subjects misclassified by disease does not depend on the subject's status with respect to other variables in the analysis, including exposure, it will occur non-differential disease misclassification. Similarly, when the proportion of subjects misclassified by exposure does not depend on subject status to other variables in the analysis, including disease, it will occur nondifferential exposure misclassification.

We will give an example to illustrate how an independent nondifferential disease misclassification with full specificity does not bias the risk ratio estimate but rather biases the absolute magnitude of the risk difference estimate downward by a factor, equal to the probability of false negatives. Suppose there is a cohort study in which 30 cases occur in 300 unexposed subjects and 60 cases occur among 200 exposed subjects. The actual risk ratio is 3, and the actual risk difference is 0.20. Assumes no false positives for disease detection, sensitivity is only 70% for both exposure groups. The expected numbers of exposure cases detected will be  $0.70 \times 60$  and unexposed cases detected will be  $0.70 \times 30$ , which means that the expected risk ratio is estimated to be  $((0.70 \times 60)/200)/((0.70 \times 30)/300) = 3$  and the expected risk difference is estimated to be  $(0.70 \times 60)/200 - (0.70 \times 30)/300 = 0.14$ . Thus, although disease misclassification did not bias the risk ratio but the expected risk difference estimate was  $0.14/0.20$  of the actual risk difference.

The effects of nondifferential misclassification of exposure are similar to the effect of nondifferential misclassification of disease. We hypothesized a cohort study comparing the incidence of liver cancer in smokers with the incidence among nonsmokers to explore nondifferential exposure misclassification. The incidence

rate was assumed to be 0.01% per year for nonsmokers, and 0.05% per year for smokers. We suppose 2/3 of the study population are smokers, but only 50% admit this. This would then result in only 1/3 of subjects being identified as smokers with a disease incidence of 0.05% per year. And the remaining 2/3 of the population is made up of equal numbers of smokers and nonsmokers. Among those classified as nonsmokers, their average incidence would be 0.03% per year rather than 0.01% per year. The rate difference has been reduced by misclassification from 0.04% to 0.02%, while the rate ratio has been reduced from 5 to 1.7.

These examples present how a nondifferential misclassification of a dichotomous exposure will produce a bias toward the null value (no relationship) if the misclassification is unrelated to other errors. The association will be completely obliterated and the direction of association will be reversed by bias, if the misclassification is severe enough (although the reversal will only occur if the classification method is worse than randomly classifying people as “exposed” or “unexposed”).

We cannot dismiss a study simply because of the presence of substantial non-differential misclassification of exposure, it is incorrect. This is because the implications may be greater if there is no misclassification, which provides a probability of misclassification that applies uniformly to all subjects. Thus, the impact of nondifferential misclassification depends heavily on whether the study is considered positive or negative. Emphasizing measurement rather than qualitative descriptions of study results can reduce the likelihood of misinterpretation, but even so, it is important to keep in mind the direction and possible magnitude of bias.

### **8.3.3 Control**

Whether differential misclassification or nondifferential misclassification is mainly due to problems in measurement or data collection methods, resulting in errors in acquired data. Therefore, we mainly adopt the following methods to control information bias.

#### **8.3.3.1 Material Collection**

The main purpose of the survey design is to standardize the tables in the study, which is crucial for internal validity, so that valid, reliable, and complete data could be collected efficiently. In addition, pretesting survey instrument in populations similar to the study population can identify flaws in the survey design and instruments before full data collection begins. We'd better use the blinding method to collect data to avoid the influence of subjective psychology of research objects and investigators on the survey results.



### **8.3.3.2 Objective Research Indicators**

Try to use objective indicators or quantitative indicators to avoid information bias, such as applying laboratory examination results and consulting the medical records or health examination records of the subjects as the source of investigation information. Suppose it is necessary to collect data by means of inquiry. In that case, we should adopt closed questions and answers as far as possible to prevent the occurrence of report bias and measurer bias. For questionnaires concerning lifestyle and privacy, the respondents should be informed in advance that all responses are confidential and will be properly kept appropriately.

### **8.3.3.3 Investigation Skills**

The investigative skills of investigators are particularly important when obtaining information, especially the research that requires the participation of investigators. We can improve their investigation level by training investigators and formulating investigators' manuals to reduce information bias.

## **8.4 Confounding Bias**

### **8.4.1 Definition**

Confounding bias is due to one or more external factors that mask or exaggerate the link between research factors and diseases, thus partially or entirely distorting the actual relationship between them. Confounding is produced by confounders (exposures, interventions, treatments, etc.).

Taking Stark and Mantel's study on neonatal Down's syndrome as an example. Population monitoring data indicated that Down's syndrome was associated with birth order. Assume the incidence of Down's syndrome in the first-born child was 0.06% while in the fifth-born child was 0.17%. The risk of Down's syndrome increased with the increase of birth sequence, which seemed birth order to be a risk factor for Down's syndrome. However, we should consider maternal age at delivery as a confounder, closely related to birth order and Down's syndrome risk. Further study found that the incidence of Down's syndrome in children delivered by pregnant women younger than 20 years old was 0.02%, and gradually increased with the age of delivery, and the incidence of Down's syndrome in children delivered by pregnant women over 40 years old was as high as 0.85%. The study indicates that the maternal age at childbirth is related to the occurrence of the disease. Therefore, it is suggested that the association of birth sequence with Down's syndrome risk may be influenced by the confounding factor of maternal age at birth.

In this part, we briefly refer to confounding bias, but we will discuss confounding and how to control it in the next part.

### 8.4.2 *Confounding*

When estimating the effect of an exposure on exposed individuals, Confounding can occur when the exposed and nonexposed subgroups of the population have different background disease risks. These subgroups can have different disease risks even if they are not exposed to any of the effects in both subpopulations. More generally, confounding occurs when the exposed and unexposed groups are not fully comparable or “exchangeable” in terms of exposure response, i.e., the exposed and unexposed groups may exhibit different risks even if both experience the same level of exposure. In general, a factor associated with both the exposure and the outcome could be a confounder. The following are three necessary but not sufficient conditions to be a confounder of the effect of the exposor.

First, confounder must be predictors of the disease without the exposure under study. Confounders are not necessarily the genuine cause of the disease under study. However, they are only “predictive” within the level of exposure apart from casual relations. For example, race, age, gender, etc., may be considered as potential confounders. Thus, one almost always sees adjustments made for age and sex.

Second, the confounder must be related to the study exposure. For example, confounder should be related to exposures in the control group in case-control study. If the factor is not associated with exposure in the control group, an association between cases may still occur because both the study factor and the potential confounder are risk factors for disease, but this is a consequence of those effects and therefore does not cause confounding.

Third, confounder cannot be intermediate variables between exposure and outcome. In other words, confounders cannot be intermediates in the causal pathway between exposure and disease, or a condition caused by the outcome. To do otherwise would introduce a serious bias. Hypothetically, in a study of overweight and the risk of cardiovascular disease, it would be inappropriate to control for diabetes as confounder if diabetes was a consequence of being overweight and is also a part of the causal chain leading to overweight and cardiovascular disease. On the other hand, assuming diabetes is studied directly as a primary interest, overweight would be regarded as a potential confounder if it also involved exposure to other risk factors for cardiovascular disease.

We discussed the misclassification of disease and exposure in information bias. Here we need to refer to the misclassification of confounders. The ability to control confounding in the analysis will be hindered if a confounder is misclassified. Although independent nondifferential misclassification of exposure or disease usually causes the study results to be biased in the direction of the null hypothesis, independent nondifferential misclassification of a confounder usually reduces the degree of control for confounding, which may lead to bias in either direction. For this

reason, misclassification of confounder can be a serious concern. If the confounding is robust and the exposure–disease relationship is weak or zero, misclassification of the confounder can yield highly misleading results, even if such misclassification is independent and non-differential.

### **8.4.3 Control**

In the study design and analysis, confounding bias can be controlled by adjusting for all confounders or a sufficient subset of them at the same time. There are usually three methods to control for bias during the design stage.

#### **8.4.3.1 Random Allocation**

The first method is randomization, where participants are randomly assigned to exposure categories (applicable to experiments only). Ideally, we can create study cohorts with the equal incidence rate and eliminate the potential for confounding. But it must be practically and ethically feasible to assign exposure subjects. If just a few factors determine incidence, and the investigation personnel are aware of these factors, the ideal plan might call for exposure assignment that would result in the identical, balanced distributions of these disease causes in each group. Nonetheless, in studies of human disease, there are always immeasurable causes of disease that cannot be forced to be balanced amongst treatment groups. Randomization is one approach that permits one to probabilistically limit the confounding of unmeasured factors and to quantitatively account for the potential residual confounding arising from these unmeasured factors. However, this is usually only one alternative that may be beneficial for potential exposures. For instance, it is impractical and unethical to conduct randomized trials of the health effects of smoking, and therefore randomized trials may fail to prevent all confounding.

#### **8.4.3.2 Restrict**

The second control method is restriction, i.e., limiting the conditions of the study subject to a narrow range of values of the potential confounders. If a variable is prohibited from changing, it will not generate confounding if it is prohibited from varying. The restriction is a promising way to prevent or at least reduce confounding by known factors, it is both extremely effective and inexpensive. However, the advantages of restricting the study must be balanced against the disadvantages of reducing the study population when potential subjects are less plentiful. This approach has several conceptual and computational advantages, but may severely reduce the number of study subjects available and ultimately limit the extrapolation of results.

### 8.4.3.3 Matching

The third control method is matching, where study subjects are matched on the basis of potential confounders. Matching may be done by subject to subject, called individual matching, or for groups to groups, called frequency matching. Individual matching refers to the selection of one or more reference subjects with equal matching factor values to those of the index subject, whereas frequency matching refers to the selection of a whole stratum of reference subjects with similar matching-factor values to that of a stratum of index subjects. Individual matching would prevent age-gender-race confounding in cohort studies but is seldom done because it is very labor-intensive. In addition, matching does not completely eliminate confounding but does facilitate its control in case-control studies because matching for strong confounder will usually improve the precision of effect estimates. We have to discuss the concept of overmatching, which is often occurred in matched studies. In case-control studies, matching may be less accurate if the match factor related to exposure is only a weak risk factor for the disease of interest. When the number of matching factors exceeds 3, finding a suitable control becomes increasingly difficult.

### 8.4.3.4 Data Analysis

The above three control methods are usually implemented during the design phase. The analysis phase can also employ a number of methods to control for confounding bias. In the most straightforward situation, controlling for confounding in the analysis includes stratifying the data according to the level of confounders and calculating an effect estimate that summarizes the association between the strata of confounding factors. In a stratified analysis, it is usually not possible to control for more than two or three confounders at the same time, because finer stratification often results in many strata that contain no exposed or non-exposed individuals. Such strata are noninformative; therefore, a stratification that is too fine is a waste of information. In addition, we can use multi-factor analysis and standardized analysis to control confounding bias.

# Chapter 9

## Cause of Disease and Causal Inference



Li Ye

### Key Points

- In epidemiology, cause and causal inference are used to explore the etiology of risk factors for diseases at a population level.
- A causal model is a concise and conceptual graphic that describes the relationship between cause and disease.
- Most epidemiologic study designs can be used for evaluating causation. The strength of these designs to evaluate causation varies.
- Mill's canons represent logical strategies for inferring a causal relationship.
- Hill's criteria are a list of guidelines to distinguish causal and noncausal associations; these criteria have been widely used and are the best known criteria for assessing causal inference.

### 9.1 Introduction

One of the major focuses of epidemiology is to find the causes of diseases or events. Understanding the causes of diseases is important not only for correct diagnoses and treatments but also for effective prevention and control strategies. Therefore, cause of disease and causal inference—the process by which we identify the cause of disease are essential in both clinical and preventive medicine. In epidemiology, cause and causal inference are used to explore the etiology of or risk factors for diseases as well as their impact on the development of disease at the population level, which can provide unique insights into the etiology of the disease and lead to a population-level understanding of the disease. This chapter describes the epidemiologic concept of cause and the approaches to causal inference.

---

L. Ye (✉)

School of Public Health, Guangxi Medical University, Nanning, China

## 9.2 Cause of Disease in Epidemiology

### 9.2.1 *The Concept of Cause in Epidemiology and its Development History*

There are many definitions of cause in epidemiology. The following widely accepted definition is from Abraham Lilienfeld: a causal relationship would be recognized to exist whenever evidence indicates that the factors from part of the complex of circumstances that increase the probability of the occurrence of disease and that a diminution of one or more of these factors decrease the frequency of that disease. Another definition from Kenneth Rothman and Greenland [10] is also widely accepted due to its simplicity and clarity: an event, condition or characteristic, or a combination of these factors that play an essential role in producing an occurrence of the disease. Cause is an important concept in epidemiology. There are many other synonyms to describe cause, including causal agency, determinant, risk factor, exposure, etiological factor, etiological agent, etc. In epidemiology, cause is often referred to as a risk factor, which means the factor that increases the risk of disease.

The cause of disease has long been explored. The most ancient idealism attributed the occurrence of diseases to the god or devil. In the fourth century BC, Hippocrates, the father of medicine, considered that diseases occurred because of the imbalance of “four body humors.” In the fifth century, Chinese ancestors founded a materialistic view of the cause, and they proposed diseases were from the imbalance of “Yin-Yang” or “Five elements (wood, earth, water, metal, fire).”

In the later nineteenth century, at the height of the era of germ theory, Robert Koch, the founder of modern bacteriology, proposed Koch’s postulates, which include four generalized principles for determining whether a specific microorganism causes a specific disease. Koch’s postulates contributed greatly to the formation of the concept of cause in epidemiology because identifying the microorganism was equivalent to identifying the cause of the disease. In fact, the discipline of epidemiology as well as the concept of the cause of disease originated from etiology and epidemic studies on communicable diseases, among which the germ theory and Koch’s postulates represent landmark achievements.

However, Koch’s postulates cannot explain the causes for most diseases, especially noncommunicable diseases, which have replaced communicable diseases as the main threat to human health since the middle of the twentieth century. More recently, the epidemiologic studies have focused more on the probability and multicausality of the occurrence of diseases, which finally led to the formation of a modern concept of cause, as described at the beginning of this subsection.

### 9.2.2 Classification of Cause

In modern epidemiology, the concept of cause actually means “multicausality”. Most diseases, whether communicable or noncommunicable, have more than one cause. Since the definition of cause, either by Lilienfeld or by Rothman, means that any “factor” or any “event, condition or characteristic” plays a role in affecting the occurrence of the disease, the cause in epidemiology covers a wide range of factors, including individual genetics, physiological influences, environmental influences, social structure, etc. According to the source of these factors, we can divide the causes into two general categories: host factors and environmental factors (Table 9.1). Host factors refer to various characteristics that are related to people or an individual, such as genetics, immune status, age, sex, race, and behavior. (Table 9.1). Environmental factors mainly include biological, physical, chemical, and social factors (Table 9.1).

**Table 9.1** Classification of the cause of disease

Factor (cause)	Description
<b>Host factors</b>	
1. Genetics	Chromosomal disorder, single gene disorder, polygenetic disorders, etc.
2. Immune status	It involves in the occurrence of most diseases, both communicable and noncommunicable
3. Age and sex	People of different age or sex may be susceptible to different diseases
4. Race	Occurrence of disease has difference in race
5. Personality	Temperament, psychological status, psychiatric status, etc. may have effects on the occurrence or progression of diseases
6. Behavior	Bad behaviors or habits such as smoking, drinking, poor diet, lack of exercise, unsafety sexual behaviors, drug abuse, noncompliance with traffic laws, etc.
<b>Environmental factors</b>	
1. Biological	Pathogenic microorganisms (bacteria, viruses, rickettsiae, mycoplasmas, chlamydiae, spirochetes, actinomyces, etc.); parasites (protozoa, worms, insects, etc.); venomous animals and poisonous plants (snakes, ergot, mushrooms, etc.)
2. Physical	Temperature, humidity, altitude, noise, light, vibration, radiation, dust, fire, etc.
3. Chemical	Pollution, agricultural chemicals, food additives, microelement, heavy metal, etc.
4. Social	Social system, socioeconomic level, war, disaster, education, religion, living condition, lifestyle, occupation, family relationship, etc.

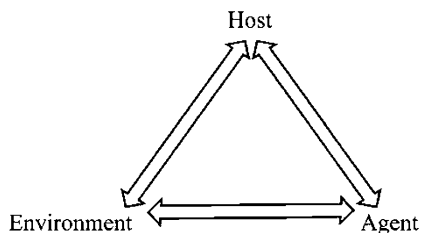
### 9.2.3 Causation Models

A causal model is a concise and conceptual graphics that describes the relationship between cause and disease. During the development of etiology, different causal models were proposed based on contemporary understanding of the diseases in different historical periods, which made a great contribution to the formation of the modern concept of cause. Casual models can be used to illuminate the association between cause and disease as well as the relationship between multiple causes of a disease and to provide direction or clues to find a new cause. Essentially, the aim of causal models is to find causes and elucidate the dominant cause and ultimately determine the best prevention or intervention strategy. The most representative causal models are the triangle model, the wheel model, the chain of causation model, and the web of causation model.

#### 9.2.3.1 Triangle Model

In 1954, John Gordon summarized the knowledge about the epidemiologic etiology of diseases at that time and put forward an epidemiologic triangle model (epidemiologic triad) to describe the relationships between multifactorial causes and a disease, especially communicable disease. The model considers that host factors (age, sex, race, genetic profile, immune status, etc.), agents (biologic pathogens, chemical, physical, nutritional agents, etc.), and environmental factors (temperature, humidity, crowding, housing, water, food, radiation, pollution, noise, etc.) are the troika of a disease. These three aspects are indispensable for the occurrence of a disease and have an equal role in the occurrence of disease. Hence, the relationships can be described as an equilateral triangle (Fig. 9.1). The three kinds of factors interact and restrict each other, and thus, a dynamic balance exists that makes the occurrence of disease in a stable state. Once the balance is disturbed, the occurrence of disease increases or decreases. The triangle model is helpful even today for finding the cause of communicable diseases and controlling the epidemic. However, it is basically unsuitable for the description of noncommunicable diseases.

**Fig. 9.1** Epidemiologic triangle model





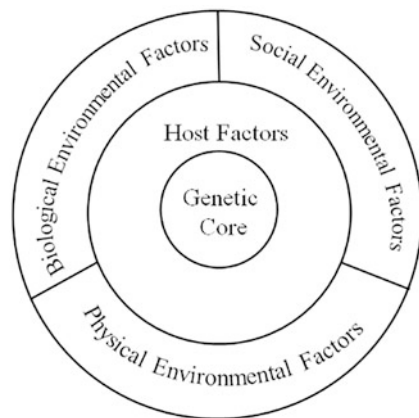
### 9.2.3.2 Wheel Model

In the middle of the twentieth century, noncommunicable diseases became the main threat to humankind. However, there is no obvious or absolute agent for most noncommunicable diseases, as a pathogen agent for communicable diseases. It is very difficult to describe the relationship between cause and noncommunicable disease using a triangle model. In 1985, Mansner and Kramer proposed the wheel model based on the triangle model. In this model, host factors play the core role in the occurrence of disease and are located in the center of the wheel, with genetic factors as the core of the center (Fig. 9.2). The host center is surrounded by three kinds of environmental factors, including biological, social, and physical environmental factors. The main difference between the wheel model and the triangle model is that the wheel model considers that different factors have different importance for the occurrence of disease. Therefore, the area sizes of the center (host factors) and surrounding parts in the wheel (biological, social, and physical factors, respectively) can be adjusted to reflect the importance of different factors. The wheel model emphasizes the core role of host factors as well as the influencing effects of environmental factors. It is considered to be better than the triangle model and suitable for both communicable and noncommunicable diseases. However, the wheel model came from etiology knowledge in the 1980s and could not truly reflect the complex interactions between various factors. It is still limited for many noncommunicable diseases, especially chronic diseases.

### 9.2.3.3 Chain of Causation Model

In multicausality theory, there are multiple causes of communicable and noncommunicable diseases. The multiple causes or risk factors can always be displayed in the form of a chain. Some factors are direct or proximal or most immediate causes, and others are indirect or distal causes. Some factors are

**Fig. 9.2** Causation wheel model





**Fig. 9.3** Chain of causation (diabetes)

independent causes; however, others are dependent causes that interact with other causes. To interpret the association of different causes and final disease as well as the complex relationship among multiple causes, a model of “chain of causation” was proposed to describe the causes in the form of a chain. For example, an accelerated life tempo can lead to an unhealthy diet or less exercise and then obesity, followed by insulin resistance, which often results from obesity and further results in elevated blood glucose. Finally, diabetes occurs when blood levels of glucose become chronically elevated (Fig. 9.3). It is worth mentioning that removing any factor in the chain can block the whole chain and thus prevent the occurrence of disease (Fig. 9.3).

#### 9.2.3.4 Web of Causation Model

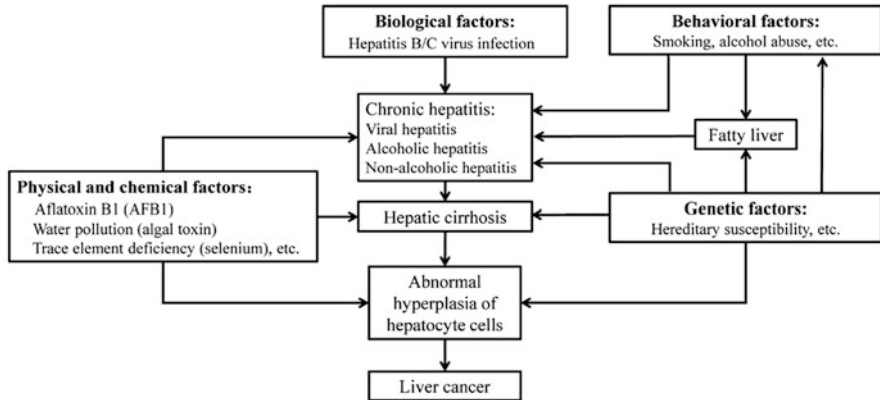
In some cases, a disease has several and interrelated chains of causation. The causation chains of the disease link and interplay with each other, and thus constitute a complex network. MacMahon proposed the web of causation model to describe the complex relationships between causes and disease as well as the interlacing chains of causation.

For example, the causation of liver cancer can be described as a network or a web. The four chains of causes of liver cancer consist of biological factors, physical and chemical factors, behavioral factors, and genetic factors (hereditary susceptibility). Multiple factors of the four chains also interact with each other and form a network, thus ultimately leading to the occurrence of liver cancer (Fig. 9.4).

#### 9.2.4 Sufficient Cause and Necessary Cause

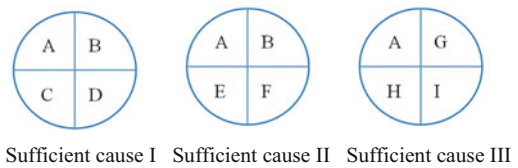
Modern epidemiology considers that the relationships between cause and effect are multiple and complex. A given disease can be caused by many factors; however, a single factor is not enough to cause the disease, as joint action from other causes is necessary. Obviously, the role and importance of different factors are different in the occurrence of a disease. From the logical view of cause and effect, all effects have sufficient and necessary conditions; thus, cause in epidemiology can also be divided into sufficient and necessary cause.

In 1976, Kenneth Rothman used a sufficient-component causal model (Fig. 9.5) to explain the complex relationship between cause and effect. Rothman proposed that a sufficient cause is a factor or a combination of several factors that will



**Fig. 9.4** Web of causation (liver cancer)

**Fig. 9.5** Sufficient cause and necessary cause



inevitably cause disease. A component cause is a factor that contributes to the occurrence of disease but is not sufficient to cause disease on its own. A necessary cause is any agent that is required for the occurrence of disease (for example, cholera bacillus for cholera occurring); without necessary cause, the disease will not occur. For instance, the three sufficient causes (I, II, III) shown in Fig. 9.5, comprise 4 component causes. In this figure, there are three sufficient causes (I, II, III), and A, B, C, D, E, F, G, H, and I are component factors. Because A is present in all three sufficient causes, it is a necessary cause.

The sufficient-component causal model interprets two paradoxes of the causation theory in epidemiology. First, why does a given disease occur without a specific cause? For example, alcohol abuse is the cause of cirrhosis; however, individuals who never drink may also develop cirrhosis. The possible reason is that cirrhosis develops through other sufficient causes, such as hepatitis B virus infection (Fig. 9.4). Second, why does a disease fail to occur in the presence of a specific cause? For example, smoking causes lung cancer; however, many smokers never develop lung cancer in their lifetime. This phenomenon can be explained by the fact that smoking is not a sufficient cause of lung cancer. In reality, most identified causes for noncommunicable chronic diseases are neither necessary nor sufficient. For example, hypertension is neither a necessary cause nor a sufficient cause for cardiovascular disease. Nevertheless, every component cause is necessary for sufficient cause that contains it. Removal of any component cause is equal to the removal of

the sufficient cause that contains this component cause, which is an important strategy for the prevention of a disease.

## 9.3 Epidemiologic Methods of Causation

### 9.3.1 Epidemiologic Study Designs for Causation

Etiological studies in epidemiology usually contain several common steps. The first step is to find the influencing factors that are associated with a disease or to develop a cause-and-effect hypothesis, usually by descriptive or analytical epidemiologic studies; the second step is to test the hypothesis often using analytical studies; and the last step is to verify the hypothesis, usually by experimental epidemiologic studies.

Most epidemiologic study designs can be used for evaluating or establishing causation. However, the strength of these designs to evaluate causation is different. Table 9.2 outlines the relative strength of the different study designs in establishing causation. These study designs have been introduced in prior chapters, and their use in providing evidence for causation will be described as follows.

#### 9.3.1.1 Descriptive Studies

The start of causation is to develop a cause-and-effect hypothesis. Descriptive studies are always used to generate hypotheses. Descriptive studies mainly include case reports, case series, cross-sectional studies, and ecological studies. Case reports and case series are useful for developing a hypothesis based on analysis of the characteristics of patients or case groups. Cross-sectional studies are always used to describe the distribution of disease in different populations, and the pattern or trend of disease occurrence over time or by geographic area, which can provide the clues regarding influencing factors. Ecologic studies explore the association of influencing factors and disease at the population or region level, which can also provide clues for influencing factors that cannot be measured at the individual level, for example, air

**Table 9.2** The strength of evidence for causation by different epidemiologic study designs

Type of study design	Strength of evidence in causation
Randomized controlled trials	Strong
Nonrandomized controlled studies	Moderate
Cohort studies	Moderate
Case-control studies	Moderate
Cross-sectional studies	Weak
Ecologic studies	Weak
Case reports	Weak

pollution. Generally, the strength of descriptive studies to evaluate causation is weak compared with analytical studies or experimental studies due to a lack of evidence on the time sequence of events. Of all descriptive studies, the weakest for causation is case reports because they have neither defined populations nor comparison groups. Nevertheless, when causal relationships have already been established, well-designed descriptive studies, especially cross-sectional studies with multiple time points or time series studies, can be very useful to quantify the effects of cause.

### **9.3.1.2 Analytical Studies (Case-Control Studies, Cohort Studies)**

Analytical studies, mainly including case-control studies and cohort studies, are more reliable methods to form a hypothesis than descriptive studies. Analytical studies can also be used for hypothesis testing. Case-control studies, which are mainly used to confirm the association between factors and disease, compare the exposure levels between the case group and the control group. Because the research direction of case-control studies is from effects (diseases) to causes (factors), this study design is vulnerable to various biases. Cohort studies are either prospective or retrospective, and they can test hypotheses more effectively by comparing the incidence rates of exposure groups with control groups, and directly calculating the relative risk (RR) of factors in the temporal order of cause and effect. Well-conducted cohort studies are a better design for causation than case-control studies because the former can minimize various biases, including selection, information, and confounding biases.

### **9.3.1.3 Experimental Studies**

Experimental studies include clinical trials, field trials, and community trials. Clinical trials are most frequently conducted among patients, with the aim of evaluating the efficacy of a new treatment or medicine. Therefore, clinical trials are known as a robust and reliable method to test or verify hypotheses, especially clinical randomized controlled trials (RCTs), which are considered the gold standard to evaluate a new treatment or medicine and the most rigorous method for hypothesis testing. Nevertheless, RCTs are subjected to many constraints, such as ethical issues, strict inclusion criteria, parallel control, and strict randomization, which greatly limit the feasibility of RCTs in causation studies. Quasi-experiments that lack parallel control or randomized assignment are also used in epidemiologic etiology, with less strength to evaluate causation. Other experimental studies, including field and community trials, are seldom used to study causation. Field trials mainly involve people who are disease-free, with the aim of preventing the occurrence of diseases. Community trials are conducted at the level of the community instead of the individual level. Therefore, although experimental studies have strong strength to test and verify hypotheses, most of the causative evidence so far has not come from this study design but comes from observational studies such as descriptive and analytical studies. For

example, most of evidence about the effects of smoking on health comes from case-control and cohort studies.

### ***9.3.2 Mill's Canons-the Logical Basis of Causation***

The causa models mentioned in Sect. 9.2.3 of this chapter are mainly used to describe the relationships between the factors and diseases or between different risk factors. They cannot be used as a method to find causes or test causation. Mill's canons were proposed by philosopher John Stuart Mill in 1843, intended to illuminate a causal relationship between a circumstance and a phenomenon, which provides the logical basis of causation studies. In epidemiology, Mill's canons provide certain guidance for causation, especially the development of hypotheses. The canons with minor adjustments constitute five methods for the induction of hypotheses.

#### **9.3.2.1 Method of Agreement**

This method means that factors in common among different instances of a disease are perhaps the cause or a necessary part of the cause of the given disease. In other words, if two or more instances of a disease under investigation have only one factor in common, which is likely to be the cause of the given disease. For example, one school had an outbreak of diarrhea. It was found that all the students with diarrhea had consumed soy milk in the same canteen in the morning, so soy milk may be the cause of diarrhea.

However, in actual conditions, it is difficult to obtain only "one common factor". There may be a few other factors shared by patients with the same disease, but most of the factors are not the cause of the disease. In the example mentioned above, most students with diarrhea may have also eaten another food in common in the same cafeteria that morning. Therefore, the method of agreement in this example is actually not sure that soy milk may be the cause of outbreaks of diarrhea. Generally, a hypothesis cannot be formed by one method.

#### **9.3.2.2 Method of Difference**

If some instances in which the disease occurs, and other instances in which the disease does not occur, they have all other factors in common except one existing only in the former. That one may be the cause or a necessary part of the cause of the disease. The method of agreement concerns whether patients share certain common factors. The method of difference compares the differences in certain characteristics between patients and nonpatients. For example, if one student had diarrhea while the

other did not, the only different food that two students consumed in the canteen was soy milk, and the soy milk may be the cause of diarrhea.

Similar to the situation of the method of agreement, in the method of difference, the assumption that “all other factors are the same” between patients and nonpatients is hard to make in practice. In the above example, the two students may be different from each other in many other aspects. The method of difference cannot exclude other factors and hypothesize that only soy milk is the cause of diarrhea.

### **9.3.2.3 Joint Methods of Agreement and Difference**

This method is actually a combination of the method of agreement and the method of difference, however, it is not a simple combination but alternately contains multiround use of two methods. Briefly, if two or more instances in which the disease occurs have only one factor in common, while two or more instances in which the disease does not occur have nothing in common except the absence of the factor commonly existing in the former instances, then, the factor may be the cause, or a necessary part of the cause, of the disease. Generally, the joint methods of agreement and difference are much more likely to find a risk factor than the method of agreement or difference alone. The main reason is that joint methods of agreement and difference essentially introduce contrast in the investigation, which greatly increases the logicity.

We return to the example of soy milk and diarrhea. If all of the students with diarrhea had consumed soy milk, the students without diarrhea must have not consumed soy milk in the same canteen. This is the joint method to indicate that soy milk was likely to be the cause of diarrhea.

### **9.3.2.4 Method of Concomitant Variations**

According to the method of concomitant variations, whatever one event varies in any manner whenever another event varies in some particular manner. The former event is either a cause or an effect of the latter; in other words, these two events are connected through cause-and-effect association. In essence, method of concomitant variations emphasizes dose-dependent relationships. When there is a dose-dependent relationship between two events, cause-and-effect associations are more likely exist.

In the above examples, if students who consumed more soy milk had more severe diarrhea, there was a dose-dependent relationship, namely, concomitant variations, so the probability of the soy milk as the cause of diarrhea was higher.

### 9.3.2.5 Method of Residue

Suppose a disease is caused by many factors; when you remove the previously known factors as well as the instances of disease caused by those factors, the residue of factors may be the cause for the remaining instances of disease. For example, in 1972, a number of dermatitis cases occurred in Shanghai, China. Possible factors, including industrial waste gas, plant pollen, blood-sucking arthropods, and poisonous moths, as well as dermatitis cases caused by these factors were excluded. The residual factor, *Euproctis similis*, emerged. Therefore, researchers suspected that this outbreak of dermatitis was caused by *Euproctis similis*, and finally confirmed this hypothesis.

Although Mill's canons are considered as a logical basis for causation studies and still have certain guidance significance for current epidemiologic etiologies, they have great limitations in actual practice in causation studies. Generally, Mill's canons are suitable to find both sufficient and necessary causes, such as acute infectious agents and to judge strong causal associations. However, for most diseases, especially noncommunicable chronic diseases, the risk factors are almost all nonsufficient and nonnecessary causes, and one disease always has multiple sufficient causes. In these cases, the Mill's canons are not suitable for effectively assessing the causal association. At most, the canons just play a role in the formation of a hypothesis.

## 9.4 Causal Inference

Causal inference is the term used for the process by which we identify the cause of disease. In other words, it is used to determine whether the observed association is causal. In essence, the relationship between cause and disease is a kind of cause-and-effect association in philosophy. Various risks or exposure factors are the cause, and diseases are the effect. The term association is another important concept in epidemiology.

### 9.4.1 Association Vs. Causation

Two events, the suspected cause and the effect, obviously must be associated if they are to be determined as causally related. However, not all associations are causal, namely, cause-and-effect associations. Various other associations, including chance association, spurious association, and noncausal association, which are caused by various reasons such as random error, bias, or confounding, should be excluded before a causal association is assessed. Figure 9.6 outlines various associations caused by different reasons.



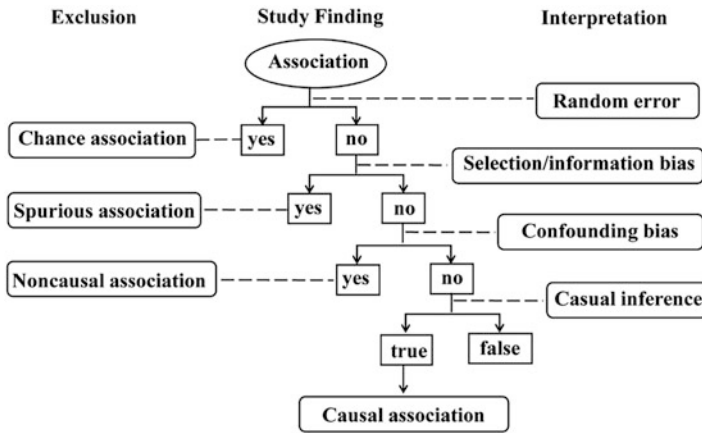


Fig. 9.6 Association and causation

### 9.4.1.1 Chance Association

First, we need to judge whether an association between two events, for example, an exposure and an outcome, is a statistically significant association rather than a chance association due to random error, e.g., sampling error and random measurement error (Fig. 9.6). In epidemiology, the strength of association can be expressed as rate ratios, odds ratios, or attributable risks, which are introduced in prior chapters. The exclusion of chance association is based on statistical comparison of these indicators between the exposed group (or case group) and the control group. When the *P* value was less than 0.05, we considered that there was a statistical significant difference, namely, a statistically significant association. The *A* value less than 0.05 is defined as a low-probability event, which means that the result is reliable in at over 95% probability, and there is less than a 5% chance that the result is caused by random error. Because random error cannot be avoided and exists in all study designs as well as each step of a study, well-designed and well-conducted studies are essentially important to effectively reduce chance association caused by random error, despite almost no study being perfect in either design or conduct practice.

### 9.4.1.2 Spurious Association

If the association is statistically significant and the probability of random error is very limited, then we could evaluate whether the association is spurious, which generally comes from nonrandom systematic errors known as selection or information bias (Fig. 9.6). Spurious association means that the association truly exists, but is not true due to selection or information bias. Thus, it is also called as false association. Selection bias is generally caused by the difference in exposure- or outcome-related characteristics between subjects selected for study and those not

selected or between the exposed (or case) group and the control group. For example, in a case-control study on birth defects, the case group of newborns with birth defects, while the control group contained consists of those without birth defects. The collection of exposure information was primarily based on the memory of mothers of the newborns. In information collection, the mothers of newborns with birth defects, who were stimulated by adverse pregnancy were able to recall various exposures during pregnancy in detail, such as taking over-the-counter drugs, fever, or cold. However, mothers in the control group were less likely to make an effort to recall the details and did not respond carefully to the relevant exposure events, because no adverse pregnancy occurred. Therefore, the results obtained may be influenced by recall bias. The association between potential exposure factors and neonatal birth defects may be overestimated, and it may be a false association.

### **9.4.1.3 Noncausal Association**

Even though a true association exists, it is still necessary to know whether the association occurs indirectly by another extraneous factor (called a confounding factor), which always leads to a noncausal association. If confounding was not found, a noncausal association could be excluded and a causal association may exist (Fig. 9.6). Confounding bias is caused by an extraneous factor (confounding factor) that is closely related to both exposure and outcome but not an intermediate link in the causal chain of the exposure and the outcome. Confounding bias can lead to underestimation or overestimation of the association between the exposure and the outcome. For example, investigation may find an association between smoking and alcoholic liver, obviously, which is not reasonable in biological plausibility. The link between smoking and alcoholic liver is a noncausal association because alcohol drinking consumption is a confounding factor, which that is often associated with smoking and directly related to alcoholic liver. Confounding occurs commonly in epidemiologic studies. However, it can be well controlled through careful designs such as matching, restriction, and randomization or through analyses such as standardization, stratification, and multivariate analysis.

### **9.4.1.4 Causal Association**

After excluding chance association, spurious association, and noncausal association caused by random error, selection/information bias, and confounding bias, the association between the exposure and the outcome is likely to be causal, and still needs to be further assessed by various judgments. We call this process causal inference (Fig. 9.6). Among various judgments, the best known and widely used is Hill's criteria. The details of Hill's criteria are introduced in the next subsection.

## 9.4.2 Evaluating Causal Association—Hill’s Criteria

The existing association between two events or two variables after exclusion of chance, spurious, and noncausal associations is just probably to be a cause-and-effect association, which is required for further judgment based on totality of evidence. Judgment of causal association is neither simple nor straightforward, and various sets of guidelines have been proposed for the judgment. In 1965, the British statistician Sir Austin Bradford Hill proposed a list of nine guidelines to evaluate causal association, which has been widely used and is certainly the best known set of criteria for the considerations of causation, sometimes with modifications (Table 9.3).

### 9.4.2.1 Temporal Relationship

In causal inference, temporal relationship is essential: the cause must precede the effect; or, in other words, an exposure to cause a disease must precede the development of the disease. Of all Hill’s guidelines, this is an absolute requirement. Different study designs have different strengths to provide evidence of temporal relationships. Cohort and experimental studies have obvious temporal relationships because they are performed prospectively. However, in cross-sectional studies, difficulty may arise in judging temporal relationships because the proposed cause and effect are measured at the same time point. In case-control studies, sometimes it is assumed that one event precedes another without actually establishing the order, and in other cases, it may be difficult to determine which the first is. Nevertheless, there are some strategies to find evidence supporting temporal relationships. For example, when the cause is an exposure that can be divided into different levels, it is essential that a sufficiently high level should be reached before the disease occurs. Repeated measurement of exposure at multiple time points or in different locations

**Table 9.3** Hill’s criteria for causation

Criteria	Comments
Temporality	The cause precedes the effect (essential criterion)
Strength	The strength of association between the cause and effect (odds ratio, relative risk)
Dose-response	Increased exposure is associated with increased effect
Consistency	Similar results are shown in other studies
Biologic Plausibility	The association is consistent with biologic mechanism
Reversibility	Removal or reduction of exposure is followed by decreased effect
Specificity	One cause leads to one effect, and vice versa
Analogy	Exposure and effect are similar to those in a well-established causal association
Experiment	Evidence from animal, intervention or mechanism studies

may also strengthen the evidence of temporal relationships. Although temporal relationships are necessary for causal inference, an existing temporal sequence alone is weak evidence for causation. Many things occur before an event: however, they have no relationship with the event. For example, someone may sneeze in Beijing city, and 30 minutes later, there's a heavy rain fall in Nanjing city. Obviously, there is no causal link between these two events.

#### **9.4.2.2 Strength of Association**

The stronger an association is, which is usually expressed by the relative risks (odds ratio, OR; relative risk, RR), the more likely a causal association is. A strong association less likely comes from either bias or confounding. Thus, the 13.70-fold higher risk of lung cancer among male smokers compared with nonsmokers is much stronger evidence than the finding that smoking is related to coronary heart disease, for which RR is only 2.00. What is a “weak” or “strong” association? There is no universal standard, but epidemiologists generally consider a relative risk (OR, RR) greater than 2.0 (or less than 0.5) to be moderately strong and a risk greater than 5.0 (or less than 0.2) to be strong. Nevertheless, a weak association does not mean that it can be overlooked for causal inference. Sometimes the strength of an association may depend on the prevalence of other possible causes. For example, the relationship between diet and coronary heart disease is a cause-and-effect association; however, the diets in populations are rather homogeneous, although greater variation may be observed among different individuals or in different stages of one person. In addition, a weak association, when combined with other guidelines, for example, consistently observed in different designs or in different settings, may also provide stronger evidence than a strong association that is only found in one or two studies.

#### **9.4.2.3 Dose-Response Relationship**

When changes in the level of possible cause are associated with corresponding changes in the incidence or prevalence of the disease, a dose-response relationship exists. Generally, the presence of a dose-response relationship in unbiased studies is considered strong evidence for causation. However, the absence of a dose-response relationship does not mean that the association is noncausal, because not all causal associations exhibit a dose-response relationship. For instance, there may be a “threshold” effect in which any exposure above a certain level will lead to disease. In addition, although a dose-response relationship is a strong evidence for causation, it cannot exclude confounding factors.

#### **9.4.2.4 Consistency**

Consistency means that, when several studies are conducted at different times in different settings or among different patient groups, the same or similar results are derived. Consistency is a kind of evidence strengthened for causal inference because the possibility that all different studies make the same “mistake” is minimized. However, a lack of consistency does not exclude a causal association. Different results may come from variation in study design or quality or different exposure levels and other conditions that may affect the impact of a causal factor on the effect.

#### **9.4.2.5 Biologic Plausibility**

A causal association generally should have biological rationality. The association between cause and effect is consistent with the current biologic knowledge and often enhances convincing causal inference. However, the lack of biological plausibility does not deny a causal association, which may simply reflect a lack of scientific knowledge or evidence. Increasing knowledge of biological mechanisms may support this association in the future. In other words, biologic plausibility, when present, enhances evidence for causation; when absent, other evidence for causation should be sought.

#### **9.4.2.6 Reversibility**

When the removal of a factor that is likely to be a cause of disease results in a decreased risk of disease, there is a greater possibility that the association is causal. An example is that people giving up smoking decreases their risk of lung cancer compared with people who continue to smoke. Reversible associations are strong but not infallible evidence for causation because they cannot exclude confounding factors, which can also conceivably account for a reversible association.

#### **9.4.2.7 Specificity**

Specificity refers to the strict corresponding relationship between a cause and a disease: that is, a certain factor can only cause one certain disease, and vice versa, the disease is just caused by a certain factor. This guideline is currently only applicable for some acute communicable or genetic diseases because for most diseases, either communicable or noncommunicable, there are many risk factors for the same effect or many diseases come from one factor. Therefore, specificity is considered the weakest evidence of all the guidelines for causation.

#### **9.4.2.8 Analogy**

Analogy is sometimes used in causal inference. Suppose there is a well-established cause-and-effect association: for instance, factor A leads to effect B. If a similar association is observed between factor C and effect D, which are also similar to factor A and effect B, respectively, we can consider that factor C is likely to be the cause of effect D. In general, analogy is weak evidence for causation.

#### **9.4.2.9 Experimental Evidence**

Experimental data, from studies in animals or other organisms, from intervention studies in humans, or from mechanistic studies, may also provide evidence supporting causal associations. In medicine, evidence from a well-conducted experimental clinical trial is always considered the strongest evidence for causation. However, in epidemiology, the results from a single or few experiments are generally not considered to be convincing or strong evidence for causation.

Among the nine guidelines described above, temporal relationships, and strength of association are necessary conditions to judge causality. That means that if there is a causal association, temporal relationships and statistically significant associations must exist; otherwise, causal associations can be denied. The other seven guidelines belong to unnecessary conditions, which are just general criteria for causal inference. A lack of any one or even all seven guidelines does not preclude causal association. Moreover, it is worth mentioning that all nine guidelines are not sufficient conditions for causation. Thus, even though a relationship between two events satisfies all nine guidelines, we cannot absolutely draw a conclusion that the relationship is causal. Causal inference is always tentative, and judgment must be made on the basis of the available evidence. Although there are no completely reliable criteria for determining whether an association is causal or not, Hill's criteria are widely accepted and have been applied in practice. Nevertheless, Hill's criteria are essentially guidelines for causal inference but not a "gold standard" to judge a causal association.

### ***9.4.3 An Example of Causal Inference Using Hill's Criteria***

Primary hepatocellular carcinoma (HCC) is one of the most common malignancies worldwide. Many researchers have investigated the causes of HCC and indicated that alcohol abuse (habitual heavy drinking) is likely to be a risk factor for HCC. The Causal inference was performed and summarized as follows.

### **9.4.3.1 Temporality of Association**

Cohort studies have indicated that habitual heavy drinking always precedes the occurrence of HCC. Some cases of HCC had several years or even several decades of heavy drinking history.

### **9.4.3.2 Strength of Association**

Many case-control studies have indicated that the risk (OR) of HCC among habitual heavy drinkers is 2–4-fold greater than that among nonhabitual drinkers or non-drinkers. Furthermore, a few cohort studies found that the relative risk (RR) of heavy drinking was 2–5-fold higher than that of nondrinking groups.

### **9.4.3.3 Consistency**

The association between habitual heavy drinking and HCC has been repetitively studied in many different countries by different study designs at different times. The results from different studies consistently indicated that heavy drinking is associated with the occurrence of HCC.

### **9.4.3.4 Dose-Response Relationship**

Previous studies have found that more alcohol consumption and a longer heavy drinking history result in a higher incidence of HCC. This is an obvious dose-response relationship.

### **9.4.3.5 Biologic Plausibility**

The relationship between alcohol abuse and HCC is consistent with the current understanding of alcohol metabolism in the liver. Alcohol is metabolized in the liver, and its metabolism produces free radicals that cause lipid peroxidation, damage mitochondria in liver cells, and lead to alcoholic liver injury.

### **9.4.3.6 Experimental Evidence**

A number of animal studies have shown a similar relationship between alcohol administration and hepatic injury.

Taken together, the evidence is strong for the conclusion that habitual heavy drinking is a risk factor for HCC.

# Chapter 10

## Disease Prevention and Surveillance



Chunhua Song

### Key Points

- The three levels of prevention of disease: (1) primary prevention: known as causation prevention, two complementary strategies are used: high-risk strategy and population-based strategy; (2) secondary prevention: it refers to early detection, early diagnosis, and early treatment; (3) tertiary prevention: known as clinical prevention or disease management.
- With the development of healthy connotations, the goal of public health is no longer just to prevent disease but also to actively maintain and promote health. The main strategies for achieving this include health protection and promotion.
- Public health surveillance is a persistent, continuous, and systematic process of information collection about health events and health problems.
- In addition to active surveillance and passive surveillance, routine reporting, and sentinel surveillance, the commonly used surveillance methods and analytical techniques are also included.

## 10.1 Prevention Strategies and Measures

### 10.1.1 Strategy and Implementation for Prevention

A strategy is a basic principle guiding the overall work under specific circumstances. It focuses on the overall situation and considers the issue from a macro perspective. The measures include the application of actual methods, steps, and plans that are used to achieve the desired objectives. The two parts are closely related. Only by taking the correct prevention strategies and under the guidance of reasonable

---

C. Song (✉)  
College of Public Health, Zhengzhou University, Zhengzhou, China  
e-mail: [sch16@zzu.edu.cn](mailto:sch16@zzu.edu.cn)



measures, can we achieve the desired preventive effect efficiently. However, without considering the feasibility of the measures, the strategy will not be implemented, and it is impossible to achieve the desired objectives. The effect of measures is often small without the guidance of strategies. Therefore, in order to minimize the influence and achieve the maximum effect, we have to consider the measures and strategy at the same time, which is the meaning of the strategy.

The successful eradication of smallpox is one of the best example that demonstrates the importance of the correct use of prevention strategy and measure. In 1796, Jenner found that Chickenpox vaccines could effectively prevent smallpox, and the rest of the world had used the same approach. However, in 1960, smallpox reoccurred and persisted in some countries and regions. There were still deficiencies in the strategy for preventing smallpox, therefore, it is necessary to modify and improve them. In addition to the vaccination rate, monitoring work should be carried out in order to detect, isolate, treat, and report the new cases, and the people who were in close contact with the cases should acquire artificial immunization (ring vaccination), which thoroughly controlled the spread of the disease. This method can not only save manpower and resources, but also achieve the same effect. Our ultimate goal is to eliminate smallpox worldwide, so we are not just emphasizing on large-scale vaccination, but also enhancing the monitoring process. This example shows that the strategies and measures to prevent disease complement each other. On May 8, 1980, WHO announced that smallpox had been eliminated in the world, and the strategy was modified again. As a result, countries around the world can stop vaccinating against smallpox and pay more attention to the risks to smallpox researchers. Especially, the detection of suspected smallpox cases strengthened laboratory testing and reservation of sufficient vaccines to meet any emergency needs.

## ***10.1.2 Disease Prevention***

### **10.1.2.1 The Definition of Disease Prevention**

Disease prevention is a series of activities that can reduce the occurrence of disease (or injury) and disability and prohibit or delay its development. The main purpose of prevention is to maintain high quality of life for patients by eliminating disease (or injury) and to minimize the impact of disease (or injury) and disability. If it is hard to do so, we must delay the occurrence of the disease or delay the development of disease and disability.

### **10.1.2.2 The Development of Disease**

The natural history of disease can be divided into four stages, including the stage of susceptibility, the stage of subclinical disease, the stage of clinical disease, and the

stage of recovery. For infectious diseases, the stage of subclinical disease commonly refers to the incubation period in which pathogen invades the body before the emergence of clinical symptoms. In chronic non-communicable diseases, the induction period indicates the time from exposure to the etiology factor to the onset of the disease, and the latency period indicates the time from the exposure factors to the occurrence of the disease manifestation. In this definition, the latency period contains the induction period. In another definition, the latency period refers to the time from the onset of the disease to the presentation of the disease, when the latency period follows after the induction period and the two periods do not overlap.

### 10.1.2.3 The Three Levels of Prevention of Disease

#### Primary Prevention

Primary prevention, also known as causation prevention, refers to taking measures against etiology or risk factors to reduce the level of harmful exposure, enhance the ability of the individual to combat harmful exposure, prevent the occurrence of a disease (or injury), or at least delay the occurrence of the disease when exposed to the disease (or injury). Primary prevention is a fundamental measure to eliminate a disease (or injury).

A variety of measures are needed for primary prevention, such as preventing harmful exposure in the environment, improving the body's resistance (e.g., immunization), or protecting the individual from harmful exposures, such as the elderly with osteoporosis wearing hip protection.

- ① **High-risk strategy:** high-risk strategy is based on clinical medicine aiming to achieve the first-level prevention strategy. High-risk strategy refers to taking targeted measures to reduce the level of risk exposure and its risk of future disease for small group of individuals with a risk of high incidence in the future. For example, adults are regularly assessed for risk factors for cardiovascular disease and appropriate measures are taken for those high-risk individuals, such as stopping smoking, controlling salt intake, eating more fruits and vegetables, and so on.

The limited availability of medical information means that healthcare is a system of limited supply and that there is a need to give priority to groups that are most likely to benefit or may benefit most. The use of resources by high-risk strategies may be more cost-effective. However, most lifestyles, such as eating, smoking, and exercising, are largely influenced and limited by the code of conduct and the behavior of the people around us in our society. The high-risk strategy, in essence, requires that the minorities behave differently; therefore, this undoubtedly limits the effectiveness of this strategy. If the vast majority of cases of a disease occur in a small group of easily identifiable populations, high-risk strategies are enough to control the disease and interventions for this group are effective and affordable. However, when the most fundamental cause of the

problem, also known as the cause mentioned above, reaches the entire population, treating only those patients and the most vulnerable individuals, the tip of the iceberg, is a palliative solution.

- ② The population strategy: targeting the entire population is based on public health thinking to achieve the first-level prevention strategy. The population-wide strategy does not need to determine that individuals are at high or low risk of future disease; only by eliminating harmful exposures, especially those that are unobservable or uncontrolled by individuals or determinants of deleterious exposures in the population, namely reasons for the etiology, taking measures to reduce the level of harmful exposure throughout the population, ultimately reducing the overall burden of disease in the population.

Both the high-risk strategy and the population strategy have their respective strengths and weaknesses. In resolving many problems, the two strategies complement each other and work together.

### Secondary Prevention

Secondary prevention refers to early detection, early diagnosis, and early treatment. The secondary prevention aims at early stages of the disease that symptoms and signs have not yet appeared or are difficult to detect. It aims at having a greater chance of achieving cure by early detection, a timely and appropriate treatment, or at least slowing down the development process of disease and reducing the need for more complex treatment if the disease cannot be cured.

Diseases can be found early via screening, case finding, regular physical examination, and so on. For example, fecal occult blood tests and colorectal screening for colorectal cancer are performed in adults over 50 years of age. In addition, there is periodic screening for HIV in high-risk populations and other routine physical examinations.

At present, most of the etiology of chronic diseases has not yet been established, so full and effective primary prevention cannot be realized. However, the occurrence of chronic diseases is due to the long-term effects of pathogenic/etiologic factors, so it is feasible to do early diagnosis and early treatment.

### Tertiary Prevention

Tertiary prevention, also known as clinical prevention or disease management, is implemented after the symptoms and signs of the disease are manifested. At the early stage of the disease, appropriate treatments are used to relieve symptoms, prohibit further deterioration of the disease, reduce the occurrence and recurrence of acute events, and prevent complications and disability. At the late stage of the disease, the greatest recovery treatments are applied in order to restore body and social function and improve the quality of life and prolong life. Tertiary prevention mainly consists

of symptomatic treatment and rehabilitation treatment. The prevention is aimed at reducing the burden of disease and disability on individuals, families, and society.

Symptomatic treatment can mitigate the adverse effects of diseases with fewer symptoms and prevent the occurrence of complications and disability. Rehabilitation treatment for loss of labor or disability can promote physical and mental rehabilitation, recovery of labor, and make sure they are able to create economic and social value.

### ***10.1.3 Health Protection and Promotion***

With the development of healthy connotations, the goal of public health is no longer just to prevent disease but also to actively maintain and promote health. The main strategies for achieving this include health protection and promotion.

#### **10.1.3.1 Health Protection**

##### Concept of Health Protection

Health protection is to take targeted measures to protect individuals or people from the external environment of harmful substances (such as biological, physical, and chemical harmful substances) on human health threat. Health care covers a wide range of health-related areas: infectious diseases, occupational health, sanitation, radiation hygiene, food hygiene, school health, medicines, medical devices and cosmetics safety, accidental injuries, and emergency public health emergency equipment and treatment.

##### Health Protection Measures

Health protection measures include medical measures, such as immunization and preventive medication as well as environmental engineering measures, economic measures, legal measures, etc. Some scholars, of course, refer to health protection as the latter. Many health protection measures are not available to individuals, and the non-medical measures can be implemented on their own, requiring the joint efforts of government and society.

- ① Eliminate harmful substances in the external environment or control them to the levels that do not adversely affect human health. Such as pasteurization and other processes on raw milk disinfection; the construction industry uses non-hazardous or less hazardous building materials and uses construction techniques, construction equipment, and tools that do not produce or produce less dust. Frequent hand washing is one of the measures of personal hygiene and interference control.

- ② Provide a barrier for individuals. For example, the cab or operation room of construction machinery is closed and isolated, and the air inlet is fitted with a filter device; use of personal protective equipment such as protective clothes, protective gloves, and protective glasses.
- ③ Enhance the individual's ability to fight harmful substances or take measures to prevent the onset or reduce the symptoms of the disease when exposed, such as vaccination, immune serum, or immunoglobulin. Rabies or animals suspected of carrying rabies virus bites generally require 24 h after exposure for the first vaccination. If the bite is in the upper limbs or head, or injury is heavier, anti-rabies virus serum or specific immunoglobulin should also be injected for passive immunization. When people in health care, public security, and other personnel are inadvertently exposed to human immunodeficiency virus (HIV) due to occupational reasons, the short-term anti-retroviral treatment can be used to reduce the possibility of HIV infection, i.e., preventive medication.

### 10.1.3.2 Health Education

Health education is used to help individuals and groups master health care knowledge and to establish a healthy concept through information dissemination and behavior intervention. Under the premise of the access to information and enhanced awareness, voluntary adoption of educational activities and processes that are conducive to healthy behavior and lifestyles.

Health education pays more attention to the process of internalizing the objects of education, highlighting the voluntary nature of individuals to change their behavior. Health education can play a role in tertiary prevention. For example, to inform the public of the common symptoms of TB and encourage timely treatment in the event of suspicious symptoms. For TB patients, it is important to provide them with information on the treatment management, standardize the benefits of treatment under the policy of national free treatment, and improve compliance with standard treatment. With the popularization of health education concept in the world, a lot of health education practice shows that behavior change is a long-term and complicated process. Though some education methods are simple, they have an effect on people's awareness and skills, and then change their way of life. However, environmental constraints and lack of policy may hinder the adoption of healthy behavioral intentions.

### 10.1.3.3 Health Management

Health management is a process of comprehensive supervision of the health-risk factors of individuals or groups; the aim is to get the maximum health for the purpose of minimizing the input. Unlike general health education, health management is based on individual health status evaluation, i.e., according to individual risk factors, individual guidance, dynamic tracking of risk factors, and timely intervention. At

present, health management is mainly used for the prevention of chronic non-communicable diseases such as hypertension, hyperlipidemia, coronary heart disease, stroke, diabetes, obesity, osteoporosis, and cancer.

Health management is the main trend of medical development today. It integrates medical science, management science, and information science and focuses on the concept, connotation, and evaluation of the standard of health, health risk factors monitoring and controlling, health intervention methods and means, health management service model and implementation path, health information technology and standards, etc. The main contents of health management services include: health examination, health monitoring, health risk assessment and intervention, health education and counseling services and medical Internet services, chronic disease risk screening, and tracking management.

#### **10.1.3.4 Health Promotion**

Health promotion is the process of enhancing people's ability to control health and improve their health. It is a comprehensive social and political process that includes not only health education that directly strengthens individual behavior and life skills, making people aware of how to stay healthy; but also through policies, legislation, economic means, and other forms of environmental engineering to improve social, economic, and environmental conditions to reduce their social impact on the public and individual health, thereby creating a socially supportive environment that promotes the maintenance and improvement of health. The WHO indicates that health promotion mainly involves five domains: (1) establishing policies that are beneficial to health; (2) changing the direction of health services; (3) improving individuals' and populations' health knowledge and skills; (4) creating a good physical and natural environment; (5) developing communities' abilities to promote health.

#### **10.1.3.5 Global Health Strategies and Practice**

The development of national health strategy has evolution strategies. In 1948, the WHO put forward that health is a fundamental right of mankind. At the 30th World Health Assembly in 1977, WHO members unanimously adopted a global strategic goal: "Health for all by the year 2000." This state of health allows individuals to live a productive life in both social and economic terms. This goal does not mean that the medical staff provides medical services for all diseases or nobody is sick or develops a disease. Its connotation lies in: (1) Whether at home, school or in other units, people can stay healthy during the period of life and work; (2) People will use more effective methods to prevent disease, reduce the pain caused by unavoidable diseases and disability, and grow healthier, grow old, and finally die happily; (3) All health resources are equally distributed among all members of society; (4) All individuals and families, through their own active participation, enjoy basic health care in an affordable manner; (5) People will realize that they are

capable of getting rid of the burden of disease that can be avoided, shaping themselves and their families' lives, keeping healthy, and knowing that disease is not inevitable.

In 1978, *the Declaration of Alma-Ata* was adopted by the WHO and the United Nations Children's Fund at the International Conference on Primary Health Care, organized in Almaty, which reaffirmed WHO's 1948 definition of health. It also formally put forward the concept of "primary health care," and clearly stated that primary health care is the key and fundamental way to achieve the goal of "health care for all in 2000." The meeting was recognized as a milestone in modern public health.

Primary health care refers to basic health care services that can be affordable in countries and regions, and the methods and techniques used in these services are viable, scientifically sound, and socially acceptable. Each individual and family in the community can access these basic services. The primary health care system should be able to provide appropriate health promotion, disease prevention, diagnosis, treatment, and rehabilitation services for major health problems in different countries and regions. The system is designed to achieve early protection and prevention. In this process, the health sector also covers all aspects of the relevant departments, countries, and social development, especially agriculture, livestock, food, industry, education, housing, public works, transportation, and other sectors, and requires all of these inter-departmental collaborations.

The basic content of primary health care can vary from country to country, but the following eight items should be included at least: (1) carrying out publicity and education for current important health problems and prevention and control methods; (2) promoting food supply and proper nutrition; (3) promoting the provision of sufficient safe drinking water and basic sanitation facilities; (4) providing women and children health care, including family planning; (5) vaccination against major infectious diseases; (6) preventing and controlling endemic diseases; (7) providing proper treatment and management of common diseases and injuries; (8) Providing essential medicine. The 34th World Health Assembly in 1981 adopted the "Global Strategy of health care for all in 2000." It also complements "Use all possible methods to prevent and control noncommunicable diseases and promote mental health through the impact of lifestyles, controlled substances and psychosocial environments."

In 1986, WHO adopted the *Ottawa Charter for Health Promotion* at the First International Conference on Health Promotion, which for the first time fully expounded the concept of "health promotion," principles, and the future development direction, directly promoted the health state; the city's strategy is put forward and practices in the global widespread.

In 2000, the United Nations Millennium Summit adopted *the United Nations Millennium Declaration*, which places health at the heart of the global agenda as the Millennium Development Goals (MDG), the eradication of extreme poverty and hunger, universal primary education, the promotion of gender equality, and empowerment of women, the reduction of child mortality, maternal health, AIDS, malaria, and other diseases' morbidity, in order to ensure environmental sustainability and

global cooperation for development. Countries have raised their health plan in ..., line with national strategies.

The 2030 Agenda for Sustainable Development adopted by all United Nations Member States in 2015 provides a shared blueprint for peace and prosperity for people and the planet now and into the future. As its core are the 17 Sustainable Development Goals (SDGs), which constitute an urgent call to action for all developed and developing countries in a global partnership; SDGs represent the successor to the MDGs, which aim to protect people's health and welfare worldwide.

## **10.2 Public Health Monitoring**

### ***10.2.1 Introduction of Public Health Surveillance***

#### **10.2.1.1 The Basic Concept of Public Health Surveillance**

##### Definition of Public Health Surveillance

Public health surveillance is a persistent, continuous, and systematic process of information collection about health events and problems. After scientific analysis and explanations of important public health information, we should provide timely feedback to people or institutions that need this information to establish a perfect process and to evaluate public health interventions and strategies. Its purpose is to provide decision-makers with basic decisions and to evaluate the effectiveness of those decisions. In a word, public health surveillance is a process of systematic collection, analysis, interpretation, management, and use of public health information which is persistent and ongoing.

Public health surveillance has three basic characteristics or consists of three phases of work:

- ① Health-related information should be collected continuously and systematically in order to discover the distributions and trends of public health issues.
- ② The raw materials should be scientifically sorted, analyzed, and interpreted, then transformed into valuable and important public health information.
- ③ Public health information should be fed back to the relevant departments and staff. In order to achieve the ultimate goals of surveillance, this information should be used timely and fully.

##### Related Basic Concepts and Terminology

- ① Passive surveillance and active surveillance
  - (a) Passive surveillance means that subordinate units routinely report surveillance data to the higher authorities, while higher units passively accept it,



such as the surveillance information system of statutory infectious disease and the surveillance spontaneous reporting system of adverse drug reaction.

- (b) Active surveillance is based on special needs. Higher units require collecting information, such as the omission or concealment of infectious diseases, surveillance of certain behavioral factors (such as smoking and drug abuse) and the US CDC established active surveillance system of foodborne disease (Food Net).

## ② Routine report and sentinel surveillance

- (a) The reporting is mainly carried out by the statutory responsible reporting agencies and professional staff, and the coverage is nationwide.
- (b) Sentinel surveillance means choosing a number of representative regions and/or groups according to the prevailing characteristics of the disease monitored, then monitoring continuously for the purpose of a better understanding of the distributions of certain diseases in different regions, different populations, and corresponding influencing factors based on a unified monitoring program. The most typical project is the sentinel surveillance of AIDS, which refers to selecting the representative areas and AIDS-related high-risk groups and then continuously carrying out the tests for fixed-point, regular, and quantitative HIV antibody according to a unified monitoring program and testing reagents. Meanwhile, workers collect monitoring information of the high-risk behaviors associated with the spread of HIV/AIDS in populations, so as to obtain the information on HIV infection status and behavioral risk factors and trends among different regions and different populations. In addition, monitoring of influenza-like illness (ILI), mainly choosing a number of selected hospital clinics as a monitoring sentinel, and weekly reporting of the number of ILI cases.

### **10.2.1.2 The Purpose and Application of Public Health Surveillance**

The information for public health surveillance may come from a variety of sources, including demographic information and disease information, health and hygiene data, many types of environmental surveillance data, animal-related data, and other relevant information.

## The Purpose of Public Health Surveillance

- ① Describe the characteristics of distribution and trends of health-related events: through continuous and systematic public health surveillance, we can fully understand the characteristics of distribution and trends of health-related events among certain areas and people, which can help to solve the following problems:
  - (I). Quantitatively assess the seriousness of public health issues and identify major public health issues. Important information on health issues needs to be recognized to develop correct and targeted public health policies, plans, or measures for the current or future periods.
  - (II). Find abnormal distribution of health-related events, promptly investigate the causes, and take interventions to effectively curb the development and spread of adverse health events. Long-term and continuous surveillance can help us identify abnormal changes in the distribution of health-related events and then quickly issue an early warning to the health agencies and related units and timely organize and carry out the necessary epidemiological surveys. Once there is an outbreak or epidemic of the disease, we can take appropriate interventions to control further spread of the epidemic.
  - (III). Predict the development trends of health-related events and correctly assess the needs of health service. Through dynamic surveillance and data analysis, it can help to predict the trends and scales of the relevant events and correctly estimate the needs of future health service.
  - (IV). Investigate the influencing factors of the diseases and determine the high-risk populations. Besides, the contents of public health surveillance also include surveillance of behavioral risk factors, environmental pollutants, food safety, and nutritional deficiencies or excesses, and the analysis of this information can contribute to a variety of factors that affect the development of the diseases and to determine the high-risk population of the corresponding diseases, which can provide a scientific basis for developing targeted interventions and reasonable, effective strategies.
- ② Evaluate the effectiveness of public health intervention strategies and measures: public health surveillance is conducted continuously and systematically, the trends of diseases or related events can provide the most direct and reliable basis for assessing the effectiveness of intervention strategies and measures.

## Application of Public Health Surveillance

According to the purposes of public health surveillance, the application of public health surveillance is divided into the following six major types by the WHO in 2002:

- ① Identify one or more cases and intervene to prevent infection or reduce morbidity and mortality.

- ② Assess the impacts of health events on public health or judge and measure its trends.
- ③ Demonstrate the need for public health intervention projects and resources and allocate resources rationally in the development of public health programs.
- ④ Scrutinize the effectiveness of prevention and control methods and interventions.
- ⑤ Identify high-risk populations and geographic areas for intervention and the research of guidance analysis.
- ⑥ Establish the hypotheses and the analytical research of the risk factors that lead to the cause and progression of the disease.

## ***10.2.2 Categories of Public Health Surveillance***

With the development of public health activities, the types and contents of public health surveillance are constantly enriched. At present, the public health surveillance includes surveillance of diseases, cause of death surveillance, surveillance of hospital infections, symptom-based surveillance, behavior and risk factor surveillance, and other public health surveillance.

### **10.2.2.1 Surveillance of Disease**

In the perspective of health issues as epidemiological studies, disease surveillance is a surveillance of outcomes, and surveillance requires an unequivocal diagnosis of the corresponding diseases and death.

#### Surveillance of Communicable Diseases

In 2005, the World Health Assembly approved the international health regulations {IHR (2005)} and began its implementation on June 15, 2007. According to IHR (2005), WHO defines four kinds of diseases that must be notified in any case and defines their corresponding cases. It includes smallpox, polio caused by wild strains, human influenza caused by a new subtype virus, and severe acute respiratory distress syndrome (SARS). It also provides 20 kinds of infectious diseases that received global alert and response, including the 2009 pandemic H1N1 influenza, Ebola hemorrhagic fever, dengue fever, hepatitis, monkeypox, Hendra virus, yellow fever, gram ramea—Congo hemorrhagic fever, Lassa fever, Rift Valley fever, influenza, Marburg, meningococcal disease, Nipah virus infection, avian influenza, plague, smallpox, anthrax, tularemia, severe acute respiratory syndrome (SARS).

According to the Law of the People's Republic of China on Prevention and Control of Communicable Diseases, legally reported communicable diseases are divided into three categories and 40 specific conditions. Any kind of communicable disease occurring and causing death must be reported to local and national Centers for Disease Control and Prevention and should be acted upon by persons responsible for preventing and controlling disease in China.

The main contents and purposes of communicable disease surveillance are as follows:

- ① Timely detect and diagnose cases in order to track and control; discover emerging infectious diseases or new public health problems.
- ② Learn about the distribution of diseases and determine the epidemic or outbreak in time so as to initiate an outbreak survey and control the epidemic situation.
- ③ Monitor the community immunity, serotypes, and genotypes of the pathogen, the species and distribution of the host and vector insects, and the carrying status of pathogens. Know about the changing trends of diseases, identify groups or regions with higher risk factors, and provide information for the formulation and adjustment of strategies and measures to intervene.
- ④ Monitoring the progress and effect of public health intervention projects (strategies and measures).

#### Surveillance of Non-communicable Diseases

With the change of diseases spectrum, the scope of disease surveillance has expanded to non-communicable diseases. According to the main health problems or monitoring purposes in different countries or regions, the monitoring content varies, including malignant tumors, cardiovascular diseases, cerebrovascular diseases, diabetes, mental diseases, occupational diseases, birth defects, etc.

The U.S. National Cancer Institute (NCI) has been monitoring cancer since the 1970s. Centers for Disease Control and Prevention have been promoting health promotion activities for chronic diseases since the 1980s, which is aimed at ten kinds of preventable chronic diseases that seriously affect the quality of life.

Non-communicable diseases such as malignant tumors, cardiovascular diseases, cerebrovascular disease, and birth defects have also been surveyed in some areas of China. For example, the Beijing cardiovascular and pulmonary vascular Medical Research Center organized 16 provinces and cities, 19 monitoring areas in China to survey the trend of cardiovascular disease and its influencing factors (i.e., MONICA project 1984–1993). In 2014, the National Cancer Registry collected data on cancer registration from 234 registries in 2011, including 220 million people. The Chinese cancer registration annual report for 2015 shows that in 2011, the number of China's new cancer cases was about 3,370,000, an increase of 270,000 compared to 2010.

## Surveillance of Hospital Infection

Surveillance of hospital infection refers to a long-term, continuous, and systematic process of observation, collection, and analysis of the occurrence, distribution, and determinates of hospital infection. A good program disseminates data and provides scientific evidence for prevention and control of hospital infection. China's current hospital infection monitoring standard (WS/T312-2009) was issued by the former Ministry of health in April 2009 and began to be implemented in December of the same year. Nosocomial infection surveillance includes comprehensive surveillance, target surveillance, surveillance of bacterial resistance, and surveillance of antimicrobial use. The standard requirement is as follows: establishing the effective hospital infection surveillance and notification system, identifying hospital infectious cases in time; analyzing the risk factors of hospital infection and taking targeted prevention and control measures; training hospital infection control full-time personnel and clinical medical staff on the awareness and ability to identify an outbreak of hospital infection. When the outbreak occurs, the source of infection and the route of infection should be analyzed by epidemiological methods and effective control measures should be taken. According to the different circumstances of the outbreak of nosocomial infection, reports should be given to the health administrative department within the prescribed time.

## Surveillance of Death

The purpose of death surveillance is to understand the mortality and distribution of death causes. Through the statistical analysis of death causes, the health level of the monitoring population can be assessed, and the main causes of death and disease control in different periods can be determined. "National Maternal Death Monitoring Network" and "National Death Monitoring Network for Children under 5 years of age" were established, respectively, in 1989 and 1992. Monitoring information is used to reflect the health status of women and children in China. China CDC formulated and issued the "National Death Surveillance System for Disease Surveillance System (Trial)" and the "National Death Registration Information Network Report Specification (Trial)" respectively in 2005 and 2007, which made the death monitoring work more standardized. The medical certificate of death is an important evidence for the death report and statistical analysis, and the correct judgment of death causes is the most important basis for monitoring the cause of death.

### 10.2.2.2 Symptom Surveillance

Symptom-based surveillance is a continuous and systematic process for collecting and analyzing data on certain groups of symptoms so authorities can mount a quick response to rapidly detect and predict the occurrence of a disease. Symptom surveillance is especially suitable for some new diseases with unknown etiology, and

the case has no definite diagnosis method. The surveillance of influenza is a component of syndrome surveillance and played an important role in the surveillance of influenza A (H1N1) pandemic in 2009–2010.

The common symptom surveillance includes influenza symptoms (cough, sneezing, surveillance, fever surveillance, diarrhea surveillance, etc.). Symptom surveillance does not depend on specific diagnosis; its emphasis is on the surveillance of diseases with specific symptoms. The content of the surveillance is not only the clinical symptoms (such as fever, diarrhea, respiratory symptoms, etc.) it also includes many disease-related phenomena.

There are mainly about:

1. The hospitalized features of hospital emergency room or outpatient.
2. Sales situation of OTC drugs (such as vitamin C, cold medicine, antidiarrheal drugs, etc.).
3. Sales volume of medical related supplies (such as medical masks, sanitary napkins, etc.).
4. Absenteeism rate of school or unit.
5. Disease or death in animals.
6. Changes in biological vectors.

The classification of symptoms and the diagnosis of symptoms are the basic components of the symptom surveillance system. More respiratory symptoms, gastrointestinal symptoms, skin symptoms, and neurological symptoms have been continuously used for symptom surveillance. Syndrome surveillance does not rely on the clinician's ability to consider and detect a specific disease or on the availability of local laboratory or other diagnostic resources. Since syndrome surveillance focuses on syndromes rather than the diagnoses and suspicious diagnoses, it is less specific and more likely to identify multiple individuals without the disease of interest. As a result, more data have to be handled, and the analysis tends to be more complex. So far, the symptom surveillance in China has been tested and evaluated in public health surveillance of many important social activities, such as the Beijing Olympic Games, Shanghai World Expo, Guangzhou Asian Games, and so on.

## ***10.2.3 Methods of Public Health Surveillance***

### **10.2.3.1 Surveillance Methods**

Public health surveillance is the core function of public health practice. Public health surveillance requires the establishment of specialized monitoring organizations, which should have the appropriate administrative and technical conditions and the funding required ensuring the operation. The surveillance system is an organized and planned surveillance system for a particular disease or a public health problem that can be used separately or simultaneously on a population-based, hospital-based, and

laboratory-based surveillance method. In addition, there are case-based surveillance, indicator-based, and event-based surveillance. Monitoring-related activities include the collection of data, analysis of data, dissemination of information, and the use of information.

### Population-Based Surveillance

Population-based surveillance is where the specific populations and the dynamics of the particular diseases are monitored. It can not only be a regular report monitoring that covers the entire population but also a monitoring point or sentinel surveillance, and with good representative sentinel surveillance, which can obtain more accurate, reliable, and timely information that is less costly and efficient. Many behavioral risk factors are monitored by population-based surveillance.

### Hospital-Based Surveillance

Hospital-based surveillance refers to the hospital as the scene and the patient as the monitoring object, mainly on hospital infection, pathogen resistance, and birth defects. The monitoring of statutory infectious disease reporting systems and passive surveillance of adverse drug reactions fall into this surveillance.

### Laboratory-Based Surveillance

Laboratory-based surveillance is mainly the use of laboratory methods to monitor pathogens or other pathogenic factors. Such as WHO and China's influenza laboratory testing system, carried out routine influenza virus isolation and classification for identification work, i.e., laboratory-based influenza virus surveillance. Rapidly developed in many countries, pathogen molecular subtyping pulse net is a laboratory-based pathogen detection that covers almost all of the world's major countries. Laboratory-based surveillance is beneficial in increasing the use of multiple reverse transcription polymerase chain reaction and faster pathogen identification in the twenty-first century.

### Case-Based Surveillance

Case-based surveillance is mainly referring to the disease prevention and control systems as the main body, jointly by clinical care institutions and other health care units on the special case cases and aggregation of cases of monitoring. Analyzing the number of episodes of disease outbreaks is often easier and more practical than investigating individual cases, especially for diseases that are potentially at risk of failure, poor quality reporting, or clinical type of disease. China's public health

emergency surveillance, food safety incident surveillance, and so on, belong to case-based surveillance.

### Indicator-Based Surveillance

A variety of surveillance systems can collect quantitative data, such as statutory infectious disease reporting systems, symptom surveillance systems, behavioral risk factor surveillance systems, etc., which can provide quantitative data for epidemic/outbreak intelligence mechanism(EIM).

### Event-Based Surveillance

Collecting information from media and web search, news analysis, domestic and foreign newsletters, public complaints and reporting, health advice, etc., can also provide clues and basis for EIM, a public health incident reporting system is an event-based surveillance system. However, there is no uniform standard for the method of verification of an incident and procedures for reporting incidents, currently, and event-based surveillance has not yet been established in most countries.

## 10.2.3.2 Surveillance Methods and Techniques

The correct use of monitoring methods and techniques in the monitoring process helps to improve the quality and efficiency of surveillance. With the development of modern information technology, computer networks, geographic information systems, and other technologies that are more and more used in public health surveillance, so that surveillance information collection, collation, analysis, transmission, and feedback become more convenient, greatly improving the surveillance system's work efficiency, and also making the development of public health strategies and the implementation of interventions more timely. In addition to active surveillance and passive surveillance, routine reporting, and sentinel surveillance, the commonly used surveillance methods and analytical techniques are also included.

### Case Registration

Case registration is the registration, detection, diagnosis, and information registration of daily work surveillance-related cases, and so on. It is the basis of surveillance work, and it is also an essential part of disease and death surveillance.



## Unrelated Surveillance

When the purpose of surveillance is to understand the prevalence of a disease in the population, rather than to identify specific cases, the information collected by other studies can be used to monitor without identifying the individual, known as unrelated anonymous surveillance. Such surveillance can waive ethical issues to a certain extent. For example, to collect blood samples from the hospital laboratory and to identify HIV status by HIV antibody testing without identifying individual identities.

## Record Linkages

The data from two different sources are connected to form a new database for statistical analysis in order to obtain more valuable surveillance information, which is called a record connection. For example, in the absence of information on the future incidence or death and no record of birth weight in the infant death data, you can get different birth weight infant mortality information through the two data link analysis.

## Collect Surveillance Information Online

Using computer-assisted telephone interviewing system (CATI) and network surveys, respondents can use short time and less cost to get more quality access to data, and the data can be used directly by various statistical software, automated data management, and automatic statistical analysis, which quickly and easily complete the surveillance information collection and analysis. CATI applies high-speed communication technology and computer information processing technology to traditional telephone interviews, more than half of the developed countries in Europe and the United States adopt CATI technology for surveys. There are also examples of the application of CATI in China, such as the use of influenza-like symptoms, the situation of seeing a doctor, and the diagnosis of influenza A H1N1 in public in Beijing. Network survey, also known as online survey, refers to the internet, and its investigation system will be the traditional paper survey and analysis methods online, intelligently. These two surveys can greatly expand the number of surveys and geographical scope, save survey costs and improve the efficiency of data management and statistical analysis.

## Network Direct Reporting System

With the rapid development and popularization of computer network technology, an increasing number of network direct reporting systems have been established and adopted in public health surveillance systems. Chinese public health emergency

surveillance system and the statutory infectious disease reporting information system have achieved a direct network, at the county and township level, greatly reduced the time of information transmission, and laid the foundation for the rapid processing of data.

#### Automatic Warning Technology

Early warning is the response based on abnormal information from the surveillance, such as an abnormal increase in a particular disease case so that the relevant departments and those who may be affected by the incident respond in a timely manner. Automatic warning technology is the use of mathematical models and computer information technology through a specific algorithm to determine the warning threshold, and an automatic detection of possible abnormal information (above the threshold), as well as an early warning signal is then issued. It is worth noting that the level of early warning threshold directly affects the sensitivity and specificity of the warning signal, the warning signal prompted by the suspicious events need further analysis and verification.

#### The Use of Geographic Information System

The use of geographic information system (GIS) makes public health surveillance data more visible in the regional distribution, which helps to analyze the geographical environment and climate factors on the impact of public health problems.

### **10.2.3.3 Attention in Public Health Surveillance**

#### Case Definition and Disease Surveillance

In large-scale surveillance work, strictly following the clinical diagnostic criteria to determine a disease case is often limited by working conditions and difficult to operate. So in order to determine a unified, highly operational monitoring standard is extremely important, cases identified using monitoring criteria are called surveillance cases. For example, the diagnosis of many infectious diseases is mainly based on clinical diagnosis and characteristics, and not necessarily needing a pathogen test.

Many of the cases were reported in Chinese legal infectious diseases belong to surveillance cases. Proportion of the actual cases in monitored cases should be increased as far as possible, and this proportion can be estimated to a certain extent.

## Static and Dynamic Populations

In the process of monitoring, the population in which no people move out and displace is called a fixed population. If there are only a small number of births, deaths, immigrations and emigrations, and relocations in a region with more population, it can also be regarded as a fixed population. For the static population to calculate the sample rate, the average population of the observation period can be used as the denominator. If people frequently move in or out of the population during the monitoring process, it is a dynamic population. For the calculation of the dynamic population rate, it needs to use the total person-hours observed as a denominator.

## Surveillance Information in-Depth Analysis, Exchange, and Sharing

The use of automatic analysis technology can be more efficient and clearer when analyzing the data and obtain more meaningful information; according to the monitoring data and analysis information, a timely and effective early warning can be released through automatic early warning technology. Different surveillance systems can achieve a certain amount of information through sharing and exchange of information and greatly improve the effectiveness of monitoring work.

## Confidentiality System/Surveillance Ethics

Many diseases involve personal privacy, and some patients or people infected with the disease may be subjected to social discrimination. The monitoring of these diseases should strictly comply with the confidentiality system. The risk of supervision may be at the individual or collective level. If information about health is not properly released, people may feel embarrassed and discriminated against. On the one hand, we need to maintain the monitoring of the dignity and rights and interests of the subject. On the one hand, we need to maintain the monitoring of the dignity and rights and interests of the subject. On the other hand, we need to enhance the public knowledge of the activities of the monitoring and the awareness of participation.

## ***10.2.4 Procedures and Assessment of Public Health Surveillance***

### **10.2.4.1 Basic Procedures of Public Health Surveillance**

Core activities of public health surveillance systems include collecting, analyzing, disseminating, and utilizing the information of monitored public health events.

### Systematically Collecting Relevant Data

Relevant data about public health surveillance are collected systematically and comprehensively according to the specific purpose of the different supervising systems. The data collection process should observe the uniform criteria and scientific methods, and normative schedule. In conducting surveillance, there are various approaches to gather data about public health problems, including demographic information, disease incidence or death certificates about population, laboratory reporting about pathogen and serology, investigations about risk factors, records of the intervention (e.g., distribution of vaccines, supply of iodized salt) measures, special surveys, and other relevant data.

### Managing and Analyzing Collected Data

Data management refers to checking and classifying gathered data to ensure data integrity and accuracy.

Data analysis refers to the transformation of various data into relevant indicators using statistics technology. Analysis of surveillance data is usually conducted to characterize the distribution, trend, and influences about the public health problem under surveillance. In the process of data analysis, on the one hand, it is necessary to select the correct statistical methods according to the nature of the data to fully appraise and apply the data. On the other hand, the influence of various factors on the results of disease surveillance must be considered in order to arrive at correct and reasonable explanations.

### Dissemination and Feedback of Information

These data and surveillance reports must be shared with those who supplied the data and those who are responsible for controlling public health problems. The timely, regular dissemination of basic data and their interpretations is very important for surveillance. For example, the World Health Organization collects and analyzes all aspects of monitoring data regularly, disseminating information on public health surveillance by “Weekly Epidemiological Report” and other publications around the world. The use of the Internet to distribute information is a new development for public health surveillance. The Public Health Information Network initiative sponsored by the Centers for Disease Control and Prevention, is heavily standards-based and based on the HL7 Reference Information Model.

The role of information feedback cannot be overlooked. Information feedback is a bridge linking public health surveillance and intervention. The channels of information feedback must be established to disseminate the information from disease surveillance in a timely fashion to all relevant organizations and individuals so that they are able to rapidly respond to health problems. The monitoring information can be disseminated in both vertical and horizontal directions. The vertical direction

refers to the monitoring information reported to higher-level health administrative departments and managers. The horizontal direction refers to the monitoring information transmitted to the lower-level monitoring institutes and experts and communities as well as residents. The content and manner of feedback should be different depending on the objectives.

### Information Utilization

Information obtained from monitoring can be used to describe the distribution characteristics of public health problems, to identify whether the prevalence exists, to predict the trend of prevalence, to evaluate the effect of interventions, and to provide a basis for making public health activities. Data and interpretations derived from surveillance activities are useful in setting priorities, planning and conducting disease control programs, and assessing the effectiveness of control efforts. It is the ultimate goal of public health monitoring to make full use of monitoring information, to develop strategies of public health in a timely manner, and to take effective interventions.

#### 10.2.4.2 Evaluation of Public Health Surveillance System

The quality and effectiveness of public health surveillance systems need to be regularly evaluated in order to improve the effectiveness of public health monitoring systems and to better serve public health activities.

#### Quality Evaluation of Monitoring System

Evaluation of public health surveillance programs can be based on a number of important characteristics.

**Completeness** It refers to the diversity of monitoring content or indicators contained in the monitoring system, which include the integrity of the monitoring forms, the integrity of the case reports, and the monitoring data.

**Sensitivity** It refers to the ability of the monitoring system to identify public health problems. This assessment mainly involves comparing the proportion of cases reported by the monitoring system to the number of actual cases and the ability of the monitoring system to detect when an outbreak or epidemic of a disease (other public health events) has occurred.

**Specificity** It is used to reflect the ability of the monitoring system to exclude non-public health problems. Such as the ability of the monitoring system to correctly identify the random fluctuations in the phenomenon of the disease population, thereby avoiding or reducing the false alarms in public health monitoring.

**Timeliness** This attribute refers to the time interval from the occurrence of a public health event to the relevant department receiving a report. It reflects the speed of information dissemination and feedback of the monitoring system. It is very important for acute communicable disease surveillance, which may directly affect the efficiency of the intervention. Measurement of timeliness in surveillance systems needs to be planned for in system design.

**Representativeness** It refers to the extent to which the public health problem identified by the monitoring system can represent the actual occurrence of the target population. Lack of representativeness may lead to a waste of health resources.

**Simplicity** Are forms easy to complete? Is data collection kept to a necessary minimum? The method for performing surveillance typically should be as simple as possible and achieve the purpose of monitoring.

**Flexibility** It refers to the ability of the monitoring system to respond to new public health issues and to adjust the operational procedures or technical requirements in a timely manner to meet new needs.

### Benefit Evaluation of Monitoring System

Benefit evaluation of monitoring system consists of important characteristics.

**Health Economics Evaluation** The establishment and operation of any monitoring system requires costs investment, sometimes even costly. The effectiveness of the monitoring system is mainly reflected in the early warning and timely responses to the disease or event, the prevention, and control of the disease, and the improvement of the people's health levels. Therefore, the evaluation of health economics is indispensable.

**Positive Predictive Value** This refers to the proportion of reported or identified cases accounted for are real cases. To what extent are the reported cases real cases? To what extent are measured changes in trends truly reflective of events in the community?

**Acceptability** It reflects the enthusiasm of individuals and organizations to participate in a surveillance system. Acceptability is influenced substantially by the time and effort required to complete and submit reports or perform other surveillance tasks.

**Interconnection and Sharing Ability in Monitoring Systems** Most of the monitoring systems are established for specific purposes, so there may be some limitations in collecting and utilizing information. The establishment of the interconnection and sharing between monitoring systems is able to greatly improve the working efficiency, information utilization, and reduce the waste of resources. It is quite important for the monitoring system to realize the importance of interconnection and sharing between monitoring systems.

In addition to the evaluation of quality and benefit, the function of the monitoring system should be evaluated. The function of the monitoring systems consists of core functions and support functions. The core functions include case monitoring, case registration, case confirmation, reporting process of pathology, data analysis and interpretation, epidemic warning, and information feedback. The support functions refer to those conditions that can facilitate the achievement of core functions, including the implementation standards and guidelines of monitoring system, the training and supervision for relevant personnel and organizations, the necessary communication equipment, the necessary resources of human and material as well as the monitoring, assessment, and coordination of monitoring systems.

# Chapter 11

## Communicable Diseases Epidemiology



Rongguang Zhang

### Key Points

- Epidemics of communicable diseases have become crucial public health problems in recent decades.
- Epidemic process is the process in which a communicable disease is transmitted among individuals in a specific population. Investigation of the epidemic process forms fundamental basis for prevention of the communicable disease.
- In most cases, both population strategy and high-risk strategy are taken for obtaining a high efficacy of prevention and full use of the public health resources.
- Surveillance of communicable diseases play an important role in prevention and control of these diseases and evaluation of efficacy of the strategies and measures used.

In history, communicable diseases had ever been the most serious threaten to human life and health until the end of World War II. Then the situation ameliorated accompanying with the advancement of biologic theory and techniques as well as the enhancement of human living conditions. In recent decades, however, epidemics of communicable diseases have become crucial public health problems once again. Certain newly identified infectious diseases and previously controlled communicable diseases have been emerging or recurring in part due to the variation of pathogens, changes in social and environmental conditions as well as human living modes.

---

R. Zhang (✉)

International School of Public Health and One Health, Hainan Medical University, Haikou, China

e-mail: [zrg@zzu.edu.cn](mailto:zrg@zzu.edu.cn)



## 11.1 Infection Process

### 11.1.1 Infection Process

Infection process refers to the process of a pathogen's invasion into and interaction with human body. By infection process, it is also meant the entire process from onset and development to the end of an infectious disease in individuals. Through the process, an infection may result in various outcomes in individuals.

### 11.1.2 Spectrum of Infection

Spectrum of infection is the composition of all the various outcomes caused by an infection in population (Fig. 11.1). Invasion by a species of pathogen may cause a variety of outcomes in an infected population, such as latent infection, incubatory infection, apparent infection and death, although only one outcome is exhibited in an infected individual. Spectrum of infection is an important epidemiologic characteristic of a communicable disease, defined on outcomes of infected populations. For instances, most persons infected with poliomyelitis or encephalitis B pathogen have only unapparent infection, majority of those with measles or crystalli become clinical patients while suffering from rabies or AIDS commonly results in death.

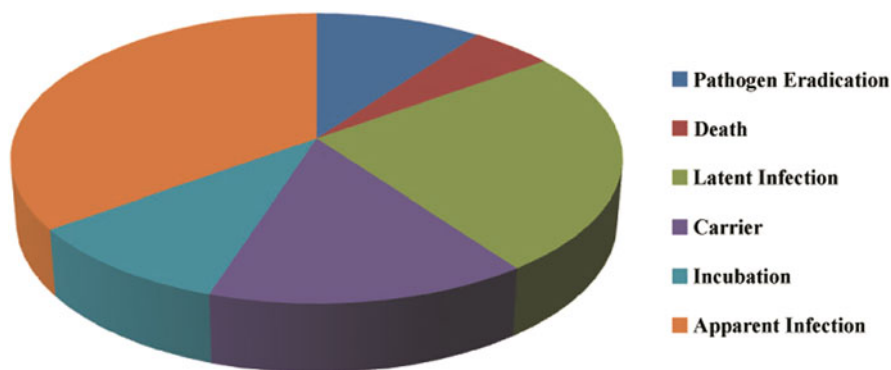


Fig. 11.1 Spectrum of an assumed infection

## 11.2 Epidemic Process

### 11.2.1 Definition

The epidemic process is the process in which an infectious disease is transmitted among individuals in a specific population. Insight into epidemic processes of communicable diseases enables us to make appropriate strategies and measures for prevention and control of infectious disease transmission.

### 11.2.2 Three Links in the Epidemic Process

An epidemic process is constituted by three links, namely, reservoir of infection, route of transmission and susceptible host.

#### 11.2.2.1 Sources of Infection

Humans and animals infected with pathogens can act as sources of infection. The process of apparent infectious diseases can be divided into such three periods as incubation period, clinical stage and recovery (convalescent) period. The infectious stage consists of the later stage of the incubation period, clinical stage and the early stage of convalescent period, in which the infected persons discharge pathogens and play the roles of infection sources (Fig. 11.2). In the early stages of the incubation periods and the later stages of the convalescent periods of most infectious diseases, the infected persons contribute no or little to the transmission of the diseases. The carriers of certain infectious diseases like hepatitis B, which have no clinical manifestations after infection, can also be important sources of infection. For

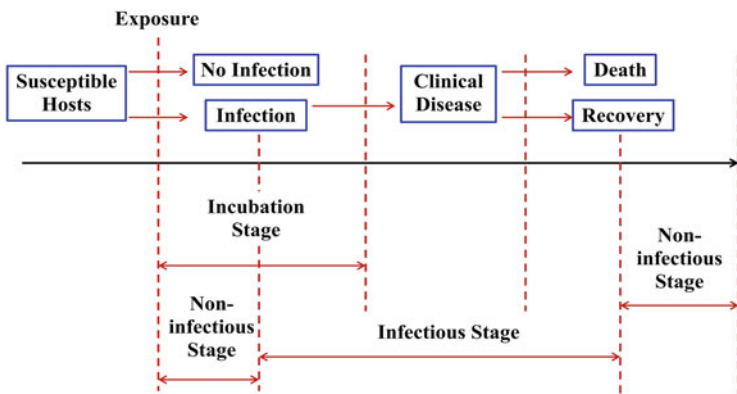


Fig. 11.2 The infectious stage in an infection process

zoonoses like schistosomiasis, clonorchiasis, paragonimiasis and rabies, the animals suffering from these diseases are also critical sources of infection.

### 11.2.2.2 Routes of Transmission

Routes of transmission are the ways by which pathogens are transmitted in the environment outside their hosts from being discharged to invading new susceptible hosts. Various communicable diseases are transported via different manners, which are of significance for formulating strategies and measures to block the transmission.

1. Airborne transmission: Respiratory infectious diseases are mainly transmitted by respiration. Droplets, produced by respiration and especially sneezes, may carry a large number of pathogens and can cause infection when in contact with the respiratory tracts of susceptible masses. Droplets will become droplet nuclei after floating in air for a time with water evaporated. Droplet nuclei are smaller than droplets, capable of floating in air for longer time, and what is more serious, droplet nuclei can reach the deep part of respiratory tracts, and thus easier to cause infection. Pathogens can also be carried by dust floating in air and infect human respiratory system.
2. Waterborne transmission: Many communicable diseases are transmitted by drinking water or contacting polluted water, involving gastrointestinal infectious diseases like cholera, giardiasis, entamebiasis and parenteral parasitic infections like schistosomiasis.
3. Foodborne transmission: Susceptible masses can catch certain gastroenteral infectious diseases and few respiratory diseases through ingestion of food containing pathogens. The epidemiological characteristics of food transmissions are as follows: (1) the patients share the same history of food taking, (2) the persons without ingestion of the food are not infected, (3) the patients have short incubation periods, (4) the outbreak may take place if a large population are exposed to the food and (5) outbreak or epidemic of the communicable diseases will subside following stopping the food supply.
4. Transmission by contact: Communicable diseases can spread via contact with sources of infection or pathogen-polluted environments. Contact transmission includes direct contact transmission and indirect contact transmission. The former means that infectious diseases like sexually transmitted diseases and rabies are transmitted by direct contact with infection sources without the participation of any other environmental media. The later refers to that susceptible masses are infected by exposure to some environmental factors like polluted public equipment and places (e.g., doorknobs).
5. Arthropod-borne transmission: Arthropods can carry or, as hosts, harbor certain pathogens and cause infection, so arthropod-borne transmission can be divided into mechanical transmission and biological transmission. During mechanical transmission, pathogens are only transported by arthropods without any changes in morphology and quantity. In biological transmission, the arthropods harbor

pathogens as their hosts, and the pathogens might develop or multiply during the transmission process. For instance, malaria and filariasis are biologically transmitted by mosquitoes, while giardiasis and cholera are mechanically spread by flies' transportation.

6. **Soilborne transmission:** Soilborne transmission refers to the routes that infectious diseases are transmitted by contacting soil polluted with pathogens, such as parasites' infective eggs and larvae and certain bacterial spores. Hookworm infection, ascariasis, trichuriasis, anthrax and tetanus are common soilborne communicable diseases. The epidemiological importance of this transmission depends on the vitality of the pathogens in soil, possibility for human contacting soil, personal hygiene habit and working conditions.
7. **Iatrogenic transmission:** Iatrogenic transmission is caused by medical and health care workers when performing diagnosis and treatment for patients without complying with the regulations and operating procedures. For instance, hepatitis B and AIDS can be transmitted by blood transfusion and invasive body examination.
8. **Vertical transmission:** Vertical transmission refers to the routes by which offspring are infected from their mothers by placenta, upstream infection or delivery. For example, infants can catch toxoplasmosis, hookworm diseases, hepatitis B, gonorrhea and herpesvirus infection by vertical transmission.
9. **Multiple transmissions:** In most cases, a communicable disease can be transmitted by more than one route. For instance, human can get infection with hookworm by contacting soil or drinking water polluted with the filariform larvae.

### **11.2.2.3 Herd Susceptibility**

Herd susceptibility is the mean level of susceptibility of a population to an infectious disease. Herd susceptibility is closely and negatively relevant to the population's immunity against a pathogen's infection. The factors for an increase in herd susceptibility includes enhanced quantity of newborn infants, immigration of easily infected masses, decreased quantity of individuals with immunity and deaths of immunized masses. Whereas, planned immunization and epidemic of infectious diseases can contribute to decrease in herd susceptibility.

### ***11.2.3 Two Factors Affecting the Epidemic Process***

Epidemic outbreak and intensity of an infectious disease depend on the coexistence and association of three links, and can vary from changes in any of them. However, both natural and social factors can influence the epidemic of infectious diseases through interaction with the three links.

### 11.2.3.1 Natural Factors

Natural factors include those related to geographic characters, climatic conditions, soil, animals and plants, and contributing to epidemics of infectious diseases through complicated mechanism. Certain geographic and climatic conditions have great influence on growth, reproduction and habits of the vectors, and thereby affect the transmission of communicable diseases such as malaria, filariasis and epidemic encephalitis B. Climatic factors can also play key roles in epidemics of infectious diseases by taking effects on human behaviors. In recent decades, accompanying the global warming, remarkable changes in the breeding area, growth and multiplication of mosquitoes and flies and enhanced virulence of the pathogens inside arthropods have been observed.

### 11.2.3.2 Social Factors

Social factors involve almost all sorts of human actions, such as hygienic habits, epidemic prevention work, health care conditions, living and nutritional conditions, inhabitant environment, production activities, occupation, culture, custom, religion, migration of populations, steady social environment and so on. In recent decades, certain epidemics of newly emerging and reoccurring diseases were affected by social factors. Social factors can aggravate or alleviate epidemics of communicable diseases. For instance, unreasonable use of antibiotics provokes resistance against antibiotics, resulting in epidemics of infections with drug-resistant *Mycobacterium tuberculosis*. Abuses of insecticides have accelerated the development of mosquitoes' anti-insecticides resistance, and thus aggravated the epidemic of malaria, dengue fever and yellow fever. Urbanization and population explosion have caused increase in epidemics of human infectious diseases, and wars, unrests, congestion of displaced people, famine and rapidly expanded global tourism and industrialization can play important roles in epidemic of communicable diseases.

## 11.2.4 Epidemic Focus and Epidemic Process

### 11.2.4.1 Epidemic Focus

An epidemic focus consists of a source of infection and a region that the pathogens discharged from the infection source can reach. Multiple epidemic foci in geographic fusion form an epidemic district.

An epidemic focus can be developed under two necessary conditions, namely, existence of a source of infection and a route of transmission. The ranges of epidemic foci vary from the species of infectious diseases, especially the transmission modes and arthropod vectors. Commonly, air transmission of infectious diseases can

produce a wider range of epidemic focus. For an arthropod-transmitted disease, the range of the arthropod's activity can become the range of an epidemic focus.

An epidemic focus will be extinguished under the following three conditions: (a) removal of the infection sources, for example, death or migration of the patient; (b) disinfection of the environment polluted by the pathogens from the sources of infection; (c) all the contactees have been proved uninfected or not become new infection cases in the longest incubation period.

#### **11.2.4.2 Epidemic Process**

A series of epidemic foci with contact to each other that appear early or late constitute an epidemic process. An epidemic process is also a process during which an epidemic of infectious disease occurs and expands in a population, and the sources of infection, routes of infection and herd susceptibility of infection link to each other. Investigation of the epidemic process forms the fundamental basis for the prevention of a communicable disease.

### **11.3 Strategy and Implementation**

Communicable diseases cause serious damage to human health, and are also the predominant causes of death in many developing countries. In recent decades, prevalence rates of global infectious diseases have, to a large extent, been rising, outbreaks of infectious disease epidemic have been continually reported, certain ever-controlled communicable diseases have reoccurred and many reemerging diseases have been identified. Therefore, prevention and control of infectious diseases should still be emphasized in public health care work throughout the world. To improve the efficacy of these works, governments of all countries have to formulate practical strategies and measures for prevention and control of the communicable diseases according to their national conditions.

#### ***11.3.1 Strategies for Control of Communicable Diseases***

For control of infectious diseases, the following policies have been suggested to implement:

1. Establishment of thoughts for long-term struggles against communicable diseases.
2. Prevention of infectious diseases should be carried out as the most predominant action.
3. Comprehensive measures should be taken according to the epidemic processes.

4. Formulate and implement laws and regulations for control of infectious diseases.
5. Prevention of transmission of certain emerging infectious diseases like HIV/AIDS from high-risk to general populations.
6. Establishment of rapid response mechanism and organization for managing emergency public health events.
7. Establishment of modern epidemical working staff to solve the problems in the present situation on communicable diseases.

#### **11.3.1.1 Population Strategy**

Population strategy refers to taking prevention measures on whole populations for lowering their levels of exposure to risk factors for communicable diseases. For instance, 15 infectious diseases including hepatitis B, tuberculosis, poliomyelitis, bronchocephalitis, diphtheritis, tetanus, leprosy, measles, parotitis and epidemic encephalitis B, hepatitis A and leptospirosis are prevented by planned immunization of all children in China.

#### **11.3.1.2 High-Risk Strategy**

High-risk strategy is meant allocation of limited public health sources to the population at high risk for certain communicable diseases, in order to enhance the cost-benefit of prevention works. In most cases, two pronged strategies are taken for obtaining a high efficacy of prevention and full use of the public health resources.

Up till now, eradication of human smallpox is thought as the best example for implementing the two strategies to fight against communicable diseases. In 1958, the World Health Assembly approved the plan for the global eradication of smallpox, and made up the vaccination strategy to elevate the inoculation rate of the whole population. In 1967, when the infection rate decreased apparently, the high vaccination rate was proved to contribute little to preventive efficacy of smallpox by public health surveillance. From then on, WHO implemented a new strategy involving enhanced surveillance of smallpox cases and encircling inoculation, namely, when finding new cases by surveillance, inoculation was performed on all the contacts of the cases and blocked the transmission of smallpox. At last, the epidemic of this serious disease was globally extinguished in 1979.

## ***11.3.2 Measures for Control of Communicable Diseases***

### **11.3.2.1 Surveillance of Communicable Diseases**

Surveillance of communicable diseases, as one of the main tasks of public health surveillance, plays an important role in the prevention and control of these diseases and evaluation of efficacy of the strategies and measures used.

Communicable disease surveillance can be divided into four levels: the local, the regional, the national and the international infectious disease surveillance, which take the responsibility for continually and systematically collecting and applying information on infectious diseases from cities (or towns or villages), provinces (or states or counties), countries and global area, respectively, in a long term.

Performance of communicable disease surveillance consists of four steps: collecting information on infectious diseases, collating and analyzing the data, giving feedback of the information to the populations who need it and applying the information in the formulation of strategies and measures and evaluation of their effectiveness.

Surveillance activities of communicable diseases include timely reporting legally notifiable infectious disease cases, morbidity and mortality, results of field investigations, laboratory isolation and identifications of pathogens and data on vaccinations and immunity levels in populations.

The legally notifiable infectious diseases consist of a variety of serious communicable diseases, which are different in different countries or times. In China today, there are 40 species of legally notifiable infectious diseases, which have been grouped into three categories: category A involves plague and cholera; category B includes 27 diseases, such as SARS, AIDS, viral hepatitis, poliomyelitis, highly pathogenic avian influenza, H7N9 avian influenza, measles, epidemic hemorrhagic fever, epidemic encephalitis B, bacterial and amoebic dysentery, tuberculosis, typhoid and paratyphoid, gonorrhoea, schistosomiasis, malaria, anthrax and so on; category C comprises 11 diseases like epidemic influenza, epidemic mumps, rubella, leprosy, leishmaniasis, echinococcosis, filariasis, hand, foot and mouth disease and so on.

The surveillance information should essentially be handled with respecting to security and confidentiality of individual privacy. Only those who need to use the individual information for the public health practice can have access to these data. Presently, although the privacy of personal information has drawn more attention than ever, more effort should be made to manage these data in a confidential and secure way.

### **11.3.2.2 Measures on Sources of Infection**

Patients, carriers and reservoir hosts infected are the sources of infection. For eradication of the infection sources, early discovery, early diagnosis, early reporting,



early isolation and early treatment should be performed for the patients. Once a person is suspected suffering from one of the category A infectious diseases or one of the two category B diseases, SARS and pulmonary anthrax, he/she must be medically treated in isolation. The length of the isolation period should be determined according to the results of medical examinations. The carriers should also be given treatment, and whether isolation is carried out depends on the species of communicable diseases they have been infected with. As for the animal host of pathogens, some of them like mice, invaluable for economic and environments, should be killed, while some like cattle and dogs, valuable for economic and production, can be medically treated or exterminated, according to the harm the infectious diseases probably present to human health.

### **11.3.2.3 Measures on Routes of Infection**

Measures should be taken for disinfection of the environments polluted from the sources of infection. Various measures have to be used for blocking different transmission routes of pathogens, for example, for prevention of intestinal infectious diseases, which are transmitted mainly by feces to mouth route, disinfection has to be carried out mainly on patients' discharges, polluted water, rubbishes, polluted goods and environments. As for the infectious diseases of the respiratory tract, improving ventilation, air disinfection and individual protection with a nose mask can be useful measures for prevention of infection; for blood-borne communicable disease like AIDS and hepatitis B, taking safe sexual action, avoiding drug taking, avoiding sharing of public syringes and strengthening management of blood transfusion are practical measures.

### **11.3.2.4 Measures on Susceptible Populations**

Susceptive masses exposed to sources of infection are at high risk for infection, can probably become new epidemic foci, contributing to the transmission of communicable diseases in populations. Preventive immunization, drug-based prevention and individual protection are commonly used measures for protection of susceptible populations.

1. Preventive immunization: Preventive immunizations are carried out prior to epidemics of infectious diseases to enhance specific immunity of populations and thereby prevent outbreaks of the epidemics. Preventive immunizations have become the most important approaches to prevention of communicable disease epidemics and classified into active immunization and passive immunization. Vaccinations that induce long-term immunity via active immunization have proved the most reasonable and effective routes for prevention and control of infectious diseases. However, for many serious communicable diseases,

developments of safe and effective vaccines have been great challenges that humans face.

2. **Drug-based prevention:** Drug-based prevention is performed on susceptible masses as an emergency measure when there is an outbreak of infectious diseases and there are specifically effective drugs for these diseases. For example, certain antimalarial drugs can be administered to susceptible populations when an epidemic of malaria takes place. Preventive administration of drugs can produce short-term and unsteady efficacy against infection, with the development of antidrug resistance.
3. **Individual protection:** Individual protection of susceptible masses plays an important role in the prevention of communicable diseases. For instance, in epidemic seasons of respiratory infectious diseases, avoidance of staying in densely populated places, improved ventilation of inhabiting and working places and wearing nose masks in contact with patients can significantly reduce the possibility of infections. Also, infections of sexually transmitted diseases like AIDS can to an extent be prevented by using condoms.

## **11.4 Immunization Program and Effectiveness**

### ***11.4.1 Immunization***

Immunization is the inoculation of susceptible populations with prepared antigens or antibodies through appropriate routes with the aim to elevate the populations' immunity and thus prevents the epidemic of communicable diseases. Vaccination has been a basic public health care service provided by governments and a public health care work of high sociality.

Immunization includes artificial active immunization and artificial passive immunization. In artificial active immunization, biological agents containing immunogenic antigens or toxoid are administered to target populations in order to induce specific immunity against infectious diseases. Vaccines are the biological agents used in preventive vaccination, which are prepared using detoxicated but antigenic microorganisms or their metabolites. Currently, vaccines can be divided into many categories including attenuated vaccine, inactivated vaccine, toxoid vaccine, subunit vaccine, synthetic peptide vaccine, conjugative vaccine, gene engineering vaccine and so on. Among them, gene engineering vaccines have drawn more attention of researchers, in part due to their high safety, low cost of production and simple preparation procedures. However, problems on relatively low effectiveness of gene engineering vaccines have to be addressed for certain communicable diseases.

In artificial passive immunization, immune sera or immune globulin are injected into human bodies to make the immunized objectives acquire specific immunity against a communicable disease. By this way, the immunity can be obtained rapidly, but only last for a short term. Therefore, passive immunization is mainly applied in emergency prevention and therapeutic treatments.

To take the advantages of both active and passive immunizations, the two modes of artificial immunization can be used in combination. For instance, both vaccines and antibodies are employed in immunization for the prevention of hepatitis B infection in newborn infants and diphtheria infection in susceptible contactees.

### ***11.4.2 Immunization Program***

Immunization programs involve formulation of the programs, plans and strategies for preventive immunization of susceptible masses according to the national programs for prevention and control of communicable diseases, and performance of planned vaccination using nationally or provincially determined species of vaccines and immunization programs, for prevention, control and eradication of infectious diseases via enhancing population immunity.

In 1974, based on the experiences from eradication of smallpox and control of measles, WHO proposed expanded program on immunization (EPI), in which it was required that all the member countries should insist on applying immunization methods and epidemiological surveillance in combination. Preventing epidemics of diphtheria, bronchocephalitis, tetanus, measles, tuberculosis, poliomyelitis etc., with emphasis on heightening vaccination coverage rates and expanding species of vaccines inoculated has been the mainstay of this approach. China joined EPI in 1981 and presently is carrying out a basic immunization program in children via using 13 vaccines for prevention of 15 communicable diseases. In 1988, the National Coalition for Adult Immunization (NCAI), a public health organization constituted by 40 countries, put forward an adult immunization proposal on programmed immunization with vaccines against epidemic influenza, pneumonitis, hepatitis B, measles, tetanus and diphtheria in adults and the populations at high risks. At present, the adult immunization program has been implemented gradually in some developed countries and some districts in China.

### ***11.4.3 Evaluation of Immune Effectiveness***

Evaluations of immune effectiveness include appraising immunological effects, epidemiological effects and immunization program management.

Immunological effects of immunization programs can be evaluated using positive conversion rate of specific antibodies, averaged titer levels of antibodies and lasting time of antibody positive status in immunized populations.

$$\text{Positive conversion rate} = \frac{\text{No. of positive conversion cases}}{\text{No. of immunized individuals}} \times 100\%$$

Immunological effects of immunization programs can be evaluated through the calculation of immune protection rates and effect index based on the data from field experiments using double-blind methods.

$$\text{Protection rate} = \frac{\text{IR of control group} - \text{IR of vaccinated group}}{\text{IR of control group}} \times 100\%$$

$$\text{Effect index} = \frac{\text{IR of control group}}{\text{IR of vaccinated group}}$$

Note: IR, incidence rate.

Immunization program management can be appraised referring to the indexes, such as qualified inoculation rate of vaccines, coverage rate of programmed immunization and record rate of vaccination.

## 11.5 Emerging Communicable Diseases

### 11.5.1 Definition

Emerging communicable diseases (ECD) refer to newly identified and previously unknown infectious diseases which cause public health problems either locally or internationally. Accompanying the decrease of traditional communicable diseases, ECD is gradually becoming the main causes of public health problems.

### 11.5.2 Main Emerging Communicable Diseases

Since 1970, more than 40 species of ECD have been identified as shown in Table 11.1, most of them are caused by viruses, and many of them do serious damage to human health, and become new research foci in medical science.

## 11.6 Summary

Communicable diseases are caused by a variety of microorganisms including bacteria, viruses, protozoan and fungus. The associations between human and the pathogens occur in two processes, that is, infection process and epidemic process, the former is ongoing in individuals, while the latter is among individuals in specific populations. A variety of outcomes resulting from infection with a pathogen in

**Table 11.1** Emerging communicable diseases identified since 1970

Year	Pathogen	Disease
1973	Rotavirus	Diarrhea of infants
1975	Hepatitis B virus	Hepatitis B
1976	Ebola virus	Ebola virus disease
1977	Hantaan virus	Hemorrhagic fever with renal syndrome
1977	Hepatitis D virus	Hepatitis D
1977	<i>Legionella pneumophila</i>	Legionellosis
1977	<i>Campylobacter jejuni</i>	Enteritis
1981	Toxin-producing strains of <i>Staphylococcus aureus</i>	Toxic shock syndrome
1982	<i>Escherichia coli</i> O157: H7	Hemorrhagic colitis
1982	<i>Borrelia burgdorferi</i>	Lyme disease
1983	<i>Helicobacter pylori</i>	Gastric ulcer and cancer
1983	Human immunodeficiency virus, HIV	Acquired immunodeficiency syndrome
1986	Human herpesvirus 6	Roseola infantum
1986	<i>Cyclospora cayatanensis</i>	Persistent diarrhea
1989	Hepatitis C virus	Hepatitis C
1989	Hepatitis E virus	Hepatitis E
1990	Human herpesvirus 7	Fever, erythra, CNS infection
1992	<i>Vibrio cholerae</i> O139	New cholera
1993	<i>Escherichia coli</i> O12:K1:H7	Urethra infection, abortion, sepsis, encephal meningitis
1995	Human herpesvirus 8	Kaposi's sarcoma
1995	Hepatitis G virus	Hepatitis G
1996	Prion	New Creutzfeldt-Jacob disease
1997	<i>Rickettsia mongolotimonae</i>	Tick-borne lymphadenopathy
1998	Nipah virus	Cephalomeningitis
2003	SARS-associated coronavirus	SARS
2006	Avian influenza A (H5N1) virus	Human avian influenza
2008	<i>Anaplasma phagocytophilum</i>	Human granulocytic anaplasmosis
2009	Influenza A (H1N1) virus	Influenza A
2013	Middle East respiratory syndrome coronavirus	Middle East respiratory syndrome
2016	Highly pathogenic avian influenza A (H5N8) virus	Human avian influenza
2019	SARS-CoV-2	COVID-19

populations is called spectrum of this infection. Communicable diseases can become prevalent based on coexistence and association of the three links, namely, sources of infection, routes of infection and susceptible populations. Epidemics of infectious diseases are also affected by multiple natural and social factors. Surveillance of public health forms the critical basis for prevention and control of communicable diseases. Effectiveness of infectious disease control is, to a large extent, dependent on reasonable strategies and practical measures. The measures employed should

vary from the epidemic characters of communicable diseases, with the aim to eradicate the sources of infection, blocking the transmission routes and protecting the susceptible masses. Programmed immunizations have been proved as effective and practical strategies and measures for prevention and control of communicable diseases, and have been developing from protecting children to now covering adults.

# Chapter 12

## Epidemiology of Noncommunicable Diseases



Jie Yang and Man Li

### Key Points

- The growing burden of non-communicable diseases (NCDs) is one of the major public health challenges facing all countries in the twenty-first century.
- The common risk factor for NCDs are tobacco use, alcohol use, unhealthy diet, physical inactivity, raised blood pressure, overweight/obesity, etc.
- Prevention of NCDs should integrate the strategies of individual-based high-risk population and population-based all-population.

## 12.1 Introduction

### 12.1.1 Definition

Noncommunicable diseases (NCDs) are the diseases characterized by multifactorial causation, long latent period, indefinite onset, and noncontagious among individuals. NCDs include broad types of diseases such as cardiovascular disease, renal disease, nervous and mental disease, musculoskeletal conditions, chronic non-specific respiratory disease, cancer, diabetes, and various other metabolic and degenerative diseases. In this chapter, we focus on cardiovascular disease, cancer, and diabetes. The risk factors of NCDs generally include tobacco use, alcohol use, physical inactivity, unhealthy diet, and some unhealthy conditions such as overweight/obesity, high systolic blood pressure, high fasting plasma glucose, and high cholesterol levels.

---

J. Yang (✉) · M. Li (✉)  
School of Public Health, Hebei Medical University, Shijiazhuang, China

## ***12.1.2 The Influence of NCDs on Health and Society***

At the broadest level of the global burden of disease (GBD) hierarchy, NCDs contributed 73.4% or 41.1 million deaths. At Level 2, the largest numbers of deaths from NCDs were 17.8 million deaths for cardiovascular diseases, 9.56 million deaths for neoplasms, and 3.91 million deaths for chronic respiratory diseases. Total disability-adjusted life years (DALY) from NCDs increased by 36.6% from 1.07 billion in 1990 to 1.47 billion in 2016. In China, NCDs are also leading cause of deaths and account for 80% of all deaths. For Chinese population in 2017, stroke, ischemic heart disease, chronic obstructive pulmonary disease (COPD), and lung cancer were the leading four causes of all-age DALYs in 2017.

Treatment, rehabilitation, and taking care of disability from chronic diseases have formed tremendous pressure on the individuals, family, society, and health system. The economic burden of NCDs is huge, involving in not only the huge extra health care spending of personal, family, and society due to chronic disease, but also the loss of productivity due to illness, disability, and premature death.

## **12.2 Epidemiological Features**

### ***12.2.1 Overall Global NCDs Outlook***

For the time distribution of NCDs, deaths from NCDs increased by 22.7% from 33.5 million in 2007 to 41.1 million in 2017 globally, while the death rate decreased by 7.9% from 582.1 deaths per 100,000 in 2007 to 536.1 deaths per 100,000 in 2017. Declines in cardiovascular disease and neoplasms are slowing in many high-income countries. For the place distribution of NCDs, there are significant difference in incidence, mortality from NCDs between low-, middle-, and high-income countries. In low- and middle-income countries, communicable disease and maternal and infant mortality remain at a high level, but deaths from chronic disease are lower than in high-income countries. Due to the large population in all low- and middle-income countries, the absolute numbers of chronic disease patients are still higher than that of those in high-income countries. About three-quarters of chronic disease death, three-quarters of death from cardiovascular disease and diabetes, 90% of chronic respiratory disease deaths, and two-thirds of cancer death overall globally occur in low- and middle-income countries. The age-standardized mortality is not affected by population size and age composition of the population. In 2012, the age-standardized mortality of chronic disease in low-income countries (625/100 thousand) and middle-income countries (673/100 thousand) was higher than that in high-income countries (397/100 thousand).

The trend of incidence and mortality of NCDs between low-, middle- and high-income countries is different. In some high-income countries, mortality of cardiovascular disease and some cancers (e.g., lung cancer) shows a declining trend. For



instance, since the 1950s, the age-standardized mortality of heart disease, stroke, and cancer declined 70%, 78%, and 17%, respectively, in the United States. From 1980 to 2010, compared to the significant reduction of age-standardized mortality of ischemic heart disease in high-income countries, Eastern Europe, Central Asia, South Asia, and East Asia have shown significant upward trends. South Asia has a larger population, and the average death age of ischemic heart disease is younger, and the years of life lost (YLLs) due to premature death is greatest. Since early 1990s, ischemic heart disease became popular in Eastern Europe and Central Asia, the crude mortality and age-standardized mortality in these regions are the highest overall world. In North Africa, the Middle East, and Southeast Asia, the average death age of ischemic heart disease is younger, and the age-standardized mortality is higher, which indicates the death of ischemic heart disease is more likely occur in the labor population.

The early onset of illness is becoming common. Irrespective of gender, persons from all age groups will be affected by chronic diseases. Chronic disease is common in older persons. However, the data from 2012 showed that 42% of NCDs death occurs in the person less than 70 years old (is regarded as premature deaths). The proportion of premature deaths in low- and middle-income countries (48%) is higher than that in high-income countries (28%).

For cardiovascular disease (CVD), most countries experienced four stages: low stage, rising stage, peak stage, and decline stage. Before the 1950s, the social economy, living, and medical conditions were at lower status, infectious disease were the major threat to human health. The incidence of cardiovascular disease was relatively low, and the number of deaths accounted for only from 5% to 10% of total deaths. Industrialization improved social economy and living condition, which result in increased nutrition, diet high in sodium, and insufficient physical activity. The incidence of CVD in the population was rising, and the death due to CVD accounted for 10–30% of total death (Stage II). High-fat, high-protein, high-calorie diets and inactive physical activity led to a rapid increase in CVD, especially coronary heart disease (CHD) and ischemic stroke. The incidence and death appeared a younger trend, and the death accounted for 35–65% of total death (Stage III). Due to public health measures such as health education and community intervention and progress in medicine, the incidence and mortality of CVD declined gradually and the composition of death reduced to less than 40% (Stage IV). Most countries and regions followed the above four stages in CVD epidemic, but different countries enter different development period, and the current stage is different. For example, Western Europe, North America, Australia, Japan, and Korea due to high degree of industrialization, CVD currently entered Stage IV. While Eastern Europe, Russia, the Middle East, and some fast-growing countries, the mortality of CVD has increased 50–100% which accounts for 40–60% of total death within recent 30 years. In Asia, Latin America, and Africa, CVD begins to enter Stage II and the composition of death accounts for under 30% of total death.

For cancer, from a global perspective, the incidence and mortality of cancer are increasing gradually, except for cervical cancer, esophageal cancer, and stomach cancer. The epidemiological feature of cancer among countries and regions is

different. Lung cancer has a higher age-standardized incidence in North America, the middle region of Western Europe, South Europe, North Europe, and East Asia, but lowest in the Middle and West Africa. Breast cancer in developed countries (except Japan) has higher incidence with 89.7 per 100,000 in Western Europe, but lower incidence in most underdeveloped countries with less than 40 per 100,000. Generally speaking, high incidence of cancer in developed countries is lung cancer, breast cancer, colon cancer, and prostate cancer.

For diabetes mellitus, the incidence between types, countries, and races are different. For type 1 diabetes mellitus, the difference in age-adjusted incidence is 350 times at the global level. Sardinia of Italy (36.8 per 100,000) and Finland (36.5 per 100,000) own the highest incidence, other European and American countries have middle incidence (from 5.0 per 100,000 to 19 per 100,000), and the lowest incidence (from 0.1 per 100,000 to 5.0 per 100,000) happens in some Asia countries (such as China, Japan, and Korea), American Indians, Mexicans, Chileans, and Peruvians. Type 1 diabetes show the increased incidence with far away from the equator. For type 2 diabetes mellitus (T2DM), the incidence is related to lifestyle. Keeping traditional way of life shows lower incidence, while some westernized developing countries show higher incidence. The prevalence of type 2 diabetes in rural Africa in adults is from 1% to 2%; however, in North America and Western Pacific Region, about from 1/3 to 1/2 of adults have been diagnosed as type 2 diabetes.

### ***12.2.2 Epidemiological Features of the Risk Factors of NCDs***

It is accepted that a set of “risk factors” are responsible for morbidity and premature mortality of NCDs. A large percentage of NCDs are preventable through the changes in these factors, which include tobacco use, physical inactivity, alcohol use, unhealthy diet, raised blood pressure, overweight/obesity, high cholesterol, cancer-associated infections, and environmental risk factors.

#### **12.2.2.1 Tobacco Use**

The number of men smokers has steadily increased in the first half of the twentieth century and up to a peak of 80% within several decades after World War II. The rate of smoking among men began to decline in English-speaking countries and North European countries, but female smokers began to ascend in these countries at first and after the last half of the twentieth century, then spread to Japan, Latin America, central Europe, and south Europe.

In 2012, the smoking prevalence was 22% and had obvious regional difference. The smoking prevalence is highest in European countries (30%) and is lowest in Africa (12%). The smoking prevalence among male (37%) is higher than that among female (7%). In 2010, the incidence of smoking among adult was 28.1% (male

52.9% and female 2.4%). Among men, the highest smoking age is 45–64 years old (63.0%) and the lowest is 14–24 years old. Smokers in rural areas accounted for 56.1% higher than those in urban region (49.2%). Prevalence of second-hand smoking is up to 72.4%.

#### **12.2.2.2 Alcohol Use**

Alcohol use is undoubtedly a risk factor for NVDs. It is reported that the harmful use of alcohol is one of the four behavioral risk factors (tobacco use, unhealthy diet, physical inactivity, and alcohol use) for three major NCDs (cardiovascular disease, cancer, and chronic respiratory disease). About 2.3 million people die from the harmful use of alcohol each year, contributing about 3.8% of the world's total deaths. The attributable DALYs is high for alcohol (85.0 million DALYs). Adult alcohol consumption is highest in Europe and America and lowest in Mediterranean countries and Southeast Asian countries. The heavy episodic drinking within the past 30 days is highest in Europe and America.

#### **12.2.2.3 Unhealthy Diet**

Unhealthy diet is a key modifiable risk factor for NCDs, which include inadequate consumption of fruits and vegetables, excessive consumption of sugar-sweetened beverages (SSBs), high sodium intake, and high consumption of saturated fats and trans-fatty acids. Evidence showed that inadequate consumption of fruits and vegetables increases the risk of CVD, stomach cancer, and colorectal cancer. High consumption of SSBs was associated with excess energy intake and was strongly linked to obesity. People with much higher levels of sodium input than recommended by WHO are at higher risk for high blood pressure and cardiovascular disease. High intake of saturated fats and trans-fatty acids has been linked to heart disease. Unhealthy diet is increasing rapidly in low-resource areas.

#### **12.2.2.4 Physical Inactivity**

Physical inactivity means the inability to achieve the recommended levels (at least 30 min of regular, moderate-intensity physical activity on most days) of physical activity for health. It is the fourth leading cause of death worldwide and is the major risk factor for NCDs. About 9% of all deaths globally are attributed to physical inactivity. People who lack physical activity have a 20–30% increased risk of all-cause death. Physical inactivity is most severe in high-income countries, but it is significant in some middle-income countries, particularly among women.

### 12.2.2.5 Raised Blood Pressure

It is estimated that raised blood pressure cause 7.5 million deaths, about 12.8% of all deaths. The percentage of populations with raised blood pressure was higher in regions with lower income level. In low-income and middle-income countries, such as eastern, western, middle, and southern Africa and Mongolia in Asia, about 30% of the population had raised blood pressure, but other countries had a lower population of raised blood pressure.

### 12.2.2.6 Overweight/Obesity

Since 1980, prevalence of overweight/obesity has a steady rise with the fastest speed in the United States, with the second fastest is in China, Brazil, and Mexico. In 2014, the prevalence of overweight (BMI (body mass index)  $\geq 25$  kg/m<sup>2</sup>) of adult was 38% and 40% in male and female, and the prevalence of obesity (BMI  $\geq 30$  kg/m<sup>2</sup>) accounts for 11% and 15%. The highest prevalence of overweight/obesity is in the American countries (overweight 61% and obesity 27%), and the lowest is in Southeast Asia (22% and 5%).

## 12.3 Risk Factors of Several Common NCDs

### 12.3.1 *Cardiovascular and Cerebrovascular Disease*

#### 12.3.1.1 Stroke

Stroke, also known as cerebrovascular accident, is an acute cerebrovascular disease, which including ischemic stroke and hemorrhagic stroke. Damage to brain tissue occurs when a blood vessel in the brain suddenly bursts or becomes blocked, then preventing blood from flowing into the brain. Stroke does not occur by chance, and there are factors that occur several years before stroke. The risk factors are as follows:

1. Hypertension

Hypertension is the main risk factor for cerebral thrombosis as well as cerebral hemorrhage. Data from prospective studies showed that the risk of stroke is increased by 49% with each increase of 10 mmHg in systolic pressure, and is increased by 46% with each increase of 5 mmHg of diastolic pressure. The geographic distribution of stroke is consistent with that of hypertension in morbidity and mortality.

2. Heart disease

Heart damage is the second highest risk factor for stroke. In Framingham heart study, the majority of stroke patients had coronary heart disease, congestive heart failure, and atrial fibrillation.

### 3. Diabetes

Diabetes is also an independent risk factor for stroke. The incidence of stroke in individuals with diabetes is 2.5–3.5 times higher than those without diabetes. Men with type 2 diabetes are three times of risk in having a stroke than nondiabetic patients, but women with type 2 diabetes have five times risk than nondiabetic patients.

### 4. Dyslipidemia

The incidence of stroke is increased by 25% for every 1 mmol/L increase in serum total cholesterol. The incidence of ischemic stroke is reduced by 47% for every 1 mmol/L increase in high-density lipoprotein (HDL).

### 5. Other factors

Additional factors include obesity, smoking, glucose intolerance, blood clotting and viscosity, and oral contraceptives.

## 12.3.1.2 Coronary Heart Disease

### 1. Hypertension

The prevention of hypertension and the improvement of blood pressure are essential and fundamental steps for CHD prevention. The famous Framingham heart study showed that prehypertension and Stage 1, Stage 2, and higher hypertension increased the risk of CHD in both men and women. In the past, emphasis was placed on the importance of diastolic blood pressure (DBP). Many investigators feel that systolic blood pressure (SBP) is a better predictor of CHD than diastolic pressure. However, both components are significant risk factors. A meta-analysis showed that the slope of the association between CHD mortality and normal SBP levels was almost constant across each age range throughout the normal SBP values drop to lower than 115 mmHg. Furthermore, for the age-specific hazard ratio between CHD mortality and DBP values drop to lower than 75 mmHg, it was equivalent to that associated with a 20 mmHg difference in normal SBP values.

### 2. Dyslipidemia

It is well known that hyperlipidemia with elevated serum total cholesterol, low-density lipoprotein (LDL) cholesterol, and triglycerides is a major risk factor for CHD. The increased serum total cholesterol and low-density lipoprotein and declined high-density lipoprotein are associated with increased risk of CHD. The risk of CHD is decreased by 2% for every 1% decrease in serum TG. The reduction of every 0.03 mmol/L in HDL-C will increase the risk for CHD by 2–3%. The prevalence of dyslipidemia was 75–85% in patients with early onset CHD, compared with 40–48% in age-matched controls without CHD. Clinical studies have shown that lowering total or LDL cholesterol can play a better role in primary or secondary prevention.

### 3. Diabetes

CHD is common in patients with diabetes, and its prevalence increases with worsening glycemic status due to a higher risk of accelerated atherosclerosis and other lipotoxic and glycotoxic effects. The risk of CHD in people with diabetes is 2–3 times higher than in those without. In industrialized countries, 30–50% of diabetic among people over the age of 40 years die from CHD.

### 4. Overweight/obesity

The overall obesity rate among adults was 12.0% in 2015, with higher rates in women across all age groups. Globally, elevated BMI causes more than four million deaths and 120 million DALYs each year, most of which are directly attributable to the subsequent development of cardiovascular disease. In fact, between 1980 and 2000, higher BMI resulted in approximately 25,905 additional deaths due to CHD. Compared with normal weight, the relative risk for those overweight/obesity developing to CHD and death is 1.5–2.0.

### 5. Tobacco use

Tobacco use has been identified as a major CHD risk factor. Nearly, six million people die each year from tobacco use, including direct smoking and second-hand smoking. Data from several studies suggest that the relative risk of CHD is 2–3 times higher among smokers. These risks have age gradient, with higher relative risk in the younger age group (5–6 times).

### 6. Physical inactivity

A sedentary lifestyle is associated with the risk of early CHD development. There is evidence that regular physical activity reduces body weight and blood pressure and increases the HDL level, which are beneficial for cardiovascular health. In a meta-analysis of 43 prospective cohort studies, compared with 600–3999 MET-min/week of total physical activity across all domains, the RR among people <600 MET-min/week increased by 19%. A total of 5.0% CHD deaths can be attributed to physical inactivity.

## 12.3.2 T2DM

### 12.3.2.1 Genetic Factor

T2DM has strong family aggregation. The prevalence of diabetic relatives is 4–8 times higher than that of nondiabetic relatives. Twin studies have shown a consistency of about 90% in identical twins with T2DM, thus demonstrating a strong genetic component. The heritability of T2DM in China is 51.2–73.8%. In addition, a person's risk of developing diabetes may also depend on the genetic susceptibility to diabetes. Many genome-wide association studies have investigated genetic variants in different populations that influence disease susceptibility through rare alleles and common variants. For example, Chauhan et al. showed the association of eight gene variants (PPAR $\gamma$ , KCNJ11, TCF7L2, SLC30A8, HHEX, CDKN2A, IGF2BP2, and CDKAL1) with diabetes in Asian Indians.

### 12.3.2.2 Overweight/Obesity

Overweight/obesity, particularly central adiposity, has long been accepted as a risk factor for prediabetes or type 2 diabetes (T2DM). Overweight/obesity leads to inflammation, endoplasmic reticulum stress, and fat factor, all of which occur in the pathogenesis of insulin resistance of the liver and skeletal muscle in the work, and increase the risk of diabetes. In a meta-analysis of 84 articles involving more than 2.69 million participants, the combined prediabetes risk of overweight/obesity versus normal weight was 1.24. Based on the race-specific BMI classification, the combined risk for type 2 diabetes relative to normal weight was 0.93 for underweight, 2.24 for overweight, 4.56 for obese, and 22.97 for severely obese. The RR of T2DM in overweight/obesity decreased with age. Another meta-analysis showed that obesity in children and adolescents was positively associated with the prevalence of T2DM and prediabetes, in which obese subjects 13 times higher than normal-weight subjects. The prevalence of prediabetes was three times higher in obese subjects (17.0% vs 6.0%, respectively).

### 12.3.2.3 Physical Inactivity

Sedentary lifestyle is an important risk factor for the development of T2DM. Lack of exercise may alter the interaction between insulin and its receptors, which can lead to T2DM. Those who watched TV for 4 h a day had a 46% higher risk of developing diabetes less than 1 h a day. Meta-analysis of 55 prospective cohort studies <600 MET-min/week versus 600–3999 MET-min/week of total physical activity across all domains: 1.17 (1.11–1.23) 4.5 (3.1–6.0) type 2 diabetes 2.7 (1.9–3.5) 4.2 (2.9–5.7) 5.9 (4.2–7.7).

### 12.3.2.4 Unhealthy Diet

Both food calories and the quality of diet components affect the risk of diabetes. Excessive calorie intake can increase the overweight, with the passage of time, the metabolism of liver glucose control and steady state would be destroyed. Poor dietary quality, such as low intakes of dietary fiber, low-sugar carbohydrates, or whole grain grains, increases the risk of diabetes, as does high intakes of saturated fatty acids and trans fats. A diet containing high-quality fats and carbohydrates rather than low-quality fats and carbohydrates is more important than the relative amounts of these nutrients in preventing type 2 diabetes.

### **12.3.2.5 Malnutrition**

Malnutrition in early infancy and childhood or undernutrition early in life (e.g., exposure to famine) can lead to partial beta cell failure, impaired carbohydrate tolerance, and increases the risk of developing type 2 diabetes later in life.

### **12.3.2.6 Impaired Glucose Tolerance (IGT)**

IGT is an intermediate state between normal and diabetes. The IGT patients have a higher prevalence of diabetes. When IGT patient was followed up to 5–10 years after the first diagnosis, about one-third of the individuals have developed into diabetes, one-third were converted to normal blood glucose and one-third remained IGT status. IGT is easily converted to diabetes when accompanied by the following factors: fasting blood glucose, 2-h blood glucose, and BMI more than 5.0 mmol/L, 9.4 mmol/L, and 25, respectively. Improved diet and increased physical activity are beneficial in reducing the chance of IGT conversion to diabetes.

### **12.3.2.7 Insulin Resistance**

Clinical studies found that insulin resistance occurred in obesity, type 2 diabetes, hyperlipidemia, hypertension, coronary heart disease, stroke, etc. Blood insulin plays a vital role in the process of diabetes development from normal or IGT. Insulin resistance is a common pathophysiological mechanism in above pathological processes.

### **12.3.2.8 Maternal Diabetes**

Offspring of diabetic pregnancies, including gestational diabetes, tend to be large and heavy at birth, tend to develop obesity in childhood, and have a high risk of developing T2DM in early life. The risk of diabetes was three times higher in children born to mothers with diabetes than in children born before their mothers. Maternal diabetes, which is associated with intrauterine growth retardation and low birth weight, appears to increase the risk of later diabetes in children if it is associated with subsequent rapid growth catch-up.



### **12.3.3 Cancer**

#### **12.3.3.1 Physical Factors**

Ionizing radiation (X,  $\gamma$ ,  $\alpha$ ,  $\beta$ -ray, etc.) can cause a variety of human cancer including lung, breast cancer, leukemia, multiple myeloma, thyroid cancer, skin cancer, etc. In occupational factors, other physical factors such as asbestos fiber, coal dust, and quartz dust can result in lung cancer and mesothelioma.

#### **12.3.3.2 Tobacco Use**

Many large prospective studies have provided that tobacco use increases the risk of cancer mortality, especially lung cancer. Tobacco use also leads to larynx, oral, head and neck, pharynx, esophagus, bladder, pancreas, cervical, breast, and probably kidney cancer. A prospective studies reported that the age-adjusted incidence rate of cancer is highest in current smoker, with the RR is up to 12.0 for lung cancer. The attributable risk for oral-bladder cancers, other cancers, and all cancers were 46%, 16%, and 29%, respectively. Another systematic review showed that men who are current smokers have a moderately increased risk of total cancer compared to those never smoked. In women, the risk is increased but less than in men. The overall relative risk was estimated at 1.53.

#### **12.3.3.3 Alcohol Use**

Excessive intake of alcoholic beverages is associated with oral, pharyngeal, esophageal, liver, colon, rectal, and breast cancer. With the exception of the American Cancer Society and the Canadian Cancer Society, all organizations state that alcohol is a class carcinogen and that even small amounts of alcohol can increase the risk of certain cancers. However, some studies showed that light or very light alcohol use was not associated with the risk of common tumors, except for mild increases in breast cancer in women and colorectal cancer in men.

#### **12.3.3.4 Dietary Factors**

A lot of studies suggested the potential role of diet in certain cancers. However, there is no guarantee of cancer prevention. The study of diet and cancer risk reduction is complicated not only by the multistage, multifactor nature of the disease, but also by the inherent complexity of any diet. Prospective cohort findings support an association between unhealthy eating patterns and increased risk of colon and breast cancer, particularly in postmenopausal hormone-receptor negative women. The limited evidence of an association between unhealthy dietary patterns and the risk

of upper digestive tract, pancreatic, ovarian, endometrial, and prostate cancers relies only on case-control studies.

### **12.3.3.5 Occupational Exposures**

Occupational exposure includes exposure to benzene, arsenic, cadmium, chromium, vinyl chloride, and asbestos polycyclic hydrocarbons. The risk of occupational exposure is greatly increased if individuals also smoke. It has been reported that occupational exposures typically account for 1–5% of all human cancers.

### **12.3.3.6 Biological Factors**

Biological factors are one of the main causes of human tumor. By now, it is identified that at least eight viruses have been linked to human tumors. For example, the increased risk of hepatitis B, C virus for hepatocellular, human immunodeficiency virus for Kaposi's sarcoma, the Epstein–Barr virus for Burkitt's lymphoma and nasopharyngeal carcinoma, and human papilloma virus for cervical cancer.

### **12.3.3.7 Genetic Factors**

It is now becoming clear that individual differences in the incidence of tumors are related to genetic background, which means that the occurrence of tumors is also related to the individual's own genetic susceptibility. Although there is probably a complex interrelationship between hereditary susceptibility and environmental carcinogenic stimuli in the causation of a number of cancers. With the completion of the Human Genome Project and the rapid development of high-throughput gene variation detection methods, genome-wide association studies (GWAS) have become a major strategy for revealing tumor susceptibility genes. In recent years, tumor researchers around the world have used GWAS strategy to conduct a series of studies on nasopharyngeal carcinoma, liver cancer, esophageal cancer, lung cancer, pancreatic cancer and other tumors in people around the world, and a large number of genetic variations and genetic loci of tumor-related chromosome regions have been discovered, which is of great significance for fully revealing the causes of tumor occurrence.

### **12.3.3.8 Other Factors**

Other factors include the immune, endocrine, and psychosocial factors. The immune system is closely related to the incidence of cancer. Tumor cells can evade immune system attacks by one or more mechanism or cannot activate specific antitumor immunity and induce the tumor development. Endocrine-related tumors include

breast, ovarian, and testicular cancer. The risk factors for breast cancer include non-procreation, early onset, late menopause, and non-lactation. Social psychological factors are also one of the important risk factors for cancer. Major adverse events and depression can cause psychological stress, which lead to the disturbance of the nervous system and the decline of immunity.

## **12.4 Prevention and Control of NCDs**

### ***12.4.1 Prevention Strategy***

The prevention and control of NCDs emphasize the primordial prevention, which controls the risk factors at the population level. Some of the risks of adult chronic disease begins with adverse exposure in pregnancy. Many unhealthy lifestyles are formed from childhood. Once formed, it is difficult to change. Therefore, the prevention of NCDs takes the life-course approach which is from early life through the whole life period.

Prevention of NCDs should integrate the strategies of individual-based high-risk population and population-based all-population. When the risk factors exist among the whole population, the all-population strategy is particularly important. Smoking ban in public places and workplaces is a successful strategy for the all-population strategy.

Member States in the WHO Western Pacific Region endorsed the “For the Future” Vision at the Regional Committee Meeting in 2019. The paper set out four thematic priorities for making the Western Pacific the healthiest and safest region in the world, one of which is NCDs and aging. The burden of NCDs and related risk factors is a major barrier to the development and achievement of the sustainable development goals (SDGs), the WHO’s Global Action Plan for the Prevention and Control of NCDs 2013–2020 and the “For the Future” Vision. Based on the Global Action Plan for the Prevention and Control of NCDs 2013–2020, WHO provides a list of “Best Buys” and other recommended interventions in 2017 for the four key risk factors for NCDs (tobacco, harmful use of alcohol, unhealthy diet, and physical inactivity) and for four diseases (cardiovascular disease, diabetes, cancer, and chronic respiratory disease) to tackle global NCDs problems.

In order to maintain people’s health and build a healthy China in a well-rounded way, China has recently issued a number of policies, plans, and national actions for the prevention and control of NCDs, which include the National Basic Public Health Service Projects (NBPHSP) to manage hypertension and diabetes in primary health facilities since 2009, community-based and comprehensive intervention projects for NCD, the China Healthy Lifestyle for All (Phases I and II) launched in 2007, the Medium- and Long-Term Plan for the Prevention and Treatment of NCDs (2017–2025), the Outline of the Plan for “Healthy China 2030,” and the Healthy China Action (2019–2030) promulgated by the Chinese State Council. Early in March 2021, China issued the 14th Five-Year Plan (FYP), a high-level development

blueprint for the next 5 years. The 14th FYP calls for “fully implementing the Healthy China Actions,” “strengthening prevention, early screening and comprehensive intervention of chronic diseases,” etc.

### ***12.4.2 Prevention Measures***

There are a variety of measures for the prevention of NCDs. Urgent action is needed to reduce the growing burden of NCDs and prevent the annual toll burden that dying prematurely before the age of 70 from heart and lung disease, stroke, cancer, and diabetes. There is not only a growing awareness and concern about the burden of NCDs on families, individuals, and public health, but also the social and economic burdens associated with the NCDs. Many interventions for prevention and control of NCDs exist. Even in the richest countries, it is essential to choose which interventions to prioritize because resources are limited, and this is especially true in most countries. In 2017, WHO launched Best Buys, which recommended three types of intervention measures, including “the most cost-effective measures,” “effective interventions with cost-effectiveness ratio of more than \$100,” and “other interventions (without cost-effectiveness analysis),” for four chronic behavioral risk factors, tobacco use, unhealthy diet, insufficient physical activity, and harmful use of alcohol.

For tobacco use, the interventions including increased tobacco excise taxes to reduce the affordability of tobacco products, implement plain/standardized packaging and/or large graphic health warnings on all tobacco packages, eliminate second-hand smoke exposure in all indoor workplaces, public places, and public transport, ban cross-border advertising, and offer smoking cessation services to all who want to quit through mobile phone apps. For unhealthy diet, the interventions include reducing salt input, limiting food package sizes and portion sizes, reducing sugar consumption through effective taxes on sugary beverages, and promoting unsaturated fats instead of trans and saturated fats through formulation, labeling, fiscal or agricultural policies, etc.

For physical inactivity, the interventions include reducing physical inactivity, promoting physical activity and national fitness, promoting travel and domestic physical activity, reducing static behavior, providing physical activity counselling and referrals within routine primary health care services using short-term interventions, etc.

For alcohol use, the interventions including increased excise taxes on alcoholic beverages, complete bans or restrictions on alcohol advertising, restrict alcohol use and promote health education, prevention, treatment, and care of alcohol use disorders and their comorbidities, etc.

# Chapter 13

## Epidemiology of Public Health Emergencies



Hong Zhu

### Key Points

- Public health emergencies have enormous impact on population health.
- Public health emergencies can be divided into natural disasters, man-made disasters and disease outbreaks.
- Epidemiology plays a crucial role in the management of public health emergencies.
- Epidemiologic investigation is usually the first step when disease outbreak and disaster occurs.
- Response to public health emergencies is inextricably linked to advance preparedness.

The health impacts of public health threats, such as emerging infectious diseases (e.g., 2003 SARS epidemic, 2009 influenza A pandemic, and 2016 Zika outbreak), terrorism (e.g., 2001 World Trade Center bombing), environmental catastrophes (e.g., 2011 Fukushima Daiichi nuclear disaster), and natural disasters (e.g., earthquake), have demonstrated the importance of strengthening the public health systems and improving the community's ability to respond effectively. Epidemiology is critical for the management of public health emergencies. This chapter introduces the application of epidemiology in the investigation of, preparation for, and response to public health emergencies.

---

H. Zhu (✉)

School of Public Health, Tianjin Medical University, Tianjin, China

e-mail: [zhuhong@tmu.edu.cn](mailto:zhuhong@tmu.edu.cn)

© Zhengzhou University Press 2023

C. Wang, F. Liu (eds.), *Textbook of Clinical Epidemiology*,

[https://doi.org/10.1007/978-981-99-3622-9\\_13](https://doi.org/10.1007/978-981-99-3622-9_13)

## 13.1 Basic Conception of Public Health Emergencies

Throughout 2018, altogether 484 public health events were recorded in WHO's event management system (a 16% increase from 2017), of which 352 (73%) were attributed to infectious diseases, 47 (10%) to disasters, and 19 (4%) to food safety.

### 13.1.1 Definition of Public Health Emergencies

A “public health emergency” refers to *a sudden-onset natural or man-made event, which poses a risk to public health*, like infectious disease outbreak, bioterrorist attack, severe food poisoning or industrial poisoning, or other significant or catastrophic events. A public health emergency, being an important part of all kinds of emergent events, seriously affects the health of a certain population and needs multi-sectoral cooperation co-assistance to cope.

Two terms “disaster” and “accident” are easily confused with “public health emergency.” The United Nations Disaster Relief Organization (UNDRO) defines disaster as *“a serious disruption of the functioning of a society, causing widespread human, material, or environmental losses which exceed the ability of the affected society to cope using its own resources.”* This definition shows that disaster has broader influences not only on health, but also on social stability and economic development. Many experts agree that the disasters which pose a threaten to human life can be called public health emergencies, while those posing danger to environment or material, rather than human health, are not public health emergencies, such as volcanic eruption in remote areas or the 1998 Asian financial crisis.

The other term “accident” means *an unpleasant event that happens unexpectedly and causes injury or damage*. Accidents can usually affect individuals or groups, while public health emergencies always affect groups. Besides, the accident cannot be anticipated or predicted. However, different from accidents, some types of public health emergencies can be predicted, for example, flood, and the activities on risk assessment, early warning, and preparation will mitigate the consequences of predictable hazards.

### 13.1.2 Characteristics of Public Health Emergencies

1. Public health emergency is a sudden-onset event. A public health emergency, as a kind of emergencies, has the characteristic of emergency, that is, it occurs suddenly and unpredictably, and its onset and development are hard to predict. However, now, development in science has made it possible for humans to predict more and more disasters, like early warning of floods, forest fires, and infectious diseases.

2. Public health emergency has a great impact on human health. The victims of public health emergencies are not limited to some certain individuals, but rather a large population or substantial proportion of people. The term also includes events that may pose a potential health threat, such as exposure to infectious agents, contaminated water, or food that may cause harm to humans.
3. Public health emergencies caused by different reasons have different characteristics, and correspondingly have different treatment and management strategies. For example, for the outbreak of an infectious disease, the main aim of the investigation is to identify the pathogen and transmission route; while for a natural disaster, the main aim of the investigation is to assess the situation rapidly.
4. Immediate action and unconventional measures should be taken to deal with public health emergencies. Multi-sectoral and multinational cooperation is usually needed. Health-care services, together with administrative institutes, media, military, traffic agencies, academic institutes, etc., may deal with emergencies more effectively and efficiently.

### ***13.1.3 Classification of Public Health Emergencies***

There are many different classifications of public health emergencies. The most commonly used method is based on the cause(s) of emergencies.

1. According to the nature of these events, public health emergencies can be divided into as follows:

*Biological emergencies:* These include communicable diseases, biological agent-related terrorisms, and vaccine inoculation-related events.

*Chemical emergencies:* These include leaks of hazardous chemicals, intentional or unintentional chemical food poisoning, and the use of chemical agents in terrorist incidents.

*Radiological emergencies:* These include the release of harmful radiation caused by the explosion of a nuclear weapon or improvised nuclear device.

*Weather and home emergencies:* These include the threats caused by abnormal meteorological conditions, like thunderstorms, flooding, tornado, and extremely hot or cold weather, and the threats from home, like kitchen fire, gas leak or explosion, and carbon monoxide poisoning.

2. According to the originating source of the disaster, public health emergencies can be divided into as follows:

*Natural disasters:* Natural disasters include weather phenomena (such as tropical storms, tsunamis, avalanches, extreme temperatures, winds/typhoons/hurricanes, and floods) and geologically related disasters (like earthquakes, landslides, and volcanic eruptions).

*Man-made disasters:* Man-made disasters include industrial accidents, traffic accidents, pollution incidents, terrorism, and armed wars. Natural disasters have long been considered the ones that cause the most deaths and economic losses, but man-made disasters are becoming more prominent.

*Epidemic diseases:* Disease epidemics or outbreaks would threaten the health of a certain population. These diseases are usually infectious or communicable, such as cholera, measles, hepatitis, influenza, malaria, SARS, H1N1, and HIV. They can spread among population through different routes of transmission, including air, food, water, direct or indirect contact, and insect or animal vectors. It is worth noting that the risk of disease outbreak usually increases following natural or man-made disaster (i.e., a flood), mainly due to poor sanitation and overly dense populations.

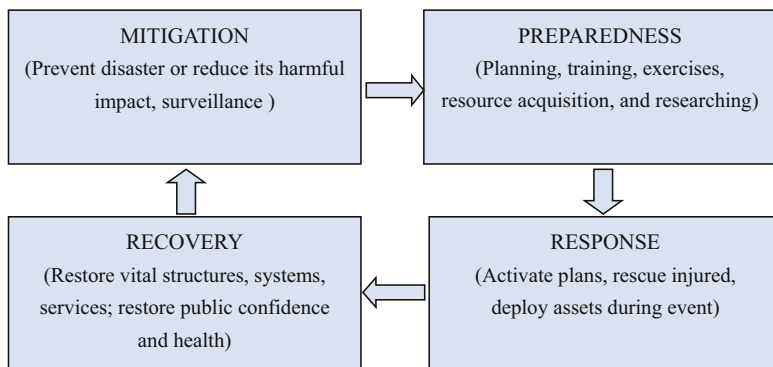
### ***13.1.4 Phases of Public Health Emergencies***

Emergency situations always change dynamically and require changes in response accordingly at different phases. The whole process of a public health emergency is often thought of as a cycle, which consists of six phases including preparedness phase (pre-emergency phase), warning phase, impact phase (emergency phase), response phase, reconstruction (rehabilitation) phase, and mitigating (preventing) phase. While the stages of dynamic change appear to be continuous, clearly identifying the end of each phase facilitates the adjustment of response strategies and better accommodation of new demands.

The whole process of an emergency may be divided into four phases: impact phase, response phase, recovery phase, and mitigation phase. Sometimes, it also includes a warning phase. Impact phase is the stage when the disaster causes real harm to the population. The duration of this phase depends on the number of people affected, and the type of incident. It may last for a few minutes (e.g., an earthquake), or several days (e.g., a flood), or several months (e.g., a disease outbreak). A response is made following the impact phase of an emergency. Preparedness actions taken in a timely manner prior to an emergency can be extremely helpful when facing an emergency. Relief activities (such as patients' treatment or victims rescue) occur during the response phase, which is followed by the reconstruction (rehabilitation) phase. During this phase, the focus is no longer on relief, but rather on development, which aims to help the affected people self-reliant. Besides, the lessons learned from the emergency are applied to prevent the recurrence of such disasters or to reduce the harm caused by such disasters, and to adequately prepare for such disasters should it recur.

The above process helps to formulate emergency response strategies and protocols. This concept suggests that four phases occur successively and unidirectionally. Actually, however, many things may occur at same time. The cycle concept ignores the reality that the serious consequence caused by emergencies may extend well beyond efforts at the reconstruction phase. The weakness of the efforts at reconstruction and mitigation phases can easily be magnified by subsequent emergencies (Fig. 13.1).





**Fig. 13.1** Phases cycle of public health emergencies

### 13.1.5 Harm Caused by Public Health Emergencies

Public health emergencies, especially those occurring in large scales not only result in human mortality, but also cause physical, psychological, and social disabilities.

1. *Physical harm*: Public health emergencies may cause deaths, injuries, and sometimes malnutrition. The impact phase is the primary phase of an emergency resulting in serious injury or death. When a public health emergency occurs, the most important and urgent task is the treatment of the wounded or patients.
2. *Psychological harm*: Psychological trauma is a unique individual mental experience caused by a serious injury event, especially when enormous pressure derived from the event is beyond one's ability to cope. Psychological effects of fear, helplessness, anxiety, depression, and terror caused by emergencies might linger for years, even decades, hence, psychological rehabilitation must be done on a long-term basis.
3. *Economic and social disabilities*: Natural disasters or terrorist attacks may lead to social disruption and infrastructure damage, such as transportation, residential building, enterprise assets, and electricity. It will directly or indirectly reduce the profits of the economy and reduce the speed of economic development, and ultimately affects social stability. Usually, developing countries are more vulnerable to emergent events than developed ones.
4. *Environmental and ecological harm*: Some kinds of emergencies, like volcanic eruptions, chemical or radiological releases, may affect the environment and disrupt the ecological balance in a certain geographic area, which in turn leads to harmful effect on human.

## 13.2 Basic Principles and Application of Epidemiology in Public Health Emergencies

### 13.2.1 Role of Epidemiology in Public Health Emergencies

Epidemiology of public health emergencies may be defined as *the application of epidemiology in public health emergency preparedness and response, with the aim of determining the nature of the events, exploring the causes and risk factors of the events, identifying the high-risk population, developing and assessing the health programs to prevent and control harmful effects caused by emergencies.*

From an epidemiological perspective, policymakers and practitioners can focus on the main issues of the entire affected population rather than on individuals, and to further develop measures to improve the health of the entire community. For public health personnel, epidemiological investigation is usually the first step when the disease or the disaster occurs.

Table 13.1 summarizes the epidemiological tools and principles applicable to responding to public health emergencies, with the aim of containing the progress of events, and reducing the harmful impacts on the whole populations.

### 13.2.2 Key Epidemiological Indicators

In epidemiology, the occurrence of disease or disaster is not in a random way; instead, follows a specific pattern which can be studied and predicted with respect to “what, who, where, when, how, why, whom, and what next.” Hence, some indicators can be used to describe this pattern, evaluate the impact of disease or disaster, and to assess the effectiveness and efficiencies of intervention programs. Table 13.2

**Table 13.1** The application and goals of epidemiology in public health emergencies

Application of epidemiology	Goals of epidemiology
Demand assessment	Evaluating the size and structure of the affect and potentially affected population, identifying the priority health issues and the health-related demands (medical devices, staffs, etc.) in the community
Population surveys	Determining the mortality and morbidity of disease (death rates and incidence/prevalence) or the number of injured; determining health status (nutrition and immunization status)
Public health surveillance	Monitoring health trends of the community; early predicting and warning
Disease outbreak investigation	Identifying the causes or risk factors of the disease; determining the source of infection, the route of transmission, and the susceptible population
Program evaluation	Assessing the coverage and the impact of health programs; economic evaluation

**Table 13.2** Epidemiological indicators for evaluating emergency intervention programs

Indicator	Examples
Public strategies	Degree of protocol commitment 1. Follow the diagnostic criteria for the case and treatment protocols 2. Degree of public and community involvement 3. Degree of cooperation and coordination between inter-sectors 4. Equity and accessibility of resource allocation
Demographics	Estimated number and structure of the affected population 1. Age and sex proportion 2. Population mobility and migration 3. Proportions and characteristics of high-risk and vulnerable groups 4. Ratio of urban and rural population
Health status	Rate of disease and death 1. Incidence and prevalence of common diseases or the infectious disease to be studied, secondary attack rate 2. Death rate (crude, age- or sex-specific, infant, under-fives, maternal) and fatality rate 3. nutritional status especially among under-fives
Program inputs	Input of the following resources: 1. Government financial guarantee funds 2. Facilities and equipment (health centers, beds, medicine, sanitize devices) 3. Staff (professional staff, volunteers, army, firefighters) 4. Basic supplies (food, water, shelter material, daily necessities) 5. Ancillary resources (fuel, charcoal, transport, communication)
Program process	Access, coverage, and quality of the following services: 1. Daily necessities sanitation 2. Environmental sanitation (feces/garbage disposal, disinfection) 3. Coverage of vaccination and prophylaxis 4. Demand for and utilization of health services (two-week attendance rate, hospitalization rate)

summarizes the epidemiological indicators that can be used to evaluate an emergency intervention program's process and outcome.

### 13.2.3 *Outbreak Investigation*

#### 13.2.3.1 Purpose of Outbreak Investigation

A disease outbreak (or epidemic) refers to the sudden occurrence of similar disease among many people in a region or a community due to the same source or carrier, with the incidence rate higher than normal expectations. Epidemiological investigations can be used to ascertain the nature of disease epidemics, for example, the cause of diseases (why), transmission routes and vectors (how), disease distribution by place, time and population (what, where, when, who), susceptible population (whom), and disease trend (what next). Determination of these natures of disease outbreak is critical to identify effective and proper clinical and public health

interventions. The objectives of outbreak investigation are to determine the existence and the severity of the outbreak, to explore the cause(s) and risk factors of the disease, to identify high-risk groups who are at higher risk of being affected and who would benefit most from interventions, and to identify and evaluate effective interventions to slow down the spread of the disease and reduce the harm caused by the outbreak.

### **13.2.3.2 Three Elemental Epidemiological Designs in an Outbreak Investigation**

In the epidemiological investigation of disease outbreaks, three elemental epidemiological designs can be used:

1. Descriptive epidemiology, especially cross-sectional design, can be used to describe the characteristics of disease distribution and to identify the difference in disease frequency among different regions and populations, which provide clues to the establishment of etiological hypotheses. Multiple cross-sectional studies can also provide information on temporal trends of disease outbreak.
2. Analytical epidemiology, mainly including case-control design and cohort design, can be used to explore the risk factor (e.g., environmental and behavioral factors) of disease by comparing the exposure proportion between cases and non-cases (case-control design) or by comparing the incidence rate between those with or without a certain exposure (cohort design).
3. Experimental epidemiology, including randomized controlled trial, field trial, and community intervention trial, can be used to assess the safety and efficacy of a certain intervention, such as a new medicine, a new vaccine, or a new public health program. For example, evaluating a cholera prevention program by comparing the incidence of cholera between two communities with or without initiating this program.

### **13.2.3.3 Key Steps in Carrying Out Outbreak Investigation**

1. Confirm the outbreak: All reports of the suspected outbreak by health-care workers or victims themselves require an immediate response from the relevant health authorities. The first step is to confirm the existence of the outbreak by local public health teams (e.g., staffs in CDC) through field investigation. Sometimes, besides epidemiologist, other specialists, such as microbiologists, zoologist, internists, and environmentalist, are needed to cooperate with the field investigation.

Initial investigations provide first-hand information on disease outbreaks, which is helpful for the follow-up investigation. Besides, the findings can also be used to develop diagnostic criteria for the disease, which is essential for the further investigation and clinical treatment.

2. Define a “case”: The main work at this step is to define a “case.” Health workers use this standard case definition to determine whether a person has a certain disease. Based on this case definition, investigators can identify how many people are cases (the numerator), and how many people are at risk of the occurrence of this disease (the denominator). Subsequently, attack rates (for disease outbreak) or incidence rate (for disease epidemic) can be calculated and compared with previous average rate level. Please note that data on the whole population is always needed as the denominators in the calculation of incidence rate or death rate.

The definition of the case includes suspected case, clinically diagnosed case, and a laboratory-confirmed case. In the early stage of the investigation, due to the lack of laboratory equipment for diagnosis, most cases are diagnosed based on epidemiological and clinical information. At this stage, more sensitive case definition (e.g., suspected case) is helpful to find more cases. In the middle stage of investigation, more specific definition (e.g., clinically diagnosed or laboratory-confirmed case) should be used in order to identify real patients and explore risk factors of the disease through case-control study or cohort study. In the end stage of the investigation, the definition used should be feasible and proper for disease surveillance, with the aim of assessing the effect of outbreak control activities.

3. Describe the outbreak by place, time, and person: Epidemic curve, which shows the number of diseases (y-axis) in an outbreak over time (x-axis), provides key information about an outbreak, including time trend (how quickly it is growing, and whether it is ongoing), what are the potential sources of disease (single or multiple sources), and how long the incubation period is. Dotted map which graphs the location of all reported cases can help to identify regional distribution of the outbreak and disease spread direction. Rates calculated by age and sex provide information on the most susceptible subpopulations and causal clue. Investigators can access to public health surveillance system to get data for rate calculation.
4. Analyze what caused the outbreak: Explore key differences between the non-cases and cases to determine which groups or individuals are more susceptible to disease, what are the potential risk factors, and what are the possible sources of disease and routes of transmission. Biological specimens are also collected from two groups for experimental test. If there has been some evidence that certain exposure is the potential cause of the disease outbreak, historical or prospective cohort design can be used to identify whether the persons with the exposure has the higher incidence rate than those without the exposure.
5. Assess environment: If disease outbreak is speculated to be related to environmental factors, such as animals or vectors, fecal contamination, or toxic chemicals, assess environment becomes necessary. Animal hosts or vectors should be investigated and some abnormal phenomenon should be observed and reported, especially for animal-borne diseases.
6. Initiate and improve prevention and control strategies: The ultimate goal of epidemiological investigation is to prevent and control disease outbreak. Hence,

intervention strategies and programs should be developed and launched based on the available information on disease outbreak. A clear understanding of the outbreak characteristics is needed for effective prevention and control. This involves three aspects: morbidity and mortality of the diseases, three elements of disease epidemic (source of infection, route of transmission, susceptible population), and context information (the amount and availability of medical resources, administrative divisions, etc.).

The main countermeasures against disease outbreak include controlling the source of infection (isolation and treatment of confirmed cases, management of asymptomatic carriers and animal hosts), curb disease spread (personal hygiene, environmental disinfection, health education, avoid gatherings, vector control, and entry-exit health quarantine) and protecting susceptible individuals (vaccination, nutrition, personal protection, and chemoprophylaxis). All these control measures should be established in accordance with national disease control regulations and policies.

Note: Taking control measures as early as possible is the most effective way to prevent the spread of disease. Hence, the process of outbreak investigation and intervention must be conducted rapidly and simultaneously. When little is known about the cause of disease, nonspecific measures for control of communicable diseases can be taken, such as isolation of the suspected patients, environmental disinfection, and personal protection. With the further understanding of the disease, prevention and control strategies will be improved correspondingly, and more targeted measures can be introduced (e.g., immunization).

7. Summarize an outbreak investigation and write a report:

Summarizing the whole process of the investigation and sharing the experience with all involved is the last step when the outbreak ends. The report should include the source of infection and possible routes of transmission, possible causes of the outbreak, case characteristics and clinical symptoms, geographical distribution, time trends, and lessons learned from epidemic control.

### ***13.2.4 Disaster Investigation***

#### **13.2.4.1 Purpose of Disaster Investigation**

In contrast to disease outbreaks, the cause and the harmful effects of most disasters (natural or man-made) are relatively easy to identify and diagnose. Hence, the objectives of investigation in disease outbreak and disaster are different. In the event of disasters, investigation is conducted mainly for identifying the amount and the trend of harmful impact, and corresponding demands for staff, materials, and services. For example, it is important to understand the need for intervention and to determine the size and type of the intervention as soon as possible. Unfortunately, several weeks are usually needed to collect and organize precise and reliable data. Hence, the urgent work for disaster investigation is to conduct a rapid needs

assessment to get less-precise and less-reliable data, based on which crucial decisions are made. Besides rapid needs assessment, further investigation is also needed to collect more detailed, precise, and complete information, which is essential for the reconstruction of the community.

A rapid needs evaluation should be initiated as soon as possible, preferably within the first 3 days after the event. The objectives of rapid needs assessment are to assess the amount of affected population, disaster situation, local rapid response capacity, secondary disaster risk, resources needed, and the recommended actions.

#### **13.2.4.2 Key Steps in Carrying Out Disaster Investigation**

##### **1. Preparing for rapid health needs assessment**

Prior to rapid assessment, adequate preparation is required, such as collect background information about the emergency location, inspect safety conditions at the site of disaster, contact with local authorities and relevant organizations, and prepare essential equipment and supplies. Besides, a plan for the field assessment should be made to determine the order and manner of information collection, the forms for recording and analyzing the collected information, time schedule, and tasks assignment to each team member.

##### **2. Making checklists for rapid needs assessment**

Making a list of information to be collected during the preparation phase ensures the quality and comprehensiveness of needs assessment. A number of rapid assessments checklists are available. The checklist chosen should be adapted to the specific culture and background of the emergency.

##### **3. Identifying methods and sources for collecting data**

The main sources of information include health records at local health facilities; satellite maps; local pictures and videos; news media and social media information; vital statistics; medical insurance records; and field interviews with local health workers, health officials, or affected population. Please note that the interviewees should be selected from different demographic characteristics, such as gender, age, and race.

The method of data collection is determined by the type of the emergency, financial, material, and staff resources for the assessment as well as time requirements. Both qualitative and quantitative methods can be used, such as reviewing past records, observation, focus group discussions, key informant interviews, patient narratives, and questionnaire.

##### **4. Conducting rapid health needs assessment**

The following steps may provide a logical approach to rapid health needs assessment:(a) Conduct preliminary observation when approaching the site to evaluate the extent of population displacement and the damage caused by the disaster. (b) Interview local authorities and personnel, such as local government leaders, public health workers, volunteers, as well as people in affected communities to identify water and food supplies, medical services, and demographic characteristics. (c) Review existing records, such as maps, census data, and

surveillance system data (d) Conduct detailed visual inspection, like field surveys around affected communities to collect information on the structure of camps, living conditions and sanitation, population movement and migration, social influences and reactions, damage to transportation and communication systems, epidemic risk of infectious diseases, and priority health issues.

Rapid needs assessment is usually conducted through cluster or convenience sampling methods. The results are useful for planning follow-up interventions; hence, the main findings of needs assessment and corresponding recommendations should be reported to policy-makers or related agencies as soon as possible.

#### 5. Conducting thorough investigation

When the efforts of disaster rescue are more focused on the reconstruction or rehabilitation of the community, more precise and complete information or data are needed, such as detailed data on death and injury, nutritional status of the residents, the damage of health-care services and other public facilities, and long-term effect of the disaster on physical and psychological health. Such information plays an important role in guiding after-disaster reconstruction. Besides, the information can also be used as evidence for determining legal responsibility in man-made accidents.

### 13.3 Public Health Emergency Preparedness

It is generally believed that managing public health emergencies is about how to respond to the occurrence of it. However, this is but a corner of the picture. The management of public health emergencies must cover all the four phases of emergency: pre-emergency phase, response phase, recovery phase, and mitigating phase. In every phase, some certain measures should be taken, that is, preparedness, response, and recovery. Efforts conducted in all these phases are essential and valuable. For example, in preparedness phase, we take measurements to prevent the occurrence of emergencies or carry out researches to achieve precise prediction and early alarming of harmful events. When these events occur unavoidably, we can minimize the harm they cause. Previous experience has illustrated the important role of adequate preparation in advance on the successful management of public health emergencies, no matter natural or man-made. Hence, without adequate preparation, it is difficult to respond effectively and successfully, even with sufficient resources.

#### 13.3.1 *Definition of Public Health Emergency Preparedness*

Public health emergency preparedness (PHEP) is the ability to prevent, prepare for, rapidly respond to, and recover from public health emergencies in coordination with local government, health-care system, health-related institutes, communities, and individuals, especially when big threatens occur that exceed the normal scale.



Preparedness involves a continuous and long-term process of planning and implementation, and its success depends on proper planning, corrective actions, coordination, and selfless dedication of many people and institutes concerned. Public health preparedness requires the collaboration of government and community leaders to ensure that the community are adequately prepared to respond to a possible emergency, do their best to mitigate its damage and recover from the emergency as quickly as possible if they cannot prevent it.

### ***13.3.2 Significance of Public Health Emergency Preparedness***

Organizing responses to emergencies ensures adequate access to mental, medical and all other health-related services for the affected population. Emergency preparedness cannot completely eliminate all hazards, but it is a strong prerequisite for ensuring rapid and effective response to emergencies, thereby minimizing morbidity, mortality, and other damages. For example, governments can develop food security programs to ensure that food is protected from pests to avoid food crises, meteorological warnings and material reserves help residents to withstand extreme heat and extreme cold weather, and wide-coverage immunization procedure and systematic surveillance system help communities prevent from outbreaks of infectious diseases.

### ***13.3.3 The Main Activities of Public Health Emergency Preparedness***

Preparedness is reflected both in the ability to prevent possible public health emergencies and in the ability to respond adequately when emergencies occur. The main activities of public health preparedness are as follows:

#### **1. Establishing an emergency management agency**

Preparedness activities must be undertaken jointly by the many relevant health sectors in order to respond more quickly and effectively to an emergency or disaster. Hence, an effective national emergency management system is needed to organize and coordinate various departments to deal with emergencies effectively and efficiently. However, there is still no such comprehensive national emergency management system in many developing countries. Instead, the military is in charge of the relief effort by default. In China, a three-level emergency management system has been set up since 2003, which include national-, provincial-, and city-level agencies, with Prime Minister as the top leader. In 2006, Emergency Management Office of the State Council was set up, which is a significant milestone representing the building of complex emergency response system in

China. In April 2018, Ministry of Emergency Management of the People's Republic of China was set up. As the national-level agency for emergencies management, it coordinates emergency response in China.

## 2. Developing laws and regulations

In the event of the SARS outbreak (2003), the Chinese government realized the necessity of legalization in emergency management, and then issued *Regulations on Preparedness for and Response to Emergent Public Health Hazards* on May 7, 2003. After that, another law was issued on August 30, 2007, which was *Law of the PRC on Response to Emergencies*. Besides, there are some other related laws or regulations, such as *Law of the PRC on the Prevention and Treatment of Infectious Diseases*; *Law of the PRC on Prevention and Control of Occupational Disease*; *Hospital Infection Management Measures*; *Management of Information Report on Monitoring of Public Health Emergencies and Infectious Diseases*; etc. These laws and regulations identify the specific obligations and responsibilities of each related institute and individual.

## 3. Establishing action protocols

The purpose of establishing action protocols is to facilitate a prompt, efficient, coordinated response in the case of an emergency. Detailed protocols should be set up for all kinds of emergencies. Given the wide scope of emergencies, there are a corresponding number of different types of emergencies. Plans and protocols are developed and established to mitigate adverse impacts of emergency events and train the team to keep them ready.

## 4. Simulating emergency events to improve readiness

Planning, preparation, and practice are the keys to achieving success in the case of an actual emergency. One kind of practice is a simulation program, like a fire drill, or earthquake drill, which is focused on training departments and residents in the timely recognition and appropriate intervention for critical emergency events.

## 5. Training public and professional personnel

On the one hand, professional training on the knowledge and skills related to emergency preparedness and response should be given to emergency-related personnel, such as public health professionals, emergency physicians, clinicians, other health-care workers, and even firefighters or military soldiers. On the other hand, appropriate training and education to community residents in advance is also relevant, which helps public know basic knowledge on common infectious diseases, emergency rescue skills, and proper response to a public health emergency.

Various ways of training can be used, including lectures, courses, and network training provided by universities, Centre for Disease Control and Prevention, and academic health centers. Moreover, emergency drills and exercises are important part of training, familiarizing the personnel with the practical application of emergency procedures, systems, and facilities.

## 6. Reserving supplies and staffs

Disaster relief goods and materials should be prepared in advance and specially used for the rescue, transfer, and arrangement of the affected population in

the emergency. These materials may be provided by governments at all levels or donated by public or private institutions/organizations and individuals. Till 2010, the Chinese government has set up 17 central-level lifesaving disaster relief materials storage warehouse, including foods, drinking water, medical goods, and daily necessities. Besides, some specific warehouses are building for certain kinds of emergencies, like fire, flood, earthquake. In addition, prophylactic medicines, such as antitoxins, antibiotics, chemical antidotes, and vaccines should be reserved. Professional staffs should also be cultivated by school education or on-the-job training.

#### 7. Monitoring and early warning

Continuous and systematic monitoring, also called “surveillance,” public health data, as well as analyzing and interpreting these data, provides policy-makers critical information for planning, implementation, and evaluation of public health interventions. As the foundation of emergency preparedness, surveillance is to monitoring any health-related changes or patterns. Based on monitoring system, abnormal signs and changes which may adversely affect communities can be detected promptly, thereby, there is enough response time for emergency systems to adequately prepare for possible disasters and minimize their harmful effects. Especially, for certain events, like disease outbreaks, local armed conflicts, and floods, the authorities may issue warnings to alert people to the impending dangers. This can reduce the material and economic losses and prevent loss of life. Different colors can be used to represent different threat alert levels. For example, in American green means low threat, blue means general threat, yellow means significant threat, orange means high threat, and red means severe threat. The main responsibilities of a comprehensive monitoring and early warning system includes data monitoring, risk analysis, event detection, warning release, communication, and feedback, with the aim of realizing complete surveillance, accurate prediction, and timely warning.

#### 8. Risk assessment

Risk assessment is a risk analysis of predefined vulnerabilities and hazards based on systematically collected data. In a broader sense, risk assessment includes daily risk assessment and specific risk assessment. The former is conducted in the pre-emergency phase to assess the hazard vulnerability in an area, and the latter is carried out after the onset of an emergency to assess the potential further risk caused by this event. In a narrower sense, risk assessment only refers to daily risk assessment.

Daily risk assessment involves gathering information about the most common risk factors affecting the area, the likelihood and risk of emergencies, the extent of damage to infrastructure, emergency response capacity and weaknesses of public facilities, and the amount and characteristics of vulnerable populations.

Specific risk assessment would be conducted once an emergency is detected or notified, and then confirmed. Specific risk assessment takes into account the actual site conditions and address only the relevant hazards, and then results in grading the event, which determines the level of response that needs to be taken and activating the appropriate emergency response. Event grading is based on

five criteria: scale, severity, urgency, respond ability of local or national government, and government's reputational risk.

## **13.4 Public Health Emergency Response**

### ***13.4.1 Definition of Public Health Emergency Response***

Public health emergency response refers to the actions, which are taken immediately following an emergency, to save lives, provide assistance, minimize economic damage, and accelerate post-disaster reconstruction, so as to mitigate the adverse impact on the public and society.

### ***13.4.2 Significance of Public Health Emergency Response***

Rapid and proper response to an emergency is important with respect to life safety (which is usually the Number 1 Goal), stabilizing the emergency and protecting the environment and property.

### ***13.4.3 The Main Activities of Public Health Emergency Response***

Emergency response is a dynamic process. The response actions should be initiated during the first 24 h of an incident. Specialized rapid-response teams are needed to respond quickly to new disasters and disease outbreaks.

#### **13.4.3.1 Ensuring Availability of Preventive and Emergency Medical Treatment**

When disease outbreak or disaster occurs, the top priority of any public health response is to save lives and control the spread of disease. Many resources are required, including isolation treatment hospital (for communicable diseases), medical supplies (antitoxins, antibiotics and chemical antidotes, laboratory agents, and equipment), staff (medical workers, health-care workers, public health personnel, volunteers, and psychological counselor), guidelines for diagnostic and treatment, transport, and stationery. Countermeasures for communicable diseases must be administered to patients, prophylactics must be provided to high-risk groups when necessary, like vaccine or prophylactic antibiotics, and psychological canceling must

be provided alongside treatment. Please note that medical workers must attach enough importance to personal protection, patients isolation, and environment and items disinfection to avoid nosocomial infections.

#### **13.4.3.2 Preventing Secondary Public Health Emergencies After Disaster**

The primary cause of mass casualties from terrorist attacks or natural disasters may not be chemical or biological in nature, but be the subsequent secondary hazards, such as infectious diseases (such as cholera and plague) epidemic, leakage and spread of toxic gas, destruction of lifeline supplies (communications, transportation, water supply, power supply, etc.), and social unrest (robbery). For instance, World Trade Center explosion, which caused severe damage to urban facilities and high levels of social panic may further lead to secondary infectious diseases and poisoning. Another example is that on March 11, 2011, the earthquake in Japan triggered a tsunami, causing a nuclear leak at the Fukushima nuclear power plant. Hence, in emergency situations, priority should be given to preventive measures to avoid further deaths or more health threatens. Local public health managers protect the affected and surrounding areas from secondary health threats caused by pest or rodent infestations, ensure adequate water, food, and living space, dispose garbage and human feces, promote hygiene practices, handle dead bodies appropriately, etc.

#### **13.4.3.3 Interrupting the Route of Transmission**

Public health authorities should take measures to prevent further spread of the disease as soon as the source of the disease is identified. Measures taken include isolating sources of outbreak, such as closing contaminated restaurants or water supply, isolating and treating communicable disease patients, and enclosing buildings. When a disaster is severe enough, the government has the right to declare a lockdown in the affected areas. In addition, there must also be an adequate contact tracing workforce to track and trace cases and contacts to prevent the spread of outbreaks.

#### **13.4.3.4 Remediating of Environmental Health Conditions**

The role of public health authorities also includes decontamination and disinfection of affected site and facilities. Decontaminating is the neutralization, removal or destruction of toxic or hazardous substances, such as toxic substances, radioactive materials, or disease pathogens, in order to avoid harm to other patients and health-care providers, and prevent secondary pollution and nosocomial infection. The manner and extent of decontamination mainly depends on the nature of the hazardous substance and its viability (microorganisms) and degradation rate (radiation).

#### **13.4.3.5 Performing Laboratory Analyses to Support Epidemiology and Surveillance**

Laboratory tests can strongly support epidemiological survey and surveillance. In many instances, laboratory work is inseparable from the discovery of pathogens or chemical poisons, the diagnosis of patients, and the determination of harmful substances in environment. In some special cases, new disease pathogens must be detected through more sophisticated laboratory analysis such as RT-PCR and high-throughput DNA sequencing. In addition, laboratory techniques can be used for the development of vaccines or new drugs, as well as for the study of pathogen resistance or homology.

#### **13.4.3.6 Communicating with Media and Delivering Message to the Public**

Information about the emergency must be communicated to the public through the media, telling them the good, the bad, the ugly, and what we do not know yet. This communication can also be called risk communication. As one of the key countermeasures for emergencies, risk communication is the timely exchange of information, knowledge, attitude, and advice between government or health officials or relevant experts and the affected people. Effective risk communication reduces mortality and morbidity by promoting good personal and home hygiene and improving self-rescue and self-care ability, it also helps government maintain political and economic stability of the country. Therefore, all countries should regard risk communication as a core part of their efforts to respond to emergencies. The ultimate goal of risk communication is to enable people at risk to respond correctly when a risk occurs to mitigate the effects of a hazard. The message delivered in risk communication must be simple, timely, accurate, relevant, credible, and consistent. In addition, more emphasis should be placed on effective public education and two-way conversation, as well as timely communication of risks to the public.

### **13.5 Summary**

In the management of public health emergencies, epidemiology has the ability to monitor, detect, and investigate potential hazards, and to maintain and improve the systems necessary to support this capability. Epidemiological functions would be especially important to address hazards that are environmental, radiological, toxic, or infectious in nature. Epidemiological functions are one of several core public health capabilities as being critical for public health emergency preparedness.

For public health emergencies, preparedness and response are inextricably linked. Preparedness is based on lessons learned from both actual and simulated response

situations. An effective response is all but impossible without extensive planning and thoughtful preparation. Public health emergency management conveys the important idea that protecting populations and property involves the estimation of risks, preparation, and activities which will mitigate the consequences of predictable hazards and post-disaster reconstruction in a way that will decrease vulnerabilities. An important goal is building a culture of awareness that preparation is not only possible but also will greatly reduce the consequences of disasters in terms of human and economic loss. In these, public health is an important partner with engineers, planners, elected leaders, and community organizations.

# Chapter 14

## Molecular Epidemiology



Hui Wang

### Key Points

- Describe the concept of molecular epidemiology
- Familiar with the concept and classification of biomarkers and know how to select biomarkers.
- Understand the relationships between molecular epidemiology and traditional epidemiological methods
- Discuss, apply and interpret application of molecular epidemiology in disease control and prevention.

In molecular epidemiology, the study of the determinants of disease will pay attention to the causative, protective, and predisposing factors (including infectious agents and various environmental exposures, e.g., chemical or physical agents and lifestyle habits) and host characteristics such as genetic susceptibility. Most of these studies are performed via molecular techniques within the molecular biology.

## 14.1 Introduction

### 14.1.1 Concept

Molecular epidemiology is defined as the study of the epidemiology of human diseases by application of the techniques of molecular biology at the population level. Molecular investigations can contribute to the elucidation of diseases' etiology.

---

H. Wang (✉)  
School of Public Health, Peking University, Beijing, China  
e-mail: [huiwang@bjmu.edu.cn](mailto:huiwang@bjmu.edu.cn)



Proposing the clear definition of biomarkers is the vital step in the study of molecular epidemiology. Biomarkers are referred to any markers which could be detected and represented the changes from the exposure to the onset of disease on a population scale. The range of biomarkers is quite broad, including cellular, biochemical, and immunological elements. Generally, all biomarkers (e.g., nucleic acid, protein, lipids, and antibody) can be investigated in the molecular epidemiology. According to the process of disease, biomarkers applied in the molecular epidemiology are divided into three categories: markers of exposure, markers of biological effects, and markers of susceptibility.

### ***14.1.2 Characteristic***

Compared with conventional epidemiology, molecular epidemiology lay emphasize on the knowledge of the pathogenesis of diseases by elucidating specific molecular pathways and pointing out specific molecules or genes that influence the risk of developing diseases. For instance, genetic biomarkers rather than family history might be more precise in characterizing host susceptibility. Molecular epidemiology can enhance the validity and reduce bias in the assessment of environmental exposures. It can predicate the onset of disease at the subclinical level and provide tools to discern heterogeneity within a disease, such as the development of breast cancer subtypes (i.e., basal, luminal A, luminal B, normal breast-like, and ERBB2<sup>+</sup>).

Not all biomarkers are suitable for molecular epidemiological studies due to expensive cost or intensive labor. It is necessary to examine those laboratory techniques used in the studies of molecular epidemiology, for their sensitivity, validity, specificity, and variability within and between laboratories before using them in any epidemiological research. Meanwhile, the acquisition of appropriate biological specimens, costs, and ethical issues need to take into considerations in molecular epidemiological research design as well.

This chapter shows the introduction of main characteristics of molecular epidemiology, the description of three major categories of biomarkers, most commonly used research methods and application and prospective of molecular epidemiology. This chapter will present an overview of molecular epidemiology. The authors refer the readers to more articles published recently and update the knowledge of molecular epidemiology.

## **14.2 Classes of Biomarkers**

### ***14.2.1 Biomarkers of Exposure***

Molecular epidemiological studies intend to establish the causal and biological associations between exposures and diseases. The exposure is defined as any contact

with physical, chemical, or biological agents by the International Programme on Chemical Safety. Exposure assessment should be continuous which requires that biomarkers of exposure should be continuous as well. Thus, this assessment will provide more exact knowledge with regard to the exposure-disease association.

Molecular epidemiology implements biomarkers to improve exposure assessments in the complexity of distinguishing respect between the effect of individual and environmental factors to diseases etiology. Validated biomarkers could measure the disease process at the individual level which leads to the causal inference or biological plausibility of an exposure-disease association. For instance, in the area of virology, antibodies have been used to identify what kind of virus which a person has been infected with. Furthermore, by measuring the accumulation of chemical agents or metals in biospecimen, such as arsenic in hair or mercury in fingernails or toenails, it can directly measure an exposure at the individual level.

Using sensitive laboratory techniques, low levels of exposure can be detected by trace analysis of biomarkers. Most biomarkers usually represent the exposure of environmental toxicants, nonetheless, they could identify the crude amount of ingested dietary components as well, such as bacterial or viral infections. Moreover, they also serve as terminuses for a determination of the success of interventional strategies.

Biomarkers of exposure are classified relying on what they measure, either an internal dose (i.e., serum vitamin D) or a biological effective dose (i.e., a dose that causes DNA damages). Biomarkers of internal dose estimate the presence of environmental chemicals and their metabolites in human tissues, excretions, and/or exhaled air. Further, measurements of dietary biomarkers either as “recovery” biomarkers, for example, measurements of sucrose and fructose in 24-hour urine samples for direct assessment of sugar consumption, or as “concentration” biomarkers, such as serum carotenoids, which indirectly indicates dietary intake since they are the results of complex metabolic processes. Exposure biomarkers also serve as evaluation indicators for the effects of interventional strategies. The utility of such biomarkers is restricted to the availability of detectable levels of the compound. Although biomarkers of exposure are often described separately from biomarkers of effect, many actually overlaps exist, which can be partially attributed to the fact that the biomarkers provide information related to both of the exposure and the effect. For example, lymphocytes could be a surrogate for exposure and also a target for the exposure’s effect.

### ***14.2.2 Biomarkers of Effects***

Biomarkers of effects measure the interaction between an agent and/or its metabolites and target cell(s) or molecule(s); they are defined as “measurable changes in the organism.” This definition consists of markers of effects that can indicate a preclinical response and is not always detected by using traditional clinical diagnostic techniques. The early effects of an individual can be used as informative markers

of disease risk. Several molecular-based assays have been developed to identify cellular response(s) activated by such exposures.

The amount of the agent that reaches a crucial cellular target can be detected by biomarkers of a biologically effective dose, for instance, DNA adducts or the amount of a chemical agent bound to a cellular receptor. Other biomarkers of effect measure the damage caused by the agent. For instance, under situation of a biological effective of UV dose, DNA damage induced by UV is measured and UV length can be further used to classify the type of damage. UVA exposure induces base excision repair, while UVB exposure induces nucleotide excision repair. Biomarkers of DNA damage include mutations, DNA strand break, adducts, micronuclei, sister chromatid exchanges, and chromosomal aberration.

### ***14.2.3 Biomarkers of Susceptibility***

Molecular epidemiology provides tools to identify the genetic and acquired susceptibility (such as DNA repair capacity). At present, association studies are the most common genetic epidemiological studies.

Abundant studies have been carried out on candidate genes based on biochemical hypotheses in terms of DNA repair, carcinogen metabolism, or cell cycle. Multiple GWAS and meta-analyses (see below) are currently evaluating large numbers of SNPs for an overview of methods and genetic loci that seems to correlate with diseases. Although these studies include thousands of cases and controls, which tend to be huge, some smaller studies with very well-designed selection of subjects have contributed to the understanding of genetic susceptibility as well. For instance, in the study by Klein et al., a genome-wide screen of 96 cases and 50 controls shown that an intronic and common genetic variant in the complement factor H gene (CFH) played a crucial role in age-related macular degeneration (OR = 7.4, 95%CI = 2.9~19).

### ***14.2.4 Biomarker Selection***

Before starting the experiments, several issues should be considered when identifying candidate biomarkers. For example, it should be known the prevalence of biomarkers of interest in the population. The ability of the biomarker representing the agent of interest and the sensitivity and specificity of the biomarker measuring low dose of exposure should be considered as well. Additionally, the validity of the biomarker and reliability must be determined. Epidemiological studies implementing biomarkers must take into consideration that environmental exposures will vary qualitatively and quantitatively over time. It also should be taken into account that part of biomarkers decay over their lifetime, thus, when selecting an appropriate design of a molecular epidemiological study, the half-life of biomarkers

must be considered. Most biomarkers are transient with a relatively short half-life period. In conventional epidemiology, case-control studies have great advantages in research of which the disease of interest is rare and the exposure is frequent and easy to identify. For instance, when investigate cancers or other chronic diseases, studies are usually focused on events that happened many years before the disease onset and often involve chronic exposures. However, implementing molecular epidemiology in chronic diseases, the method of case-control is restricted if the biomarker has a short half-life time. Therefore, it indicates an acute exposure that occurred in a short time before disease onset. However, such studies are often confined in the sense that they are using a “one-time” biological sample, which cannot certainly represent the common exposure or of changing exposures. For those biomarkers, prospective molecular epidemiological studies are more suitable.

In molecular epidemiological study, biomarker validation contains several issues that need to be considered when assessing the utility of a biomarker. Analytical validity, clinical validity, clinical utility, and ethical, legal, and social implications and safeguards are often regarded as the *ACCE* evaluation of a biomarker. The analytical validity points at the ability of a test to reliably and accurately measure the genotypes/markers of interest which includes its sensitivity and specificity. The ability of a genetic test to detect or predict the phenotype is the clinical validity or the positive predictive value. The clinical utility component of this assessment considers the risks and advantages related to the incorporation of the test into routine clinical practice. Recently, the legal, ethical, social implications and safeguards are other issues that need to consider when assessing the utility of a biomarker.

Betsou and colleagues recommended several methods to evaluate the vulnerability of a biomarker to pre-analytical variation. These evaluations can be conducted to ensure that association with clinical end points is not because of uncontrolled pre-analytical variation. For example, the characteristics could change rapidly (e.g., vitamin C is light-sensitive) for serum is not processed rightly. Other reasons of pre-analytical variations include fasting conditions, specimen collection's time, the position of the patient when collecting, the patient's diet, or other life habits, all of which are necessary consideration when choosing appropriate biomarkers. The use of inappropriate biomarkers may partly due to the publication bias which causes false-positive associations. Many biomarkers were tested but never published because of the unfavorable results. Publication bias might be a result of time consuming and costly assays, such that the positive findings of manuscripts are more likely to be published than negative findings, although they may have been acquired by chance.

## **14.3 Main Research Methods Used in Molecular Epidemiology**

### ***14.3.1 Study Design in Molecular Epidemiology***

In the past several years, it might be the most revolutionary changes in molecular epidemiology for the emerging of discovery technologies that can be put into use in many study designs, such as genome-wide scans of common genetic variants, messenger RNA (mRNA) and microRNA expression arrays, proteomics, and metabolomics (also referred to as metabonomics). These approaches are helping investigators to explore biological responses to exogenous and endogenous exposures, to evaluate potential modification of those responses by variants in essentially the entire genome, and to define tumors at the chromosomal, DNA, RNA, and protein levels. Biomarkers of genetic and environmental factors referred to human disease have been applied to cross-sectional studies, case-control studies, and cohort studies.

#### **14.3.1.1 Cross-Sectional Studies**

Cross-sectional studies can be used to assess allele and genotype frequencies, exposure levels in the population, and the relationships among genotypes, exposures, and phenotypes. Although cross-sectional studies cannot infer causality between incidence and natural history, they can provide information on genetic variants and environmental exposures that may help to guide research and health policy at population level. For example, a population-based prevalence study analyzes two common mutations in the hemochromatosis gene (HFE) (C282Y and H63D variants of HFE) in the U.S. population. Steinberg et al. genotyped 5,171 samples from the Third National Health and Nutrition Examination Survey (NHANES III) of Center of Disease Control and Prevention (CDC), a nationally representative survey conducted in the United States from 1992 to 1994. Genotype and allele frequency data were cross-classified by sex, age, and race/ethnicity. The CDC provides an ongoing assessment of the U.S. population's exposure to environmental chemicals by the analysis of NHANES surveys. The first National Report on Human Exposure to Environmental Chemicals was issued in 2001 and presented exposure data for 27 chemicals from NHANES 1999–2001. In 2003, The second report presented exposure data for 116 environmental chemicals stratified by age, gender, and race/ethnicity. Furthermore, by cooperating with the National Cancer Institute (NCI), the CDC applied the NHANES III survey to measure prevalence of variants in 57 genes and correlate the resulting genotypes with clinical, medical history, and laboratory data. When completed, such studies will provide valuable information on the association between genetic variations and numerous health end points.

### 14.3.1.2 Case-Control Studies

The case-control approach is especially well suitable to study genetic variants in that (a) unlike other biologic markers of exposures such as DNA adducts and hormonal levels, genetic markers are stable indicators of host susceptibility; (b) case-control studies can implement an all-sided search for the effects of several genes, along with other risk factors, and look for gene-environment interactions; and (c) case-control studies are suited for plentiful unusual disease end points (e.g., specific cancers and birth defects). Furthermore, because the environmental exposures change over time, cohort studies with repeated biomarkers of exposures and intermediate outcomes may be preferable to case-control studies, unless case-control studies are nested within an underlying cohort of a well-defined population for which biological samples stored at the start of the study are later analyzed for exposures. Case-control studies can synchronously support gene discovery and population-based risk characterization. For instance, registries of population-based incident disease cases and their families offer a platform to conduct family-based linkage and association studies. The reflection of this philosophy is the NCI sponsors Cooperative Family Registries for Breast and Colorectal Cancer Research. Population-based case registries can support many study designs, including extended family studies, case-parent trios, and case-control family designs. One type of family-based association study is the kin-cohort design in which researchers access the genotype-specific risk of disease occurrence in first-degree relatives of study participants (proband), inferring genotypes of relatives from genotypes measured in probands.

### 14.3.1.3 Cohort Studies

Efforts are now being done to integrate genomics into cohort studies started in the pregenomic era to study disease incidence and prevalence, natural history, and risk factors. Well-known cohort studies include the Framingham study, the Atherosclerosis Research in Communities study, the European Prospective Investigation on Cancer, and the newly designed National Children Study, a planned U.S. cohort study of 100,000 pregnant women and their offspring to be followed from before birth to age 21 years. In addition, the genomics era is enlightening the development of very large longitudinal cohort studies and even studies of entire populations to set up repositories of biologic materials (“biobanks”) for discovery and characterization of genes relevant to common diseases. There are adequate number of studies could be listed, which range from large random samples of adult populations such as the UK Biobank ( $N = 500,000$ ) and the CartaGene project in Quebec ( $N = 60,000$ ) to populations of entire countries such as Iceland ( $N = 100,000$ ) and Estonia ( $N = 1,000,000$ ; Estonian Genome Project), to a cohort of twins in multiple countries (GenomeEUtwin). It is worth mentioning that the China Kadoorie Biobank was launched in 2004, which recruited 0.5 million people with blood data and then collect their health information for at least two decades. These biobanks can help

epidemiologists to quantify the occurrence of diseases in multifarious populations and to understand their natural histories and risk factors, including gene-environment interactions.

Longitudinal cohort studies allow for repeated phenotypic and outcome measures of individuals over time, including intermediate biochemical, physiologic, and other precursors and sequels of disease. Cohort studies can also be applied to nested case-control studies or even as an initial screening method for case-only studies (as explained before). Such studies will generate abundant data on disease risk factors, lifestyles, and environmental exposures, and make preparations for data standardization, sharing, and joint analyses. An example of data standardization across international boundaries is the global P3G (Public Population Project in Genomics), which, to date, includes three international studies from Europe and North America. “Harmonization” is vital for creating comparability across sites on measures of genetic variation, environmental exposures, personal characteristics and behaviors, and long-term health outcomes.

### ***14.3.2 Main Molecular Methods Used in Molecular Epidemiology***

Numerous molecular biological techniques are implemented in the epidemiological studies. This section listed main methods used in the molecular epidemiological studies.

#### **14.3.2.1 Electrophoretic Mobility Shift Assay (EMSA)**

The EMSA or mobility shift electrophoresis referred as a gel mobility shift assay, gel shift assay, gel retardation assay, or band shift assay as well, a usual affinity electrophoresis techniques, is used to study protein-DNA or protein-RNA interactions. This procedure can confirm if a protein or mixture of proteins is able to combine with a given DNA or RNA sequence. Sometimes, it can be used to indicate if more than one protein molecule take part in the binding complex. Gel shift assays are often performed in vitro concurrently with DNase footprinting, primer extension and promoter-probe experiments when studying transcription initiation, DNA replication, DNA repair or RNA processing and maturation. Precursors can be found in earlier literature, but most present assays are based on methods described by Garner and Revzin and Fried and Crothers.

The EMSA technique is based on the observation that protein-DNA complexes migrate more slowly than free linear DNA fragments when subjected to non-denaturing polyacrylamide or agarose gel electrophoresis. Because the rate of DNA migration is shifted or retarded when bound to protein, the assay is also defined as a gel shift or gel retardation assay. The ability to resolve protein-DNA complexes

depends greatly on the stability of the complex during each step of the procedure. During electrophoresis, the protein-DNA complexes are quickly resolved from free DNA, providing a “snapshot” of the equilibrium between bound and free DNA in the original sample. The gel matrix provides a “caging” effect that contribute to stabilize the interaction complexes: even if the components of the interaction complex dissociate, their localized concentrations remain high, promoting positive reassociation. Additionally, the relatively low ionic strength of the electrophoresis buffer helps to stabilize transient interactions, permitting even labile complexes to be resolved and analyzed by this method.

Protein-DNA complexes formed on linear DNA fragments lead to the characteristic retarded mobility in the gel. However, if circular DNA is used (e.g., mini-circles of 200–400 bp), the protein-DNA complex may actually migrate faster than the free DNA, analogous to what is observed when supercoiled DNA is compared to nicked or linear plasmid DNA during electrophoresis. Gel shift assays also help to resolve altered or bent DNA conformations that induce by the binding of certain protein factors. Also, gel shift assays are suited for protein-RNA and protein-peptide interactions by using the same electrophoretic principle as well.

#### 14.3.2.2 Dual Luciferase Reporter Assay

The wide applications of genetic reporter systems help to study eukaryotic gene expression and cellular physiology, including the study of receptor activity, intracellular signaling, transcription factors, mRNA processing and protein folding, and so on. Dual reporters are usually applied to enhance experimental accuracy. The term “dual reporter” refers to the simultaneous expression and measurement of two individual reporter enzymes within a single system. Generally, the “experimental” reporter has relation to the effect of specific conditions of experiment, while the activity of the co-transfected “control” reporter offers an internal control that act as the baseline response. Normalizing the activity of the experimental reporter to the activity of the internal control minimizes experimental variability caused by differences in cell viability or transfection efficiency, which also can effectively eliminate other sources of variability, including differences in pipetting volumes, assay efficiency and cell lysis efficiency, and so on. Hence, dual-reporter assays often permit more reliable interpretation of the experimental data by reducing extraneous influences. The Dual-Luciferase Reporter (DLR™) Assay System offers an efficient method of performing dual-reporter assays. In the DLR™ Assay, the activities of firefly (*Photinuspyralis*) and Renilla (*Renillareniformis*, also known as sea pansy) luciferases are measured sequentially from a single sample. The firefly luciferase reporter is measured first by adding Luciferase Assay Reagent II (LAR II) to generate a stabilized luminescent signal. After quantifying the firefly luminescence, this reaction is quenched, and the Renilla luciferase reaction is simultaneously initiated by adding Stop & Glo Reagent to the same tube. The Stop & Glo Reagent also produces a stabilized signal from the Renilla luciferase, which decays slowly over the course of the measurement. In the DLR™ Assay System, both reporters



yield linear assays with subattomole sensitivities and no endogenous activity of either reporter in the experimental host cells. Furthermore, the integrated format of the DLR™ Assay provides rapid quantitation of both reporters either in transfected cells or in cell-free transcription/translation reactions. Promega provides the pGL4 series of firefly and Renilla luciferase vectors designed for use with the DLR™ Assay Systems. These vectors may be used to co-transfect mammalian cells with experimental and control reporter genes.

#### **14.3.2.3 The Comet Assay**

The “comet” assay was developed in the late 1980s/early 1990s and used only some lymphocytes. The lymphocytes are frozen at a very low temperature to ensure their viability, and then treated and run out on a gel that was spread on a glass slide. DNA from the cell “migrates” to form a “tail.” If DNA is “broken” (i.e., single-strand breaks), then the length of the tail is relative to the amount of breakage. This assay tends to measure DNA single-strand breaks, cross-links, base damage, and apoptotic nuclei. Cells could be subject to damaging agents first, then allowed to repair, and placed on the gel on the glass slides. In this situation, this assay measures DNA repair “capacity” by the length of the comet tail. The comet assay is commonly used in assessment environmental toxicant-induced DNA damage. The application of this assay exponentially increased based on its high sensitivity and specificity. This method also enables researchers to detect increased risk for different health outcomes. Massive validation efforts have been taken on optimizing standardization and reliability of the comet assay by the European Standards Committee on Oxidative DNA Damage.

#### **14.3.2.4 Micronucleus (MN) Assay**

MN assay, which is used to detect MN, extracellular bodies, after the cells go through first cell cycle, has the ability to discern chromosome breaks from aneuploidy (abnormal number of chromosomes) and can detect chromosome loss. Since MN are formed from acentric chromosomal fragments or chromosomes that are not involved in either daughter nuclei, they are classified relying on whether they include chromosomal fragments or whole chromosomes.

This assay is suited for use in molecular epidemiological studies for the relative ease of scoring, limited costs and personnel requirements, and the precision that scoring larger numbers of cells provides. The MN assay can be proceeded in peripheral blood lymphocytes, alveolar macrophages, erythrocytes, epithelial cells, and fibroblasts. In this assay, the cells under investigation must survive at least one round of nuclear division, so some of the damaged cells are lost before the analysis begins, and the survivability of the damaged cells is not known with this assay.

A review of published evaluated the occurrence of MN and the influence of genotoxic exposures on MN frequency in children and adolescents. This review

indicated that this cytogenetic assay is a helpful and sensitive tool which is suitable for biomonitoring studies of children including those with low-dose exposures to environmental agents. The confounding effects of age, sex, and chronic and infectious diseases on MN levels were evaluated in these studies, and the only variable irrelevant to MN frequency was sex.

### ***14.3.3 Genome-Wide Association Studies (GWAS)***

GWAS are designed to identify the entire human genetic associations with detectable traits or the presence or absence of a disease of interest. The precondition is the entire genome can be assessed for variation and a few SNPs would stand out as key risk factors of disease. The comparison is carried out between individuals with and without the disease of interest. Since the genome is large and the number of SNPs is countless, participants by the thousands are required to suitably investigate the associations. The method of GWAS takes advantages over candidate gene studies and it enlarges the potential of exploration of genetic analyses. GWAS recruit numerous study subjects with a disease or phenotypic trait of interest. The study subjects usually originate from ongoing collaborative scientific work including different institutions or over all the continents. These studies take benefits of high-throughput genotyping technologies, DNA isolation, automated collection of biospecimen, and high-quality-control practices, and then employ statistical analyses to determine associations between qualified SNPs and diseases or phenotypic trait of interest. Great efforts from the laboratory and biostatistical have contributed to thousands of GWAS so far, which, no doubt, conduce to the knowledge base of molecular epidemiology around the world. Accurate GWAS would replicate their results in different populations or in experimental animals, when the biological pathways have mechanistic modeling. Regarding the “common disease, common variant” hypothesis, GWAS depending on SNPs as markers of allelic variants that indicates over 1–5% of each human genome. By genetic characterization, and then fine mapping and analyses, researchers are capable of determining common genetic variations of chronic diseases.

Generally, genome-wide scanning is conducted on an initial group of cases and controls, and then a smaller standout SNPs are assessed to replicate findings in a second and a third set of cases and controls. The possibility of false-positive or false-negative findings will be reduced by the performance of such multistage study design. Furthermore, it reduces the genotyping costs as well. Additionally, with employed quality controls, the replicated genotyping provides the essential validation, particularly for SNPs of intron or unknown functional region.

Biases are inclined to happen in GWAS. Especially, population stratification is one of the most crucial confounders. For instance, a potential population structure leads to false-positive associations when the detected SNPs are also linked with unknown factors which reflect geographical origin or ethnicity of study individuals. The vast data produced by GWAS is prone to false-positive associations. Effective

statistical skills must be used to decrease the possibility of false positives raised by a lot of multiple comparisons.

Another common limitation is that current statistical methods used in GWAS capture a large number of common variants, which derived from the concept of linkage disequilibrium or other statistical algorithms validated according to the Human Genome International HapMap databases. These methods establish on the theory of human genome is constituted by blocks of nucleotides named haplotypes. Haplotypes are inherited together. Some SNPs within a given block define and explain or “tag” within block variability. These tagging SNPs get popular in GWAS. However, certain variants may not be captured by the current genotyping chips while they potentially represent crucial but unknown function. Besides, it is necessary to consider that some genetic variants might be influenced only when combining with exposures that initiate or modify expression of that gene. Without considering exposures to assess risks of chronic disease, we cannot successfully reveal the complicated patterns of gene-environment or gene-gene interactions which contribute to a great degree chronic disease risk. “Next-generation GWAS” probably should combine with more detailed analyses of common exposures (e.g., smoking, alcohol, dietary patterns, air pollutants, over-the-counter medications (like common non-steroidal anti-inflammatory drugs (NSAIDs)), and recreational drugs), which influenced the chronic disease etiology and pathogenesis.

#### ***14.3.4 Mendelian Randomization (MR)***

Confounding, selection bias, and reverse causation are major problems in building causal relationships between exposures and diseases, which may lead to spurious associations. MR is a method by using genetic variations of known function to detect the causal effect of a modifiable exposure on disease in nonexperimental situation. A vital characteristic of observational epidemiology is to identify the causes of common diseases which public health takes interest. For the purpose of confirming the favorite effects of a recommended public health intervention, the association of observation between the certain risk factor and a disease must prove that the risk factor indeed causes the disease. Well-known successful examples are that causal relationships are identified between smoking and lung cancer, and between blood pressure and stroke. However, there are failures when identified exposures were later demonstrated by randomized controlled trials (RCTs) to be noncausal. For example, hormone replacement therapy (HRT) was previously thought prevent cardiovascular disease. However, it did not and may even have other adverse effects in health. In observational epidemiological studies, the confounders such as social, behavioral, or physiological factors result commonly in such spurious findings. They are easy to uncontrol and especially difficult to measure accurately. Furthermore, many findings repeat unlikely by RCTs for ethical reasons.

MR allows one to test for a causal effect from observational data in the presence of confounders by taking common genetic polymorphisms with well-understood

effects on exposure patterns. Necessarily, the genotype must only affect the disease process directly through its effect on the exposure. Since genotypes are assigned randomly when inherit from parents to offspring during meiosis, if we hypothesized that option of mate is unrelated with genotype (panmixia), the genotype distribution among population should be irrelevant to confounders that commonly trouble observational epidemiological studies. Therefore, MR can be considered as a “natural” RCT. From a statistical perspective, it is a use of instrumental variables, with genotype serving as an instrument/proxy for the exposure.

The same with all studies of genetic epidemiology, trouble exists in the requirement for large sample sizes, the non-replicable results, and the lack of functional proof on genetic variants. In addition to these limitations, genetic findings could be confounded by other genetic variants by linkage disequilibrium with the variant under study or by population stratification. Moreover, pleiotropy of a genetic variant may contribute to null associations on account of canalization of genetic effects. If correctly performed and carefully interpreted, MR studies can offer valuable evidence to identify causal hypotheses between environmental exposures and common diseases.

## **14.4 Application and Prospection**

### ***14.4.1 Control and Prevention of Infectious Diseases***

The aim of molecular epidemiology of infectious diseases is to apply molecular (amino acid or nucleotide) sequences to study the ecology and dynamics of pathogens. For infectious diseases, it includes the transmission system (source of infection, transmission route, and susceptible population), pathogenesis and virulence of the microbe, the interaction between microbe and the human (or other) host(s), and the microbiota of the host (the area microbes usually live on and in the human body).

#### **14.4.1.1 Outbreak Investigation**

In all outbreak investigations, setting the definition of a case is a key step. Molecular techniques are the standard tool in an outbreak investigation for clarifying case definitions, enhancing specificity, and decreasing misclassification. During an outbreak of disease, it is commonly assumed that a single microbe causes the clinical symptoms. A microbe of the same genus and species but different strains is possible cause of disease during the same period. Case definitions can be refined by including the molecular typing which would increase the specificity, reduce misclassification of non-outbreak cases with outbreak cases, and potentially increase the possibility to identify the outbreak source. Only based on clinical symptoms, we are hard to distinguish between diseases. This could make outbreak investigations complicated, especially if the symptoms are not very typical. For instance, lots of viruses could

cause flulike symptoms; however, classification of influenza based on clinical symptoms is specific only during an epidemic when a large number of flulike patients suffer from influenza. Even during an epidemic, the confirmation from laboratory is required as well, since there may be not only one strain of influenza in transmission. In 2008, there were two predominant influenza A strains in circulation: H1N1 and H3N2. Laboratory test is particularly helpful for identifying individuals with mild or atypical symptoms, and determining the specific type. A variety of methods of laboratory tests could provide a molecular fingerprint based on the microbial genotype. For example, pulsed-field gel electrophoresis (PFGE) is applied as the standard method for foodborne outbreaks investigations.

#### **14.4.1.2 Trace Dissemination of a Specific Subtype of Pathogen Across Time and Space**

Microbes that cause human disease are constantly emerging and reemerging. In order to prevent and control the spread of infection, we must be capable to trace the origin and source of entry of pathogens into the population. By comparing strains, we can determine if there have been single or multiple points of entry, and if emerging resistance is from multiple spontaneous mutations or from dissemination of a single clone. For example, *Streptococcus pneumoniae* (*S. pneumoniae*), a major human pathogen and one of the most common indications for antibiotic use, results primarily in pneumonia, but also gives rise to meningitis and otitis media. However, resistance to penicillin emerges relatively slowly, once it emerged it was widely disseminated in relatively few clones as defined by multilocus sequence typing (MLST). By contrast, the recent emergence of *S. pneumoniae* resistant to fluoroquinolones has been due to various genetic mutations, suggesting spontaneous appearance after treatment. Because the resistance of *S. pneumoniae* to fluoroquinolones rapidly followed the introduction of fluoroquinolones, alternative antibiotics will be needed in relatively short order to treat *S. pneumoniae* infections.

#### **14.4.1.3 Determine the Origin of an Epidemic**

Molecular tools help us to trace an outbreak or epidemic return to its origin in time, and return to its reservoir in space. Knowing the origin in time is critical to predict future spread, and identification of the reservoir for infection is the key to control disease spread. For instance, the prevalence of methicillin-resistant staphylococcus aureus (MRSA) has been a steady increase in America's hospitals. In 2004, among some intensive care units, the prevalence was as high as 68%. Nevertheless, in the early 2000s, the emerging of new strains of MRSA in population from community could not be traced back to hospitals. Genetic typing of the strains verified that strains isolated from those who had no linkage with hospitals on epidemiology were genotypically different from hospital strains. More recently, community-acquired MRSA has been transmitted into hospitals. When comparing with hospital-acquired

MRSA, community-acquired MRSA has different virulence factors and different patterns of antibiotic resistance so there is a clinical benefit in enabling us to distinguish between the two.

#### **14.4.1.4 Follow the Emergence and Spread of New Infections**

Severe acute respiratory syndrome (SARS) is a new infectious disease emerging firstly this century. Before its identification, coronaviruses were not regarded as primary pathogens in that only 12 known coronaviruses can be able to infect humans or other animals. The identification of SARS resulted in a search for other coronavirus pathogens, and horseshoe bats were identified as the reservoir and civets as the amplification hosts at last. The time from the initial observation to the sequencing of the virus and development of a diagnostic test was 5 months. The story of the rapid isolation, identification, and sequencing of the coronavirus causing SARS is illustrative of the synergistic effects of the combination of molecular methods with epidemiology. This effective combination enables scientists to follow the emergence and spread, and to identify ways to prevent transmission and further introductions of the virus into human populations.

#### **14.4.1.5 Identify Previously Unknown or Uncultivable Infectious Microbes**

The most microbes could not be cultured using standard laboratory techniques. The ability of replication of genetic material and determination of genetic sequence, which can then be compared to known genetic sequence, has brought about a fundamental reevaluation of vast life around, in, and on us. Noncultural techniques have enabled us to describe the microbial communities living in the mouth, vagina, gut, and other body sites, and the body sites thought to be sterile by previous detection, such as the blood. Epidemiological data may suggest an infectious origin for a disease. Previously, if an organism cannot be cultured, it remained only a suggestion. Molecular tools have altered this by the achievement of detecting uncultivable microbes. It is now known that human papillomavirus (HPV) types 16 and 18 can cause cervical and other cancers, and vaccines are licensed to prevent acquisition. HPV 16 was first identified in 1983 before the virus could be grown. When discovering HPV 16, we realize that papillomavirus could give rise to cancers in cows, rabbits, and sheep, but it was unclear whether the HPV can lead to human cancers. HPV was a suspected cause of genital cancer for the similarity to Kaposi's sarcoma, and the epidemiology suggested that an infectious agent was involved. But other genital infections, especially herpes simplex virus, were also suspects. HPV had been excluded by many, but a new molecular technique, the hybridization assay, detected in cancerous tissue a new subtype, HPV 16, which was specifically

associated with cervical and other cancers. The correlation of HPV 16 with cancers was verified by comparing presence of HPV 16 between cancer patients and controls. Notwithstanding this evidence can be very suggestive, it does not differentiate temporal order, because the cancer might happen before the infection of HPV 16. Demonstrating temporal order required large-scale prospective cohort studies. These studies also offered crucial perspectives supporting the possibility that vaccination could protect against HPV because of rare occurrence of reinfection with the same HPV subtype, and antibody could prevent reinfection and persistence of low-grade lesions.

### ***14.4.2 Control and Prevention of Chronic Diseases***

With the development of economics and the implementation of vaccines, most infectious diseases were controlled, while the incidence and mortality of chronic diseases were increased dramatically, such as cancer, cardiovascular disease, and type 2 diabetes mellitus. Molecular epidemiology played an important role in the discovery of the cause, mechanism of pathogenesis, and individual susceptibility.

#### **14.4.2.1 Improving the Understanding of Mechanism of Pathogenesis**

Previously, most cancer epidemiological studies were restricted to evaluating possible causal relationships between two types of events: exposure to potential causative “environmental” agents (cigarette smoking, dietary factors, specific chemicals from workplace, etc.) and disease outcome (i.e., clinical cancers incidence or cancers mortality). However, the specific mechanism was unknown. Increasingly, molecular epidemiological studies are combining panels of biomarkers related to exposure, preclinical effects, and susceptibility using samples of exfoliated cells, blood cells, body fluids, or tissues. These biomarkers are now being widely used in cross-sectional, retrospective, prospective, and nested case-control epidemiological studies, for the purpose of improving our cognition to the causes of specific human cancers. For example, the cotinine in serum or urine represented cigarette smoke exposure, which was a valuable supplement to traditional means of evaluating exposure. Moreover, assays have been implemented to measure “biologically effective dose” of a compound, for example, the amount that has reacted with key cellular macromolecules. The metabolite of cotinine could form the carcinogen-DNA adducts which related with lung cancer. Other molecular epidemiological studies in Chinese populations have prospectively linked DNA damage induced by aflatoxin B1 (AFB1) to liver cancer risk.

#### **14.4.2.2 Evaluating the Susceptibility of Individual and Defining the Risk Population**

Human beings evidently differ from one another in physical characteristics, personality, and other factors. They are also different in genetically determined susceptibility to disease. When we investigate the etiology of a disease, we cannot help asking the question: How much of the incidence of the disease is due to genetic factors, how much is due to environmental factors, and how do these types of factors interact with each other to increase or decrease the risk of disease? Obviously, not everyone who exposed to an environmental risk factor will necessarily develop disease. Even though the relative risk for exposed to a specific factor is very high, the notion of attributable risk implies that not all occurrence of a disease is due only to the specific exposure in question such as the relationship between cigarette smoking and lung cancer. It is demonstrated that lung cancer does not develop in every smoker, and it does develop in someone who does not smoke.

People often accept a fatalistic approach when they are told that a disease is primarily genetic in origin. But even in diseases originate primarily from gene, a good deal of environmental interaction often occurs. For example, phenylketonuria is characterized by a deficiency of phenylalanine hydroxylase for genetic reason; the child who affected cannot metabolize phenylalanine, an essential amino acid, and the excessive phenylalanine accumulation causes irreversible mental retardation. Can we prevent the genetic abnormality? No, we cannot. Can we decrease the likelihood that a child manifest mental retardation because of this genetic abnormality? Yes, we can do so by providing a diet with low phenylalanine to reduce or eliminate the child's exposure to phenylalanine. As shown in this example, we can prevent the adverse effects of a genetic disease by controlling the affected person's environment so that the manifestations are not expressed. Hence, in viewpoints of both public health and clinical medicine, it is crucial that bear in mind the interrelationships between genetic and environmental factors in disease causation and expression.

#### **14.4.3 Conclusions**

Traditional epidemiology has achieved greatly vital goals by means of simple tools such as interviews and questionnaires. Even a difficult issue, for example, the relationship between air pollution and chronic disease, has been successfully disposed by time-series analysis and other means not depended on the laboratory. Hence, it needs to be evaluated carefully for the application of molecular techniques combined with epidemiological designs.

As the examples above demonstrated, molecular epidemiology is not different with conventional epidemiology, but represents an endeavor that commence to achieve specific scientific goals: (1) a better description of exposures, especially when exposure doses are fairly low or different sources of exposure should be



integrated in a single measure; (2) the study of gene-environment interactions; (3) the application of markers of early response, for the purpose of overcoming the main limitations of chronic disease epidemiology, that is, the relatively low frequency of specific forms of disease and the long latency period between exposure and the onset of disease. Also limitations of molecular epidemiology should be acknowledged: the complicacy of various laboratory methods, with scanty knowledge of measurement error or interlaboratory variability; the lacking recognition of some sources of bias and confounding; in some situations, the lower degree of accuracy (such as urinary cotinine compared to questionnaires on smoking habits); and the indefinite biological meaning of markers, like some circumstances of some types of adducts or some early response markers.

# Chapter 15

## Pharmacoepidemiology



Xiaotian Liu and Jian Hou

### Key Points

- Pharmacoepidemiology is the process of the application of the principles and methods of epidemiology to study the uses and effects of drugs in human population.
- The post-marketing pharmacoepidemiology study not only can supplement the information available from pre-marketing studies, but also provide the new types of information that cannot be obtained from pre-marketing studies.
- Pharmacoepidemiology has taken advantage of the principles and methods of epidemiology and developed sophisticated methods to deal with problems in the field.

In the past decades, there was an enormous progress in the medical sciences, which has contributed to developing a great number of new powerful pharmaceuticals to provide better medical care for the patients. However, the new pharmaceuticals caused harm and led to the increase of serious adverse reactions that were unexpected in preclinical studies or premarketing clinical trials occasionally. Therefore, pharmacoepidemiology was developed as a scientific discipline at the interface between clinical pharmacology and epidemiology against this background.

## 15.1 A Brief History and Definition

### 15.1.1 A Brief History of Pharmacoepidemiology

More than 2000 years ago, there was a record of drug poison in Chinese medical literature, “Shen Nong tasted 100 herbs and encountered 72 poisons one day. The

---

X. Liu (✉) · J. Hou (✉)

College of Public Health, Zhengzhou University, Zhengzhou, China

international community really paid attention to the safety of drugs about 70–80 years ago. Since the beginning of the twentieth century, there were already adverse drug events that occurred occasionally, caused illness, disability, and death, and even led to the deformity and death of offspring. In 1935, pharmacists found the effect of sulfanilamide on the antibiosis, then various types of sulfanilamide (such as tablet and capsule) came out one after another. To improve the taste, diethylene glycol was used to replace ethanol as solvent by the pharmacist of the Massengill company of the United States in 1937, then sulfanilamide oral liquid agent was put into market to treat the infectious diseases without the premarketing clinical trials. As a result, a total of 107 people, including more than 30 children, died from renal failure. In response, the US Congress drafted and passed the Food, Drug, and Cosmetic Act in 1938, which stipulated that preclinical toxicity testing was required for both the marketing and clinical trial. In addition, manufacturers needed to collect clinical data on drug safety and submitted the data to the Food and Drug Administration (FDA) before drug marketing. There were 60 days for the FDA to audit and object to an application of marketing, otherwise, it would be proceeded. However, the proof of efficacy was not required in this process.

Until chloramphenicol was found to cause aplastic anemia in the early 1950s, more attention was paid to adverse drug reaction (ADR). In 1952, the American Medical Association (AMA) Council on Pharmacy and Chemistry established the first official registry of ADRs, which was used specifically to collect the information about cases of blood dyscrasias caused by drugs. Then in 1960, the FDA began to collect reports of ADR and sponsored the hospital-based drug monitoring programs for new drugs. In addition, the Johns Hopkins Hospital and the Boston Collaborative Drug Surveillance Program developed the combination application of in-hospital monitor and cohort study to assess the short-term effect of drug used in hospital. The approach was transported to the University of Florida-Shands Teaching Hospital later.

In 1961, the former Federal Republic of Germany and other European countries witnessed the infamous “thalidomide disaster.” Thalidomide was taken to treat pregnancy vomiting in the first 3 months of pregnancy. Shortly its post-marketing, a dramatic increase was observed in the prevalence of phocomelia, which was a previously rare birth defect with the characteristics of the parts, or even absence of limbs, sometimes with the presence of flippers. The epidemiological study was used to establish the cause, and it was discovered that exposure to thalidomide in utero was a risk factor of phocomelia. After the event that shocked the world, the United Kingdom, the United States, and other countries in western Europe all began to strictly check and examine the drug qualification before marketing. The United Kingdom set up the Committee on Safety of Medicines in 1968. These facilitated the establishment of bureau to collect and collate information from the national drug monitoring organizations in the World Health Organization (WHO) in 1970. However, there were still the epidemic of subacute myelo-optic neuropathy (SMON) in Japan in the late 1960s. A collaborative study by Japanese epidemiologists and clinicians confirmed that the SMON was caused by clioquinol taken to prevent traveler’s diarrhea. In 1971, Herbst et al. found that the use of diethylstilbestrol in the

early stages of pregnancy to preserve the fetus could cause vaginal adenocarcinoma in their daughters. By the 1980s, the occurrence of ADR after post-marketing prompted governments and drug administration to strengthen the management of new drugs, which facilitated the development of pharmacoepidemiology.

With the increase in variety and quantity of drugs, the evaluation and management of drug use have become essential. Sweden was the first country to establish the major of clinical pharmacology in 1956. In 1964, WHO fully affirmed the necessity of the clinical pharmacology specialty. After more than 20 years of effort, clinical pharmacology had become a mature profession with the main function to monitor the ADR in developed countries by the 1980s. However, in the early 1980s, the United Kingdom medical profession showed that the existing medicine management methods, clinical pharmacology, and other specialties still could not meet the needs of ensuring the safety of drug users, and then drug surveillance was put forward. Against this background, pharmacoepidemiology was developed as a scientific discipline at the interface between clinical pharmacology and epidemiology. In 1984, the word “pharmacoepidemiology” was first appeared in the *British Medical Journal*, and then well-known by the public.

### ***15.1.2 Definition of Pharmacoepidemiology***

Pharmacoepidemiology is defined as the application of the principles and methods of epidemiology to study the uses and effects of drugs in human population, and to optimize the benefit risk ratio of drugs, vaccines, and medical devices through the development and evaluation of risk management strategies, so as to improve the medical care. The study subjects of pharmacoepidemiology are human populations, the research contents are the distribution of drug use and drug effect in the population, and the aims are to provide information on the safety and efficacy of drug use in population, as well as to form a scientific basis for clinical rational drug use and policy-making. Pharmacoepidemiology studies both beneficial and adverse effects of a drug. Its focus is to assess the risk of uncommon, at times latent, and usually unexpected ADR that occurs for the first time after post-marketing.

### ***15.1.3 Drug-Related Concepts***

#### **1. Drugs**

Drug is defined as a substance used for the prevention, diagnosis, and treatment of the diseases, and to purposefully regulate human physiological functions, with specified indications, usage, and dosage, including traditional Chinese medicinal materials, radioactive drugs, serum, vaccines, blood products, and diagnostic drugs, excluding drugs that are under the premarketing clinical trials.

#### **2. Adverse drug reaction**

Adverse drug reaction (ADR) is defined as a harmful reaction that occurs under normal dosage of the qualified drug and is not related to the purpose of drug use. The definition excludes ADR caused by intentional or accidental overdose or medication errors. ADR has traditionally been separated into Type A reaction, Type B reaction, and Type C reaction according to the pharmacological effect. Type A reaction is the result of an exaggerated pharmacological effect under the normal drug dosage. It is characterized by be predictable, common, dose-related, and can be treated by reducing the dose or cessation of the drug. The common reasons are the recipients receive overdose of a drug, or they cannot metabolize or excrete the drug normally resulting in high level of the drug, or they are sensitive to the drug for some reasons despite normal drug level. In contrast, Type B reaction is an aberrant effect, with the characteristics of be unpredictable, uncommon, not related to dose, and potentially more serious requiring cessation of the drug. Type B reaction represents the major focus of pharmacoepidemiologic study. The reasons of Type B reaction may be attributed to idiosyncratically inherited reactions to the drug or hypersensitivity reactions to the drug or some other mechanisms. Type C reaction generally occurs after long-term use of the drug and is characterized by a long incubation period, unpredictability, and no clear time relationship. The mechanism of Type C reaction is not clear and may be related to the teratogenesis, carcinogenesis, and changes in the cardiovascular system and fibrinolytic system after long-term usage of the drug.

### 3. Adverse drug event

Adverse drug event (ADE) is any adverse clinical event that occurs during the period of drug use, but it is not necessarily causally related to drug use. Although ADE occurs during the period of drug use, the causal relationship between drug use and ADE needs to be verified through investigation and evaluation of pharmacoepidemiology.

## 15.2 Main Research Contents

### 15.2.1 Drug Safety Evaluation

To explore the incidence and risk factors of ADE and ADR as well as to provide scientific basis for drug risk management, to quickly find the adverse reactions to ensure the safety of drug users through the data mining techniques of the *observational* database and the analysis of safety signals, to standardize the monitoring methods of drugs after post-marketing and improve their practicality, and to develop the *flow chart* of establishment of a causal association of ADR.

### ***15.2.2 Drug Effectiveness Evaluation***

Comparative effectiveness research (CER) aims to study the effects of interventions and strategies on prevention, diagnosis, treatment, and health monitoring, and to compare the health-related outcomes of different disease groups through the data mining techniques, then to provide the evidence on which kind of intervention was the safest, easiest, and most effective for the patients, medical staff, consumers, and policy-makers.

### ***15.2.3 Drug Utilization Study***

The definition of drug utilization is that “marketing, distribution, prescription and use of drugs in a society, with special emphasis on the resulting medical, social and economic consequences” by WHO in 1977. Thereby, the aims of drug utilization study are to examine drug utilization, study the effects of drug utilization on the population, identify problems of drug utilization in relation to its importance, causes, and consequences, provide a scientific basis for decision-making, and assess the effects of actions taken. It involves pharmacy, pharmacology, pharmaceutical management, social anthropology, behavior, economics, and other fields.

### ***15.2.4 Pharmacoeconomic Study***

Pharmacoeconomics includes two levels: broad and narrow. In the broad sense, the pharmacoeconomic study is defined as applying the principles, methods, and analytical techniques of economics to study the economic behavior of drug supply and requisitioning parties, drug market price under the interaction between supply and requisitioning parties, as well as various intervention policies and measures in the field of drugs. In the narrow sense, the pharmacoeconomic study is the economic evaluation of drug utilization based on the comprehensive analysis of drug efficacy, safety, and utilization, and provides the theoretical basis for clinical drug use, the prevention and therapy of disease and medical insurance payment decision-making. By collecting and comparing economic data related to drug utilization, pharmacoeconomics is to carry out the cost-effect analysis, cost-benefit analysis, cost-utility analysis, or minimal cost analysis from the cost and benefit considerations.

## 15.3 Aims and Significances

### 15.3.1 *Aims of Pharmacoepidemiology*

According to different purposes, different organizations and individuals decide whether to conduct pharmacoepidemiological study. Generally, one study is conducted for multiple purposes and can be refined in four respects: regulatory, marketing, legal, and clinical needs.

#### (1) Regulatory

(a) Requirements of the pharmaceutical administration. (b) Manufacturers want the drug to be approved for marketing as soon as possible. (c) Answer questions from the pharmaceutical administration. (d) Producers want to the drug apply for marketing in other countries.

#### (2) Marketing

(a) Assist in entering and occupying markets by verifying the safety of medicines. (b) Raise the profile of the drug. (c) Assist in repositioning of marketing. For example, adopt different outcomes, such as life quality evaluation and economic evaluation; for different patients, such as children or the elderly; discovery of new therapeutic indications; or to reduce the restrictions on drug labels. (d) Protect developed and tested drugs from adverse reactions.

#### (3) Legal demands

Prepare for possible lawsuits of drug liability.

#### (4) The clinical need

1. To generate hypotheses. Whether or not the hypotheses need to be generated depends on the following factors: is it a new chemical monomer? Safety of similar drugs, relative safety of the drug among similar drugs, and drug formulation.
2. To test hypotheses. Conducting hypotheses testing is to solve the following problems: problems based on drug structure, questions raised by preclinical animal tests or premarket human studies, questions raised by voluntary reports on adverse drug reactions, and to better quantify the frequency of adverse reactions.
3. The disease to be cured. In addition, the characteristics of disease to be cured, such as course, prevalence, severity, availability of alternative therapies, and so on, determine whether pharmacoepidemiology study should be conducted [11].

## ***15.3.2 Significances of pharmacoepidemiology***

### **15.3.2.1 Improve the Quality of Premarketing Clinical Trials**

To ensure the efficacy and safety, new drug must undergo human clinical trials no matter how many *vitro* and animal trials each drug has undergone before premarketing. Premarketing clinical trial of new drugs is one of the main types of experimental epidemiology. Mastering the theoretical bases and methods of epidemiology can help to design clinical trials in a standardized way, collect and analyze the experimental data, identify and control bias, thereby improve the quality of premarket clinical trials.

### **15.3.2.2 Post-Marketing Study of Drug**

Premarketing study of drug effect is necessarily limited in size, time, and sample (e.g., elderly, pregnant women, and children were not included). Moreover, the disease or drug used is single during clinical trial, some low incidence of adverse reactions, delayed reactions or drug interactions caused by the combination use of a variety of drug, are difficult to find. Thus, the nonexperimental epidemiological study is needed to evaluate the effect of drug administered as part of ongoing medical care after marketing. Apart from verifying the results information in the clinical trials of pre-marketing, the post-marketing pharmacoepidemiology study can not only supplement the information available from premarketing studies, but also provide the new types of information that cannot be obtained from premarketing studies.

The potential contributions of pharmacoepidemiological study are as follows:

1. Supplement information available from premarketing studies (to better quantify the incidence of known adverse and beneficial effects).
  - ① Study the incidence of adverse reactions or the frequency of effective effects during the prevention or treatment under use of drugs through the epidemiological survey in a large number of population.
  - ② Understand the effects of drugs on special groups, such as the elderly, pregnant women, and children.
  - ③ Explore the modified effects by other drugs and other illnesses.
  - ④ Evaluate whether the new drug is better than other drugs used for the same indications.
  - ⑤ Evaluate the drug safety, effectiveness, quality standard, etc.
2. Provide the new types of information that cannot be obtained from premarketing studies.
  - ① Discover previously undetected delayed adverse and beneficial effects, and verify them by epidemiological methods and reasoning.
  - ② Study the characteristics and influencing factors of drug utilization.
  - ③ Assess the effects of drug overdoses.
  - ④ Evaluation of the economic benefits of drug utilization.

In addition, pharmacoepidemiology study is also conducted to fulfill the ethical and legal obligations as the general contributions [11].



## 15.4 Methods of Pharmacoepidemiology

Pharmacoepidemiology is a scientific discipline at the interface between clinical pharmacology and epidemiology. Therefore, the methodological approaches in epidemiology can be used to generate and test hypotheses on drug risks or benefits according to the purpose of the study (Fig. 15.1). Thus, both the primary study (such as descriptive study, analytical study, and experimental study) and secondary study (e.g., systematic review and meta-analysis) can be applied in the pharmacoepidemiology study. Meanwhile, the multiple epidemiological methods might be conducted to explore the associations between the drug and ADR/ADE in one study, especially studying in the post-marketing monitoring and major drug harm events.

### 15.4.1 Case Report and Case Series Study

Case report is the simplest report of events observed in single patient. As used in pharmacoepidemiology, a case report describes a single patient who was exposed to a drug and experienced the effect, especially an adverse outcome. For example, a published case report about a young woman suffered a pulmonary embolism who was taking oral contraceptives. Case report could be used to generate hypotheses about ADR, but more rigorous study designs are needed to test the hypotheses. Nevertheless, unreliable conclusion might be resulted in the information bias by patients or medical staff in case report. After drug marketing, case series study is the most useful for two related purposes. Firstly, it can be used to explore the incidence of ADR/ADE. Secondly, it can be beneficial to finding some special or delayed adverse reactions. However, causal correlation cannot be inferred due to the absence of control group. Thus, case series study is useful in providing clinical descriptions of a disease or patients who receive an exposure, but not in determining causation.

### 15.4.2 Ecological Study

Ecological study is used to explore the relationship between drug use (exposure) and outcome (both in terms of beneficial and adverse effects) in different populations. Ecological study includes two types: ecological comparative study and ecological trend study. The study on the association between thalidomide and phocomelia is a typical ecological trend study. The sales curve of thalidomide was consistent with the incidence of phocomelia, and there was about one pregnancy interval between them. All these proofs suggested that thalidomide was the cause of phocomelia. However, the results of ecological study should be carefully discussed to avoid ecological fallacies. For lack of data of individuals, only group data is utilized in the

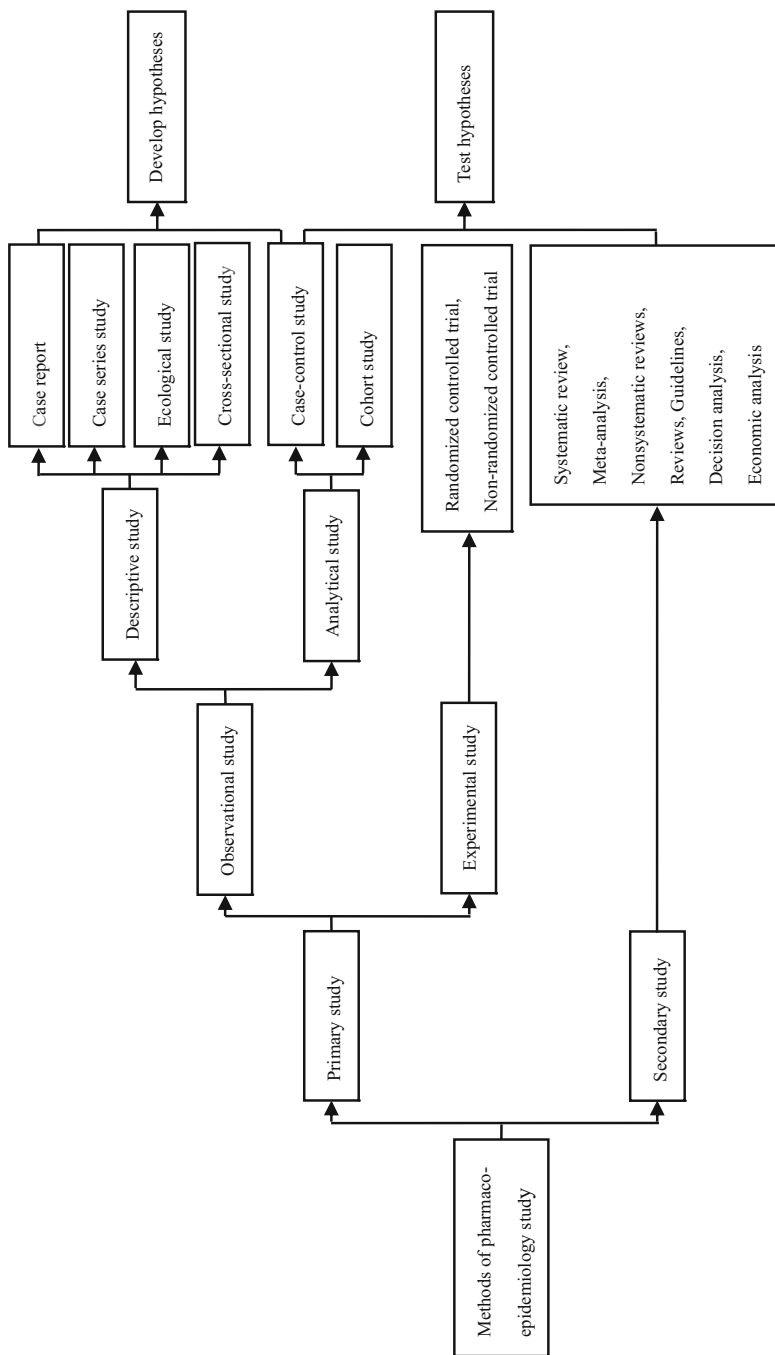


Fig. 15.1 Hierarchy of pharmacoepidemiological study design

ecological study. Moreover, the confounding variables could not be adjusted in this study. Therefore, the ecological study is unable to distinguish which factor is the real cause among exposures coinciding with the outcome.

### ***15.4.3 Cross-Sectional Study***

Cross-sectional study is to describe the distribution of drug use-related states or events among a specific group of population at a specific time and explore the influencing factors of ADR/ADE, which would provide clues and scientific bases for further disease study and the rational drug use.

### ***15.4.4 Case-Control Study***

Case-control study is to look for the exposure differences between cases with a disease of interest and controls without the disease in antecedent exposures. This design can be extremely useful when the disease is relatively rare, or when time or resources are limited. As a classic example, information was collected from only eight cases with vaginal adenocarcinoma, and each case was matched with four patients without vaginal adenocarcinoma. The information on the cases, controls, and their mothers were collected. Through comparing the data between case and control groups, it was found that the use of diethylstilbestrol to preserve the fetus in the early stages of pregnancy caused vaginal adenocarcinoma in their daughters. Nevertheless, the case-control study generally collects information on exposures retrospectively by referring to medical records or by questionnaires or interviews. Therefore, the exposure information retrospectively collected is one of limitations in the case-control study. As such, the proper selection of controls is a crucial task to reduce selection bias. If the case-control study is done well, the subsequent cohort study or randomized clinical trial (RCT) will generally verify the results of the case-control study.

### ***15.4.5 Cohort Study***

Cohort study is used mainly to test etiological hypotheses. It generally is used to compare the incidence of certain outcome (both beneficial and adverse effects) between the drug (exposed) group and no drug (unexposed) group. Cohort study might be performed either prospectively or retrospectively. In prospective cohort study, the participants are divided into two groups according to whether they take drugs at baseline. The outcomes are not available at the beginning of the cohort, and should be collected during the follow-up among a period of time. For example, two

groups of women of childbearing age who took oral contraceptives or other contraceptives were followed up to collect and compare the incidence of venous thrombosis. However, the main disadvantage of the prospective cohort study is that it needs a relatively long time until a sufficiently large number of events occur. For rare outcomes or delayed drug effects, the follow-up period may span one, or even several decades. In this circumstance, the prospective cohort might be not suitable, especially when some drug might be harmful which contraries to the ethics. In the retrospective cohort study, the outcomes under study had already occurred, and the exposure history of drug is obtained using medical records, questionnaires, or interviews. Although there is a long period between drug use and outcome, the collection and analysis of data can be completed in a short period. Moreover, there is no ethical issues. Thus, the retrospective cohort study may be a better choice for ADR study. Significantly, the information on the exposure history of drug and outcome should be complete and reliable.

Cohort study is useful in post-marketing drug surveillance study, which looks at any possible effect of a newly marketed drug. For example, cimetidine was marketed in 1976, and post-marketing monitoring began in the United Kingdom since 1978. During the follow-up period, there were 9928 patients using cimetidine and 9351 controls without using cimetidine with complete hospitalization and death records in four regions. With the improvement of drug post-marketing monitoring and database sharing, a “computerized” cohort study will play a significant role in ADR study.

The exposure is determined before the outcome occurs in cohort study. Therefore, the causal association is more convincing compared with case-control study. However, there is still confounding bias because there is no good comparability between the drug group and no drug group. Thus, the stratified or logistic regression analysis are necessary to control confounding bias in cohort study.

#### ***15.4.6 Experimental Study***

Experimental study, especially randomized controlled trial (RCT), is the gold standard for evaluating the efficacy of drug. Nevertheless, it cannot be used to verify the causal association in pharmacoepidemiology due to ethical issues. Sometimes, under certain conditions, the reverse verification of causal association can be conducted in the population. Reverse validation based on experimental study in population is that remove the hypothesized etiology (drug), and follow-up to observe the trend of incidence of related outcomes of drug use. For example, in 1982, the Ministry of Health in China eliminated a batch of drugs (containing tetramisole). Then, the number of encephalitis syndrome in Wenzhou city decreased in 1983–1984 after the elimination of these drugs.

### 15.4.7 *Systematic Review and Meta-Analysis*

Secondary use of published data has become an important means of active monitoring of drug safety in many countries. Systematic review and meta-analysis have been widely used in the medical research in the past 20 years, especially when there are doubts about the efficacy or safety of drugs, and when there are few studies with large samples.

### 15.4.8 *Real-world Study*

With the widespread use of computers, mobile devices, wearables, and other biosensors to gather and store huge amounts of health-related data, real-world study has been paid more and more attention to the real-world data evidence. Under the real medical conditions, the clinical effects between drugs and drugs, vaccines and vaccines, surgical treatment and drug treatment, inpatient and outpatient treatments, etc. can be observed and compared in real-world study in which no other intervention factors are added other than the test factor. See Chap. 19 for details.

### 15.4.9 *Newly Derived Study Design*

1. *Derivative study of case-control study*: Nested case-control study, case-cohort study, case-crossover design (to explore the effect of short drug use on the acute adverse events), and case-time-control design (to solve the effect of drug use on outcome when there are mixed indications caused by the disease severity or time of taking medicine changes).
2. *Pharmacogenetics and pharmacogenomics*: They aim to explore the relationship between genetic factors and drugs at the molecular level, and estimate the effect of genetic factors on drug efficacy. They can not only promote the development of new drugs, but also provide scientific bases for effective individualized treatment plan in the clinical practice.
3. *Propensity score and instrument variable*: To solve the incompatibility between exposure group and unexposed group in observational study, especially in real-world study, statistical techniques (such as propensity score and instrument variable) are developed to adjust for confounding factors in pharmacoepidemiology.

## 15.5 Data Collection and Analysis

### 15.5.1 Data Collection

Data collection in the pharmacoepidemiology is mainly divided into primary data collection, existing (secondary) data collection, as well as the combination of two types of data collection. The sources of data mainly include routine data, literature data, and surveillance system of ADR.

#### 15.5.1.1 Routine Data

##### 1. Vital Statistics

(1) Demographic data: Demographic data can be obtained through population census, sampling survey, and household registration system. These data are mainly used for ① the denominator, used to calculate some relative numbers, such as per capita consumption of drugs, drug expenditure, the proportion of drug users in the population, etc.; ② the standardized rate or the standard population composition, so as to compare the results of pharmacoepidemiology among different regions; ③ demographic characteristics, such as age and gender, are important factors that influence drug use in quantitative study on the relationship between these factors and drug use.

(2) Death data. The death data provides the distribution of the causes, sex, age and other information of deaths. Exploring the relationship between mortality and drug use or sales could provide clues for the future study.

(3) Disease data: The disease data is an important source of data frequently used in pharmacoepidemiological study. Such data can be obtained from literature published by medical institutions or professional prevention and treatment institutions. The main applications of disease data include the following: ① to provide the background data, ② to evaluate the effect of drug, ③ to provide clues for the future study, and ④ to assess the causal association between drug use and outcome (both beneficial and adverse effects).

##### 2. Data collected by relevant agencies

This data comes from medical and drug administration and institutions for academic research, such as FDA and State Administration of Traditional Chinese Medicine collect and preserve data on the pharmaceutical production and sales, customs preserve data of import and export of pharmaceutical products, National Center for ADR Monitoring has data of ADR monitoring, the medical insurance agencies have data of prevalence, medication use, expenses, and other related information of insured population.

##### 3. Data from pharmaceutical companies and manufacturers

Pharmaceutical manufacturers generally obtain materials related to their own products through the accumulation of daily working materials, special investigation and literature search, etc. Pharmaceutical manufacturers often have the data

of drug purchase and sale. The integrated pharmaceutical data is of great significance in the study of pharmacoepidemiology. However, the protection of commercial intelligence makes it very difficult to obtain such data.

#### 4. Hospital information

Because hospitals mainly carry out the diagnosis and treatment of disease, most of the data obtained from hospitals can be used for pharmacoepidemiology study. The data mainly includes drug warehousing records, prescription, outpatient and inpatient medical records, and drug expenses. But the data of hospital does not represent the entire population, and the information of each hospital has its own characteristics. All these circumstances may lead to the selection bias. Simultaneously, when analyzing data, more attention should be paid to the hospital level, nursing quality, hospital facilities, medical expenses, and so on.

### 15.5.1.2 The Literature

The literature in medical journals play an important role in discovering ADR and preliminarily assessing the causal relationship. Published literatures are important sources of systematic review and meta-analysis. Based on mathematical and statistical methods, bibliometrics analysis is used to quantitatively analyze the data of literature.

### 15.5.1.3 ADR Monitoring and Reporting System

ADR monitoring and reporting *system* refers to the process of discovery, reporting, evaluation, and control of ADR, with the purpose to effectively control ADR, to prevent the occurrence of ADE, as well as to ensure the safety of drug use. The common methods of monitoring ADR in the world include spontaneous reporting system (SRS), intensive hospital monitoring, intensive medicines monitoring, and expedited reporting. The main international monitoring agencies for ADR include the Uppsala Monitoring Centre (UMC), International Society of Pharmacovigilance (ISOP), the US Food and Drug Administration (FDA), and others, such as the Council for International Organization of Medical Sciences (CIOMS), the European Medicines Agency (EMA), and National Center for ADR Monitoring, China.

## 15.5.2 Data Processing and Analysis

After data sorting, checking and processing of missing data in pharmacoepidemiology study, appropriate statistical analysis method should be adopted for data analysis according to different study designs, research purposes,

and data types. However, data analysis methods of pharmacoepidemiology study also have their own characteristics.

### 15.5.2.1 Mining and Analysis of ADR Monitoring Database

The measure of disproportionality is used to analyze ADR monitoring data. This method is based on the classical  $2 \times 2$  fourfold table (Table 15.1). The basic idea is to estimate the ratio of the actual amounts of ADR related to a certain drug in SRS to the expected amounts or the number of adverse reactions caused by other drugs. For example, the center for pharmacovigilance in the Netherlands uses the ratio as the reporting odd ratio (ROR), which is calculated as  $ROR = AD/BC$ . If the ratio is large enough (“out of balance”), there is likely to be certain association between the suspected drug and the suspected ADR, and the association is not caused by an opportunistic factor or “noisy background” of the monitoring database.

### 15.5.2.2 Mining and Analysis of Prescription Database

The prescription database is also a resource that can be fully mined and analyzed. Prescription sequence analysis (PSA) is a method to monitor ADR based on reliable and complete drug prescription records. When the adverse reaction of one certain drug is the therapeutic indicator of other drugs, the prescription record will show a specific sequence of drug use, and a specific frequency distribution in a large prescription database. For example, drug A and drug B, drug A is the original prescribed drug. If drug A leads to some adverse reactions which require drug B to treat. In this way, the frequency distribution of the two drugs in the prescription database will change.

Prescription sequence symmetry analysis (PSSA) develops based on PSA. The method is to assess whether a drug is associated with an event by evaluating the symmetry of the event distribution before and after taking a specific drug. For example, drug A may cause some adverse reactions which need to be treated by drug B. Firstly, if there is no causal correlation, the patients who took drug A and B are equally ranked in the database within a certain period of time, in other words, the number of patients who first took drug A and then took drug B is the same as the number of patients who first took drug B and then took drug A. However, if drug A really can cause adverse reactions requiring drug B to treat, then the prescription of drug A will lead to an increase in drug B, which will result in an asymmetric sequence distribution.

**Table 15.1** Measure of disproportionality

	Suspected event	Other events
Suspected drug	A	B
Other drugs	C	D

Shen and Qi [12]



In summary, pharmacoepidemiology is still a relatively new scientific discipline although there has been tremendous progress in the development of the methods. Pharmacoepidemiology has taken advantage of the principles and methods of epidemiology to study the drug safety and effectiveness in human population. It is of great interest to explore why the reactions to the same drug use vary from individual to individual. The genome study will be greatly improved with the development of the pharmacogenomics and molecular biology. Besides, new study designs and statistical approaches may emerge as consequences of these developments.

# Chapter 16

## Evidence-Based Medicine and Systematic Review



Qi Gao and Huiping Zhu

### Key Points

- EBM is conscientious, explicit and judicious use of current best evidence in decision-making of the care of individual patients.
- Systematic review aims to answer a specific research question through retrieving the relevant evidence satisfying the pre-defined eligibility criteria.
- Meta-analysis is a quantitative analysis that combines the results of two or more studies on a given research issue.

## 16.1 Evidence-Based Medicine

### 16.1.1 Concept

Evidence-based medicine (EBM) is a clinical discipline that bridges the gap between research and clinical practice. EBM is dedicated to make decision-making more objective and structured by better reflecting the evidence from researches, especially from studies on clinical epidemiology. It facilitates a transformation of clinical practice and medical education by introducing the research evidence in clinical decision-making. Epidemiologists and clinicians have been intensively involved in developing evidence-based practice. It has been implemented in almost all fields of medicine including general practice, pathology, surgery, pharmacotherapy, dentistry, and nursing.

EBM is defined as “the conscientious, explicit and judicious use of current best evidence in decision-making of the care of individual patients.” As its application expanded from individual patients to health-care services and health professions,

---

Q. Gao (✉) · H. Zhu (✉)  
School of Public Health, Capital Medical University, Beijing, China  
e-mail: [gaoqi@ccmu.edu.cn](mailto:gaoqi@ccmu.edu.cn); [zhuhuiping@ccmu.edu.cn](mailto:zhuhuiping@ccmu.edu.cn)

EBM is also named as evidence-based health care (EBHC) or evidence-informed health care (EIH) or evidence-based practice (EBP). EBM involves two essential principles: (1) it delineates that scientific evidence alone is not sufficient in making a clinical decision, and decision-makers need to take into account the patient's values when assessing the benefits and risks of any treatment strategy; (2) EBM displays a hierarchy of evidence to guide clinical decision-making. Whereas, the hierarchy is not absolute, and should reflect what different levels of evidence refers to, and describe the context and agents.

EBM suggests that a valid set of rules can be a complement to medical training and common sense for clinicians to interpret the results of clinical research correctly. EBM requires careful, systematical, and specific application of knowledge acquired through the combination of individual clinical expertise with the best available external evidence obtained from systematic research. Clinicians should be clearly aware of the systematic evidence and make pragmatic and ethical decisions in terms of patient care. These bring about a question of whether evidence can only be taken as proof, say, from randomized clinical trials (RCT), conducted by a particular population in a particular country, or from a particular magazine. Actually, the evidence is obtained not only from the evidence derived from RCTs conducted by academic medical centers or publications from two or three admitted top journals, but as well as a large amount of evidence, including the patient's preference for treatment and acceptable resources. This contradicts the question of making a decision just by means of proof. EBM underlines the need to actively seek out all valid and relevant data and to continually evaluate such data to ensure its accuracy and applicability.

EBP is decision-making process by which someone makes clinical decisions using the best available research evidence, his/her clinical expertise and patient preferences, in the context of available resources. The concept of EBM has been increasingly accepted in the field of health care. There is no doubt that the practice of EBM by using the best available scientific evidence in medical research can solve the specific problems in clinical practice. EBM provides a sound scientific basis so as to achieve efficiency, consistency, high quality, and safety in medical care. Correspondingly, EBM experts set a focus on a clinical guideline; retrieve, evaluate, and synthesize the evidence; summarize the benefits and risks, and determine the fitness of interventions.

### ***16.1.2 Development of EBM***

The term "evidence-based medicine" has existed for a long time. Clinicians who received formal medical education make decisions during clinical practices based on the clinical features of patients, combined with their clinical expertise. Thus, to some extent, the clinical procedures for diagnosis and treatment are certainly evidence-based, although there may be some shortages when adopting the latest and best

evidence. The present clinical decision-making process adopted by clinicians should not be considered as “empirical medicine” simply.

In essence, clinical medicine is a branch of practical science, and constantly develops along with the development of natural science and clinical science. Thus, it is necessary for the clinicians to continually update their knowledge, learn, master, and apply advanced skills and theories to guide their clinical practices in order to improve their clinical work. Just as Dr. Sydney Burwell, dean of Harvard Medical School put it: “Half of what you are taught as medical students will in 10 years have been shown to be wrong. Moreover, the trouble is, none of your teachers knows which half.” That also explains the importance of constant learning and knowledge updating.

Since the late 1970s, with the increasing development of clinical epidemiology and advanced clinical research methodology, and the emphasis on scientific design, measurement, and evaluation, clinical research has improved dramatically, which produced abundant high-quality clinical outcomes and elicited a series of methods and standards of critical appraisal. All of these have been accepted and applied in the international medical field, which greatly improve the development of clinical medicine and the practice of EBM.

In the early 1980s in McMaster University – one of the places where the clinical epidemiology originated, Dr. David Sackett and his colleagues from the department of clinical epidemiology and internal medicine held a workshop on “How to read clinical literature” for young resident doctors. The residents incorporated the clinical problems of patients, medical literature retrieval and evaluation, and application of the current best evidences being generated by medical researchers worldwide into their clinical practices. They received the training of EBM and achieved great success on the basis of studying the principle and methods of clinical epidemiology. Since 1992, *JAMA* and other journals published a series of review articles, in which Sackett et al. named the new method as “evidence-based medicine.” Afterward, initiated by Haynes and Sackett, the American College of Physicians established a journal club, namely, ACPJC. In order to enhance the development of EBM, experts in clinical epidemiology and clinical medicine selectively and systematically analyzed and evaluated the articles that were published in 30 famous medical journals around the world since 1991. They refined the selected articles, rewrote the abstracts and comments, and then published them in *Annals of Internal Medicine* as supplements. All these were recommended to clinicians so as to use the best evidence during the practice of evidence-based medicine. In 1995, Sackett in Oxford established the Evidence-Based Medicine Center of the United Kingdom. Sackett and his colleagues successively published EBM monographs, and some EBM journals which are cohosted by BMJ and American College of Physicians. What is more, the Cochrane Collaboration was established in 1993. Cochrane Collaboration extensively collects the study results of RCT, and then conducts systematic review or meta-analysis based on a strict evaluation of the quality of the RCTs. It also recommends valuable study findings to clinicians and other professional practitioners to help them to practice EBM.

In 1996, the Evidence-Based Medicine Center of China and the Chinese Cochrane Centre were established under the support of the Ministry of Health of China. The two organizations provide training programs for medical professionals, develop extensive national and international cooperation, and publish two journals on evidence-based medicine. In addition, monographs on EBM and national evidence-based medicine teaching materials have been published. All of these accelerated the practice of clinical medicine and preventive medicine, and improved the quality of medical care in China.

### ***16.1.3 Categories of EBM Practice***

EBM practitioners can be grouped into two types: one is the producer of the best evidence, and the other is the user of the best evidence. Producers of the best evidence include clinical epidemiologists, clinical experts, health statisticians, medical sociologists, and scientific medical information workers. They collect, analyze, evaluate, and integrate the best evidence from more than two million articles of biomedical literature around the world. They aim to provide evidence for clinicians to practice EBM. For the time being, the best resources of clinical evidence are Clinical Evidence, ACPJC, EBMJ, and Cochrane Library, which are published by the BMJ. Evidence producers are the important components of EBM, and EBM practice would not be processed without their hard work.

These experts will not finish their work until they push this best evidence to be applied to EBM practice. They dedicated to provide EBM education for medical students and clinicians. The only way to achieve the real purpose of EBM is to transform the best research evidence into health prevention and treatment services for patients at the highest level, and to enable the clinicians to learn and apply these theories and methods of EBM practice.

Users of the best evidence are the medical personnel engaged in clinical medicine, including policy-makers. In order to make a diagnosis or treatment decision for patients, or to make health management and policy decisions, they should consider their practical issues, and seek, identify, understand, and apply the best of the latest scientific evidence.

Both the producers and users should not only have clinical expertise, but also possess the knowledge of the related subjects. The difference between them is simply based on the level of the requirements. Of course, evidence producer can also be an evidence user, meanwhile, an evidence user can also become an evidence producer.

### ***16.1.4 Procedures of Practicing EBM***

The practice of EBM is composed of five steps which can broadly be categorized as follows:

Step 1: Translate the indetermination (causation, diagnosis, therapy, prognosis, prevention, etc.) into an answerable question.

What a doctor needs to do first is to identify the problem of his/her patients. It is necessary for him/her to deal with the problem using different useful knowledge. Doctors could track down published evidence and contemporaneous research review as the basis for clinical decisions. The primary task is transforming clinical problems into questions. Without an answerable question, no more exploration and research can be done. Also, a good question can help clinicians to make a good strategy in collecting evidence to resolve clinical problems. However, since the practitioners of EBM are highly qualified clinicians who have varying degrees of expertise and backgrounds, clinical questions should be different in clinical practice. And even when different doctors face the same patient, the questions they raise will be different as well.

#### **“PICOS” Model**

“PICOS” model is usually used in building a specific clinical question. “P” means patients or population; “I” refers to intervention or exposure; “C” means the control group or other interventions; “O” refers to the outcome; and “S” means study type.

#### **Patients or Population**

A clinical question must be used to identify a problem of patient. When defining the “P,” it is essential to ask the important characteristics of the patient, including (a) primary problem; (b) patient’s main health concern; (c) health status; (d) age, sex, and race; and (e) current medications.

For example, ascites with or without infection is an important clinical problem for a patient with liver cirrhosis. If it is not determined whether the condition is complicated with spontaneous bacterial peritonitis, then a timely and appropriate medical treatment cannot be performed.

#### **Intervention or Exposure**

The second step in the PICO model is to identify “I.” What intervention that the doctor chooses for the patient is the main consideration. The interventions include the use of diagnostic test, treatment, medication, and so on.

And how to make an appropriate intervention? There are plenty of factors affecting the impact of intervention, including exposures, etiology, treatment, prognostic factors, the patient’s understanding, and compliance. For instance, when treating a patient with peptic ulcer, doctors must consider the cause firstly, since the treatment plan varies depending on whether the patient’s stomach ulcer is due to

H. pylori (HP) infection, or due to the use of nonsteroidal anti-inflammatory drugs, or due to stress.

### **Control**

The third step is to identify the “C.” Mostly, there will be a control, which is the contrast used to compare with the intervention. The control is the only optional component in a “well-built” question. Then, how to make a choice? For example, cancer can be treated by surgery, chemotherapy, radiation therapy, or other types of therapy like intervention. The choice of therapy should be considered according to the disease condition and the economic status of patients, as well as the views of their family members.

### **Outcome**

The last step is to identify the outcome that a doctor desires to achieve. The outcome should be measurable. Outcomes can be defined as an improved sign of symptom or function, survival, mortality, and disability. Different “types” of outcomes refer to the different clinical questions.

### **Study Type**

What is the best study design to find the evidence to answer a clinical question? Systematic review of double-blind, randomized controlled trials, cohort studies, case-control studies, or case series? Which is it depending on what type of clinical question that a doctor is asking.

Step 2: Systematically retrieving the available best evidence to answer the clinical question

### **Quality of Evidence**

Users of clinical evidence need to know how much confidence they can place in the evidences. EBM classifies clinical evidence as several types and rates them in order from the strongest to the weakest levels according to the strength of their freedom from the various biases that exist in medical research (Fig. 16.1). For instance, the systematic review of randomized, placebo-controlled, triple-blind trials with allocation concealment and complete follow-up involving a homogeneous patient population and medical condition provides the strongest evidence for therapeutic interventions. By contrast, expert opinions and case reports have little value due to the biases inherent in observation and reporting of cases, the placebo effect, etc.

### **The 5S Model**

The first “S” refers to original studies which is at the bottom of the “5S” model; the second “S” syntheses include systematic reviews and meta-analysis; synopses (brief comments of original research articles and reviews) is at the third level; then summaries (concise descriptions of an individual study or a systematic review) is

**Fig. 16.1** Levels of evidence



at the fourth level; and the top of “5S” model, systems like computer decision-making system which links individualized patient characteristics to the current evidence. Original studies, syntheses, and synopses often evaluate one aspect of health-care problems, but summaries integrate the best evidence available from the lower layers, and form a complete chain of evidence relating management options for a given health issue. Summaries can be made universally available, and is more feasible to keep up with the latest evidence.

Step 3: Critical appraisal of evidence for its validity that can be categorized into the following aspects:

Systematic mistakes stem from information bias, selection bias, and confounding variables;

Aspects of diagnosis and treatment that are quantitative;

Scale of the effect;

Clinical importance of the result;

External validity or generalizability.

Step 4: Application of the critically appraised evidence into clinical practice while taking into account of doctors’ clinical expertise, the patient’s unique biology, and values.

Step 5: Evaluation of performance (effectiveness and efficiency in taking 1–4 steps and seeking ways to improve for the next time)

## 16.2 Systematic Review and Meta-Analysis

### 16.2.1 Systematic Review

Systematic review is a method to synthesize literatures. It aims to answer a specific research question by collecting all relevant evidence that satisfies the predefined



criteria. It begins with a specific question and predefined criteria for studies, and then uses systematic and reproducible methods to search and select eligible studies. All studies searched are evaluated for possible bias before they are synthesized. The entire process of systematic review is clear and can be repeatedly performed.

Although systematic review can provide an array of information in a given area, it is limited by the quality of original literatures, the method of conducting systematic review, and the reviewers' levels of background knowledge. Therefore, it is necessary to assess the reliability of systematic reviews before applying them into practice.

### **16.2.1.1 Cochrane Systematic Review**

A Cochrane systematic review is done by reviewers from the Cochrane Collaboration according to the standard Cochrane Handbook under the guidance of Cochrane review groups. Cochrane review is of higher quality since it adheres to a strict process and quality control system. Cochrane reviews are updated every 2 years or when new evidence becomes available. Thus, Cochrane review is considered to be the single best source of evidence of effectiveness of interventions. For now, Cochrane systematic reviews focus on the prevention and treatment of diseases.

### **16.2.1.2 Importance of Systematic Review**

#### Meeting the Challenge of Increasing Information Supply in Medicine

Researchers, clinicians, or decision-makers need a large amount of information to make scientific decisions; however, there are too many messages for them. Every year, over two million medicine-related literatures are published in more than 20,000 biomedical journals, and the growth rate is 6.7% annually. A clinician has to read 19 professional papers every day to keep up with the latest development in their fields. Systematic review can select the essential information and discard the dross using strict selection and evaluation method. Also, it combines the true and reliable information with clinical application value, which can provide scientific basis for decision-making.

#### Timely Transforming and Applying Study Results

Evaluation of the treatment of malignant tumor, cardiovascular and cerebrovascular diseases, and various other chronic diseases needs to conduct large-sample clinical trials as far as possible, especially RCT. However, large-scale RCT will take a lot of manpower and time, material resources and financial resources, which is usually beyond the capacity of an institution. Although there are numerous clinical studies now, most of them do have small sample sizes, and the results of them are unreliable. Hence, using systematic review to synthesize results of several homogeneous

clinical trials which are of higher quality can produce relatively more reliable results which can be applied for clinical practice and decision-making.

### **16.2.1.3 The Difference Between Systematic Review and Traditional Review**

Systematic review differs from traditional review in several ways. Traditional review usually tends to be descriptive, and they are liable to bias. Systematic review, on the other hand, needs a comprehensive protocol and search strategy to obtain all eligible studies on a specific topic. Also, systematic review often includes a meta-analysis, and can be updated continuously when new research becomes available, but this is not the case of traditional review.

### **16.2.1.4 How to Do a Systematic Review?**

Systematic review can assess and synthesize several clinical studies with contradictory results in a strict way so as to resolve disputes and make a more reliable conclusion, which provides appropriate guide for clinical practice and decision-making. Nevertheless, the poor quality of included primary studies or inappropriate methods of conducting systematic reviews can introduce bias in the review. Therefore, the methods and process of conducting a systematic review are crucial to ensure the accuracy and reliability.

Systematic review is just a research method and is not limited to RCT or evaluating the efficacy of interventions. A systematic review can be done to investigate causes, diagnosis, treatment, prognosis, or health economics of disease, and it can be divided into systematic review of controlled trials and of observational study according to different study design of the included primary studies. In addition, systematic reviews can be qualitative or quantitative according to whether reviewers have used statistical method (meta-analysis) to analyze data.

#### **Determining a Title and Formulating a Protocol**

A systematic review usually stems from important but controversial clinical questions about the treatment and prevention of diseases encountered in clinical practice. For instance, whether using low doses of aspirin among high-risk population can prevent the occurrence of cardiovascular disease? What are the differences on the efficacy and safety between early laparoscopic cholecystectomy (within 7 days after the onset) and delayed laparoscopic cholecystectomy (6 weeks after hospitalization) for patients with acute cholecystitis?

To avoid duplicate work, it is necessary to conduct a comprehensive search first to find out whether there is an existing or ongoing systematic review or meta-analysis addressing the same clinical issue. If yes, then what is the quality of the

review? If the existing systematic review is out of date or poor quality, then it is useful to update the existing review or conduct a new one.

Studies included in a systematic review should have similar design and interventions within the similar population. Therefore, when determining the title, four factors have to be confirmed around the research question: (1) participants, the type of disease, diagnostic criteria, characteristics, and locations of the study; (2) intervention and comparator of the study; (3) the key findings (primary and secondary results) and serious adverse effects; and (4) study design. These factors are of great importance to guide the search, screening, and assessment of each study, to collect and analyze data, and to explain the application value of results.

After determining the title, a protocol should be developed that includes title, background, objectives, methods and literature search strategy, eligibility for inclusion criteria, assessment of bias in included studies, as well as methods to collect and analyze data.

Overall, the objective and the clinical question about a systematic review has to be determined before protocol formulating and literatures collecting, which can prevent the reviewer from manipulating the title and contents according to the collected data and results in analysis.

### Retrieving Literatures

Systematic and comprehensive collection of all relevant evidence is one of the distinctions between traditional literature review and systematic review. Reviewers need to use various ways and systematic searching methods to avoid publication and language bias, which is based on the established search strategy in the project plan. Both published papers and unpublished materials in several languages such as graduation thesis, academic reports should be collected. For those published papers, professional evaluation groups from the Cochrane Central Register of Controlled Trials and the Register of Controlled Trials use computer search and manual search, which can not only compensate the shortcoming of searching tool that is unable to completely label RCT, but also help systematic reviewers to retrieve relevant original literatures in a rapid and comprehensive way.

### Selecting Literatures

Selecting literature refers to identifying literatures that can answer the constructed question from all collected literatures in line with the established inclusion and exclusion criteria. Thus, the selecting criteria should be drawn up based on the established research problem and the four factors including the research problem (subjects of study), interventions, main study results, and the study design.

There are three steps to carry out literature selection: (1) Preliminary screening: Screen for literatures which are obviously ineligible according to citation information. (2) Reading the full text: Read literatures that are probably eligible one by one.

(3) Contacting the author: For literatures providing information in the text is obscure, then the reviewers can get in touch with the author and retrieve relevant information for further assessment to decide whether or not to include the literatures.

### Evaluating the Risk of Bias for Included Studies

Bias is a phenomenon that study results deviate from the true values, and it is essential to avoid bias. It can occur in every single step from allocating subjects to intervention groups, following up the subjects, measuring, and reporting results. Evaluating the bias risk in the included studies means to assess the degree to which an individual study can eliminate or minimize bias. Literatures are supposed to include three aspects contents: (1) Internal validity: It refers to how close the results of study are to the true value or the influences of various bias such as selection bias, performance bias, and measurement bias; (2) External validity: It means whether the study results can be applied to other study populations; (3) Factors affecting the results: Factors such as the drug dosage, period of treatment, and compliance in therapeutic trials.

There are five major types of bias: (1) Selection bias: It occurs in the process of selecting and allocating participants when randomization is not perfectly implemented. (2) Performance bias: It happens when one group of subjects in an experiment gets more attention from investigators than another group, and it also refers to the fact that participants can change their behavior or responses if they know which group they are allocated in. This type of bias can be minimized or eliminated by using blinding, which prevents the investigators from knowing who is in the treatment or control group. (3) Attrition bias: It refers to bias arose from systematic differences in the way that participants are lost from a study (differences between people who leave a study and those who continue, particularly between study groups). Over-recruitment can prevent important attrition bias. Also, tailored replenishment samples and sampling weights can compensate for the effects of attrition bias. (4) Measurement bias: It is caused by measuring exposure or disease different between participants in the intervention and control groups, and it might be avoided by adopting standardized measuring methods and blinding the participants as well as result measurers. (5) Reporting bias: It occurs when chance or selective outcome reporting rather than the intervention contributes to group differences. The prevailing concern about this type of bias is the possibility of results being modified toward specific conclusions.

There are many ways to appraise the quality of literatures, such as the list or checklist (there are many items that are not given a score) and scale (each item has a score and weight according to its importance), but is still no consensus. Since these assessment methods can be easily influenced by the quality of literature in combination with some information irrelevant to internal validity as well as the fact that the score of scales is limited by some subjective factors, Cochrane handbook 5.0 does not recommend any checklist or scale, and recommends to use a new risk-of-bias assessing tool created jointly by the methodologists, editors, and systematic

reviewers from the Cochrane Collaboration. The new tool includes six aspects: (1) random allocation methods; (2) allocation concealment; (3) blinding to the participants, implementers of treatment regimen, and results measurers; (4) integrity of the data; (5) selective outcome reporting; and (6) other sources of bias. Result of each single study has to be assessed from the six aspects above according to the standards of “yes” (low bias), “no” (high bias), and “unclear” (lacking relevant information or the bias condition is uncertain). The first, second, and fifth items are used to evaluate risk of bias in the included studies while the other three items are used to assess the different results of the included studies and reveal how biases influences different results from the same study. The result of risk-of-bias assessment can not only be described by words and tables, but also by graphs which can display the bias more plainly.

To avoid the selection bias when reviewers select and assess the quality of literatures, it is useful to adopt blinding methods or considering more researchers (professional and nonprofessional personnel) to do these works. With regard to the disagreements existing in selecting and assessing literatures, then the reviewers can discuss them together or the third party can be asked for help. Also, the consistency (Kappa value) can be calculated when there are several reviewers to select literatures.

### Extracting the Data

When conducting a systematic review, data is not only the statistical figure, but also the collection of information such as the basic information about the study, intervention, outcomes, and results. Extracting data should be comprehensive and accurate, and avoid bias, mistakes, and duplication. The extracting process steps are as follows: determining the data type, designing data collection form, carrying out a pretest to modify and perfect the data collection form, extracting data, checking data, and handling disagreements.

The extracting data include four aspects:

- ① General information: such as author, title, date, original literature number, and source;
- ② Characteristics of study: including the research method, characteristics of participants, settings of study, the type of design, intervention, and preventive and control measures for bias;
- ③ Results: outcome measurement and the results, loss to follow-up and dropout, and adverse effects.

### Analyzing Data and Reporting Results

Results can be obtained using nonquantitative synthesis or quantitative synthesis methods.

### ① Nonquantitative synthesis (NQS)

NQS adopts description method including the participants, intervention strategy, result of the study, the quality of the study, and design technique. And all these presents using forms so that readers can know the process of the study, whether the study method is rigorous, the difference among set of single studies, how to do a quantitative synthesis, and the explanation of the results.

### ② Quantitative synthesis

Quantitative syntheses include heterogeneity tests, meta-analysis, and sensitivity analysis.

*Heterogeneity tests:* Heterogeneity in the results is common. Heterogeneity among a number of different studies may be produced by differences in study design, study populations, and chance, and the extent of heterogeneity might have an impact on the conclusions of a meta-analysis. Heterogeneity can be divided into three categories: Clinical heterogeneity refers to the difference in a set of single studies like the patient factors (such as disease severity, age, and so on), intervention (such as drugs, dosage, and so on), and so on. Methodological heterogeneity means differences in different studies (clinical trials), such as blindness in trials, methods of measurement, etc. Statistical heterogeneity refers to the difference in the effective value of intervention strategies, which is the result of the first two kinds of differences. Heterogeneity test is used to check the degree of variation on the results of the original studies. If the test result is statistically significant, people should explain the reason and consider whether it is appropriate to combine results/findings. There are two methods to make sure whether the respective results have the same property. Firstly, it is to inspect by chart to check whether the effect value of the results overlap between credibility interval. If the credibility gap is very significant, then synthetic analysis cannot be conducted and stochastic effect model should be taken. Q test is another method that can be taken directly. Based on this, the quantitative method can be used to estimate heterogeneity. For example, 0–40% indicates that heterogeneity is not significant, 30–60% means moderate heterogeneity, 50–90% refers to a notable heterogeneity, and 75–100% means a major heterogeneity.

*Meta-analysis:* Quantitative analysis of synthesis can be done according to the type of data and the objective assessment of the amount of selection effects. For instance, odds ratio, risk difference, relative risk, and number needed to treat as the effect size to represent the result of synthesis that can be chosen for categorical variables. And for continuous variables, mean difference can be chosen when measuring the result using the same unit of measurement, while standardized mean difference should be selected when using the different units of measurement. When conducting a meta-analysis, it is always to select the fixed effect model or random effect model, and choose the forest plots to display results.

*Sensitivity analysis:* Sensitivity analysis is to test those important factors (such as inclusion criteria, quality of study, attrition, and statistical methods), which could affect the integration results. Then which factors might influence the integration results and how to do sensitivity analysis based on actual conditions should be considered according to the various studies included in the integration.

## Results Interpretation

The interpretation of the results of systematic reviews must be based on results of reviewer's conclusions. It should address the following issues:

① The strength of evidence

The strength of evidence depends on the study design and the quality of included studies, whether there are important methodological limitations, whether there is a dose-response relationship, how large and significant the observed effects are, etc.

② The applicability of the results

Value of the systematic review results is determined by considering the relationship between pros and cons of interventions in patients. In addition, researchers should consider whether the participants included in the systematic reviews are similar to the conditions of his/her patients. Is there any discrepancy in biological, social, or cultural variations, variation in baseline risk, variation in compliance, etc.?

③ Clarification of important trade-offs between benefits and potential harms as well as costs of the interventions.

## Updating Systematic Review

Carrying out systematic reviews is time- and resource consuming, and provide a snapshot of knowledge at the time of data incorporation from studies identified during the latest search. Newly identified studies can change the conclusions of reviews. The validity of reviews can be threatened if they have not been included, and even the reviews could mislead. Thus, there are clear benefits to updating reviews when new evidence emerges or new methods develop. An update of a systematic review refers to a new edition of a published systematic review with changes that can include new data, new methods, or new analyses to the previous edition. Updating a systematic review requires assessment and revision of the question, background, inclusion criteria, methods of the existing review, and the existing certainty in the evidence. In particular, methods need to be updated, and search strategies might be reconsidered.

### 16.2.1.5 Evaluation and Application of Systematic Review

Systematic reviews have played an increasingly important role in healthcare in recent years, with about 2500 systematic reviews published around the world each year. They are usually used as a starting point for developing clinical practice guidelines. Clinicians, nurses, health-care workers, health policy-makers, and even the general patients read the systematic reviews that were rigorously appraised and scientifically integrated, which can not only greatly save time, but also improve the efficiency of using the related useful information. However, the existing

evidence-based medicine research literatures have good and bad quality. And only high-quality systematic reviews and/or meta-analyses can provide a scientific basis of clinicians, patients, and other decision-makers. Thus, a key stage in a systematic review is to assess the quality of studies to ensure the application and conclusions are based on sound evidence.

It is a basic skill for medical workers and/or researchers to be able to determine whether a systematic review is a high quality. This section describes the basic principles and methods of assessing the quality of a systematic review.

### Assessing the Quality of Systematic Reviews

A high-quality system review should enable the practitioners to understand and critically judge the authenticity and clinical application value of its results. The principles of assessment of the quality are slightly different for different types of systematic reviews. The following section describes the basic principles of assessment of quality for therapy systematic reviews and meta-analysis.

### Principles of Quality Assessment for Therapy Systematic Reviews

Assessment of therapy systematic reviews mainly focuses on three areas, that is, whether the results of a systematic review are true, whether the results of a systematic review are important, and whether the results of a systematic review can be applied to an individual patient.

#### ① The authenticity evaluation on a system review

The evaluation of authenticity of a systematic review result covers four aspects:

- (i) Was the systematic review based on randomized controlled trials? Randomized controlled trials can control significant sources of bias very well, and have high homogeneity. The systematic review conducted on randomized controlled trials with high homogeneity is identified as the highest level of evidence, while it conducted on nonhomogeneous randomized controlled trials or non-randomized controlled trials is indicated as lower level of evidence due to biases.
- (ii) Did the systematic review conduct a thorough literature search and describe retrieval strategy clearly? A systematic search for research studies (systematic reviews) means the more comprehensive literature collected, the less affected by publication bias, and the higher credibility. By reading the description of search strategy, readers can judge whether the literature collection of the systematic review is comprehensive. Inadequate literature search, especially in the case of publication bias, may lead to false-positive results and affect the conclusions of the systematic review. The search



strategy should follow Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement, and present the flow diagram according to it. The flow diagram displays the process for selecting studies, and helps the reader to judge whether the literature search of a systematic review is complete and appropriate.

- (iii) Did the review appraise the authenticity of individual studies? The systematic review is a type of study that analyzes and summarizes the results of the original studies. Thus, the quality of systematic review is not only affected by the methods used, but also affected by the included studies. Systematic reviews need to clearly state the evaluation methods of the studies included, and describe whether it uses multiplayer independent evaluation and appraises the consistency among them.
- (iv) Did a meta-analysis of individual patient data is conducted? Individual participant data from multiple studies are used to conduct a comprehensive meta-analysis. Meta-analysis has more advantages than other various studies. For instance, it can examine heterogeneity in included studies at patient's level by consistently defining the results of different studies and unifying the threshold, test hypotheses by using subgroup analysis, determine the randomization method clearly and judge the quality of assessment through contacting the authors, and reduce the effect of system bias and opportunity errors using existing medical record information.

② Are the results of systematic reviews important?

Applying the results of a systematic review to clinical practice should not only emphasize the importance of using the systematic review findings in clinical practice, but also make sure whether the systematic reviews include high quality of individual studies? What main outcomes are reported in the systematic review? Is the precision of results reported? Specifically, the assessment covers the following:

- (i) Are the results consistent across the included studies? System reviews usually include high-quality original articles, with sufficient individual study and better homogeneity of the study results. The authors should test the heterogeneity first. If the results of a single study are homogenous, then the findings of different studies can be integrated. On the contrary, if result of homogeneity test is significantly different, the author needs to explain the differences and consider whether it is appropriate to combine the individual study results.
- (ii) How accurate is the systematic review? In systematic reviews, more accurate conclusions are considered mostly. Given the impact that quality assessment can have on the findings of systematic reviews, it would be reasonable to give different weights according to the quality of the studies when synthesizing the results of the included studies. The application of a systematic review is associated with the outcome indicators. If a systematic review uses low related intermediate outcomes, it may limit the findings of the systematic

review being applied to guide clinical practice. If the outcome indicators are the event rates, such as mortality and the incidence of serious adverse events, it may bear important clinical significance even if the combined effect did not generate statistically significant difference.

Indicators used in systematic reviews include risk difference (RD), odds ratio (OR), relative risk (RR), weighted mean difference (WMD), number needed to treat (NNT), and number needed to harm (NNH). Among them, NNT and NNH are easy to calculate and are among the most clinically useful statistics. NNT is the number of patients which need to treat to prevent one additional bad outcome (death, stroke, etc.). When NNT is smaller, preventive effect is better. NNH is an epidemiological measure that indicates how many patients need to be exposed to a risk factor to cause harm in one patient that would not otherwise have been harmed. Intuitively, the lower the number needed to harm, the worse the risk factor. The NNH is a crucial EBM metric that aids doctors in determining if it is wise to proceed with a specific treatment that can endanger the patient while yet having therapeutic advantages. Drugs with a low NNH may still be indicated in specific circumstances if the number needed to treat (the opposite of side effects, or the advantages of the medicine) is less than the NNH if a clinical end point is devastating enough without the drug (e.g., death and heart attack).

③ Can the results of systematic review be applied to an individual patient?

The findings of systematic review are the average effect among all participants. For a particular patient, to answer whether the findings of a systematic review could be applied, the following four aspects should be assessed:

- (i) Is the patient similar to those in the studies included in the systematic review? It can compare the differences between the patient and the cases in the systematic review, focusing on the age, gender, race, comorbidities, severity of disease, duration, socioeconomic status, cultural background, complications, compliance, etc.
- (ii) Is the therapy feasible and safe in the clinical trial setting? Since there is difference in socioeconomic conditions, technology, and equipment conditions, the conclusions of systematic reviews sometimes can hardly be applied to a particular patient, even if the intervention effects are remarkable.
- (iii) What are the potential benefits and harms of therapy for the patient? Only when the advantages outweigh the disadvantages, results of the system review have applicable value.
- (iv) What are the patient's values and preferences? Incorporating patient values and preferences as an essential input for decision-making has its potential merits. EBM stresses that any medical decision-making should take into account of the combination of personal experience and expertise, current best research evidence, and patient choice.

## Assessing the Quality of a Meta-Analysis

Assessment of the quality of meta-analysis should consider the following questions:

- ① Does the question conform to the principle of DOE?
- ② Does the question state clearly?
- ③ Is the search strategy clearly described?
- ④ Are appropriate inclusion and exclusion criteria used to select articles?
- ⑤ Is there worrying homogeneity? Has it been appropriately addressed?
- ⑥ Are the statistical methods used correctly? Has a sensitivity analysis been undertaken? If not, should it have been?
- ⑦ Does the combined analysis clarify the difference between the test group and the control group?
- ⑧ Are the conclusions appropriate?
- ⑨ Do you put the recommendations of the systematic review in your clinical practice?

### 16.2.1.6 Methods of Evaluating System Review

Effective quality assessment is very important for properly using the conclusion of a system review/meta-analysis. Quality assessment instruments of systematic reviews mainly include two types: one is to assess the methodological quality of systematic reviews and the other is to appraise the quality of reporting of meta-analyses. The poor quality of reporting will affect the applicability of the results of systematic reviews. Thus, it is recommended to use both of the quality assessment tools when assessing the quality of systematic reviews/meta-analysis.

#### Tools for Methodological Quality Assessment

Methodological quality assessment aims to understand whether the review follows the scientific standards and effectively controls the biases in the process so as to get true and reliable results. Biases are the main challenge for the quality of systematic reviews. When assessing the authenticity of systematic reviews, the focus should be put on examining the quality of the methods used in systematic reviews, and how to control biases. Usually, methodological quality assessment instrument of systematic reviews includes the following: A Measurement Tool to Assess Systematic Reviews (AMSTAR), Critical Appraisal Skills Programme (CASP), Sacks Quality Assessment Checklist (SQAC), and Overview Quality Assessment Questionnaire (OQAQ).

## Tools for Reporting Quality Assessment

Reporting quality assessment aims to ensure complete, accurate, and transparent reporting of research studies. The main content of the process of assessment is to assess the integrity and comprehensiveness of the systematic review. Report specification is one of the quality assessment tools used to appraise the reporting quality of the systematic review. It includes asking a research question, review methods, presentation of the results, discussion, and conclusion. The reporting specification of the systematic review is mainly used for assessing the quality of a system review report, and cannot be used to assess methodological quality of systematic reviews.

The Quality of Reporting of Meta-analysis (QUOROM) is the first report specification used to assess the quality of reporting of meta-analyses of clinical randomized controlled trials. QUOROM covers 6 aspects with a total of 18 items, that is, title, abstract, introduction, methods, results, discussion, and conclusion section. QUOROM has been considered as a “gold standard” to assess the quality of systematic reviews/meta-analysis. In 2006, QUOROM was updated as PRISMA. The PRISMA Statement was released in 2009, and an official update of it is currently under development. It consists of a 27-item checklist and a four-phase flow diagram (Fig. 16.2). PRISMA Statement focuses on randomized trials, but can also be used as a basis for reporting systematic reviews of other types of research, particularly evaluations of interventions. PRISMA may also be useful for critical assessment of published systematic reviews. The full score of PRISMA scale is 27 points: “full report” gets 1 point, “part of the report” gets 0.5 points, and “not reported” gets 0 points. When the score is between 21 and 27, the report is considered relatively complete; when the score is between 15 and 21, the report is considered to have some flaws; and when the score is equal to or less than 15, it is considered that there may be serious missing of information.

### 16.2.1.7 Application of Systematic Review

Sound and reliable evidence providing by a high-quality systematic review is of great importance for clinical decision-making. It is necessary for decision-makers to concern whether the results of a systematic review can be applied to clinical practice. A high-quality systematic review should enable readers to judge whether the results of systematic reviews can be used to solve a specific clinical problem. For example, in a systematic review of disease prevention study, the following aspects are the main concerns:

1. The clinical problems (the review questions);
2. The basic information such as age, gender, disease, and diagnosis;
3. The inclusion criteria and exclusion criteria;
4. The interventions, such as the intervention method of experimental group and the control group;
5. The information on outcome indicators.

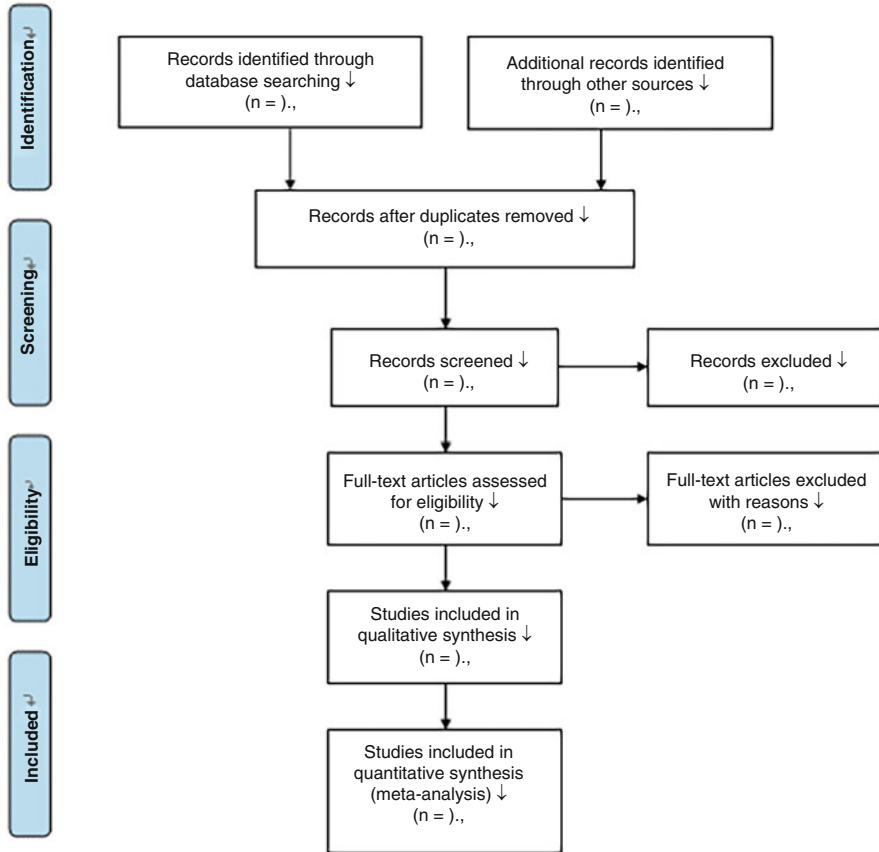


Fig. 16.2 PRISMA 2009 flow diagram

Systematic reviews encompass all fields of clinical medicine, including the studies on disease etiology and risk factors, disease diagnosis, treatment, disease prognosis, rehabilitation, and prevention. Results of systematic reviews have been widely used to develop clinical practice guidelines, facilitate health policy decision-making, guide clinical practice, and enlighten scientific research. In addition, there is a series of guidelines to help interpret and use the systematic reviews.

The application of systematic review should meet the needs of clinical practice, medical education, scientific research, and health policy-making. Results of systematic reviews should reflect the latest evidences and developments in a given area. In clinical practice, when a clinician needs to know whether a systematic review conclusion can be applied to solve clinical problems, firstly, he/she must consider the following questions by checking the results of the systematic review: Are the benefits important and necessary for clinical decision-making? Are there less potential side effects than other interventions? Are the costs less than other methods? Is my patient similar to the participants in the systematic review?

## 16.3 Meta-Analysis

### 16.3.1 Introduction

A meta-analysis is defined as a systematic review that uses statistical methods to aggregate quantitative data from at least two and ideally several studies on a given research issue to produce an overall quantitative estimate of effect. Narrowly speaking, the meta-analysis refers to the quantitative analysis of a system review. However, only a few systematic reviews can be quantitatively analyzed because of the differences in quality of research, design, methodology, etc. The general steps of a meta-analysis are as follows: formulation of the research questions, literatures retrieve, making inclusion and exclusion criteria, describing the basic information, doing comprehensive quantitative analysis, and a series of other steps.

At present, meta-analysis has been widely used in clinical medicine. In particular, it has been used for small effect size or controversial research (mainly for RCTs), etiology study, the dose-response relationship research, diagnostic trials, and prognosis research.

#### 16.3.1.1 Basic Concepts

1. Effect size is related to different variables or values caused by treatment effect.
2. Heterogeneity test refers to a test that evaluates whether variation between the results stemmed from a random error.
3. Bias means a system error in study designs, trial implementation, data analysis, and results interpretation. It brings about difference between the results and the real situation, and thus wrongly describes the relationship between exposure and disease.
4. Publication bias refers to studies which present negative results or insignificant results are less likely to be published.
5. Sensitivity analysis is that researchers can eliminate lower quality studies or use different statistical methods to analyze the same data by changing the inclusion criteria so that to observe the changes of the meta-analysis results.

### 16.3.2 Steps to Perform a Meta-Analysis

The general steps of a meta-analysis are as follows:

#### 16.3.2.1 Data Extraction

Accuracy and reliability of the data is the key of a meta-analysis. Thus, when conducting the data extraction, a research should collect data onto many channels

to ensure complete data included. At the same time, effective quality control measures should be used to prevent selective bias.

### 16.3.2.2 Data Types and Effect Size

Data used for a meta-analysis mainly include the following five types: (1) Continuous numerical variable data: often have the units and can accurately be measured, such as height and weight. (2) Binary variable data: have two incompatible categories, such as survival or death. (3) Data in hierarchical classification of variables: can be divided into multiple classes, and have degree or level differences, like never/rarely, sometimes and often, (4) Count data: individuals in a certain observation time experiences many adverse events, such as myocardial infarction and fracture. (5) Survival data: observation of two types of data at the same time, the occurrence of adverse events, and the time of adverse events occurring.

Different data types determine different effect size expressions. When the outcome is a binary variable, the effect size commonly used is odds ratio (OR), relative risk (RR), absolute risk (AR) or NNT, etc. When outcome is continuous variables, the effect size is the mean differences (MD) or standardized mean differences (SMD). For hierarchical data or count data, it can be converted to binary variables or continuous variables in light of the actual conditions. For survival data, the effect size used is a hazard ratio (HR).

In addition, the important information, such as sample size, analysis methods, design, main outcome variables, publication year, and quality control measures should be included in the study.

### 16.3.2.3 Heterogeneity Test

When carrying out a meta-analysis, strict literature inclusion and exclusion criteria should be used to control the heterogeneity source maximally. However, because of the differences in the research participants, design and statistical analysis model, heterogeneity will occur inevitably. In such case, there will be a mistake if the results are combined. Thus, heterogeneity test should be done before conducting a meta-analysis, and determine whether to estimate the combined effect size according to its results. If the heterogeneity is obvious, the source of the heterogeneity should be addressed. The heterogeneity test includes Q test and visual graphic method. Also, there are some other methods to show the heterogeneity, such as standardized Z score, radial figure, forest figure, and L'Abbé figure. The most commonly used method to determine the heterogeneity is to observe confidence interval overlapping degree in forest figure. If most confidence intervals overlap and there is no obvious abnormal value, then it can be recognized as a high homogeneity.

#### 16.3.2.4 Combining Effect Size Estimates and Hypothesis Testing

Next, appropriate statistical methods should be chosen based on the results of heterogeneity test. If the heterogeneity is not obvious, namely, assuming that the theoretical effect size is a fixed value and differences between effect variables caused by opportunity, then the fixed effect model can be used to estimate the combined effect size. If there is heterogeneity, and the assumption theoretical effect size is not fixed, but follows a certain distribution pattern, such as the normal distribution, then the random effect model can be adopted to estimate the effect. If the heterogeneity is too great, subgroup analysis and meta regression analysis can be considered, or only describe the results.

There are many methods to estimate combining effect size, like Mantel–Haenszel method, Peto method, and variance inversion method. According to the characteristics of data, different methods can be used. For binary variable, Mantel–Haenszel method can be used, and for continuous variables, variance inversion method can be adopted. Z test is used to test whether there is a statistical significance of combined effect variables.

### 16.3.3 *Fixed Effect Model and Random Effect Model*

Model selection depends on the results of heterogeneity test and the theoretical effect hypothesis. If the heterogeneity test is not statistically significant, then it can be deemed that theoretical effect size is fixed. Then, a fixed effect model can be selected to estimate it. On the other hand, if the heterogeneity is larger, and assuming the theoretical effect size follows a normal distribution, then a random effect model need be chosen. The random effect model takes variation factor  $\tau^2$  of the study as the correct weight, so the result is more robust than the fixed effect model.

#### 16.3.3.1 Fixed Effect Model

For binary variable data, MH method fixed effect model can be chosen to estimate the combined effect size. If it is continuous variable data, and there is no statistically significant heterogeneity, fixed effects model can be used for meta-analysis, and the process is the same as that for binary variable data. The variance inversion method should be used for combining effect size estimation. For effect size expression of continuous variable data, mean difference (MD) or standardized mean difference (SMD) need to be used. When all studies use the same way to measure outcome variables, MD can be used as the effect size. If those outcome variables have the same definition, but not have the same measuring scale, then SMD should be selected, and authors need to explain these results carefully.



### **16.3.3.2 Random Effect Model**

When the heterogeneity test is statistically significant, and assuming the real effects size is not fixed, but follows a normal distribution pattern, then the random effect model can be selected to estimate and combine the effect variables. The random effect model adds DerSimonian–Laird correction to fix effect model on the basis of the variance inversion method or MH method. Weight between the two models is different. The fixed effect model takes the reciprocal of variance as the weight in individual studies. The random effect model takes the reciprocal of the sum of the variance in the study and variance between studies as the weight, and the adjustment results give less weight for larger sample size study.

## ***16.3.4 Evaluating the Result of a Meta-Analysis***

### **16.3.4.1 Heterogeneity Test**

If the research has adequate homogeneity, then a fixed effect model, random effect model, or both can be used to estimate the combined effect size. If the research has adequate heterogeneity and the source of heterogeneity is known, then meta regression model or subgroup analysis can be selected. If the heterogeneity test is statistically significant, but heterogeneous source is unknown, then random effect model estimation is more conservative. When assuming that the effect size is not fixed, but obey a normal distribution, then random effects model can be adopted. If heterogeneity is too great, then meta-analysis cannot be conducted.

### **16.3.4.2 Robustness of Meta-Analysis Results**

Sensitivity analysis is often used to check the robustness of meta-analysis results. In sensitivity analysis, by changing the inclusion criteria to eliminate lower quality study, or using different statistical methods to analyze the same data, which can be used to observe the changes of the meta-analysis results so as to evaluate the resulting stability. For instance, after excluding some lower quality studies, the combined effect size is estimated again, then to compare the results with original meta-analysis results and explores the influence of these lower studies on the combined effect size and the stability of the results. If the result changes a little, it has lower sensitivity. On the contrary, after excluding some lower quality studies, the difference become bigger, or even the meta-analysis gets the opposite conclusions. It is of high sensitivity and poor robustness, and the results need to be explained carefully.

### **16.3.4.3 Applicability of the Meta-Analysis Results**

The application of meta-analysis avoids the limitations of individual clinical trials with small samples, and makes the results of analysis more comprehensive and reliable. Therefore, it can provide a good basis for medical decision-making.

# Chapter 17

## Disease Prognosis



Fang Wang

### Key Points

- Prognosis is a prediction of the outcome and influencing factors of the disease following its onset.
- Risk factors and prognostic factors are often considerably different. The most common design of prognosis is cohort study.
- Case-fatality and five-year survival are often used to express prognosis. The life table approach and the Kaplan-Meier method can be used to calculate observed survival over time.
- Assembly bias, migration, zero bias, and survival cohort bias are major sources of bias in prognosis study. Randomization, matching, COX proportional hazards model and other methods may help control bias in prognostic studies.

When a person developed a disease, doctor, patient, and even the patient's family members may have lots of questions about the disease. Will it go further to be worse? Could it be cured? How about the possibility of a worse outcome? How long do the patients have to continue with their normal activities? All those questions are discussed about the prognosis of a disease. In this chapter, we will introduce qualitative and quantitative ways that prognosis can be described.

---

F. Wang (✉)  
School of Public Health, Shanxi Medical University, Taiyuan, China  
e-mail: [wfang@sxmu.edu.cn](mailto:wfang@sxmu.edu.cn)

## **17.1 Basic Concepts**

### ***17.1.1 Concept of Prognosis***

Prognosis is a prediction of the outcome and influencing factors of the disease following its onset. The outcomes can be recovery, relapse, disability, deterioration, complications, and death. Prognostic studies are to identify the probability and possible influencing factors of those outcomes. Understanding of prognosis helps to predict the future of the patients better. Clinicians may choose more appropriate decisions on the following aspects: (1) What kind of treatment guidelines should clinicians follow? and (2) What kind of treatment should be adopted?

A better understanding of influencing factors of disease outcomes can alter the outcome of a certain disease. If there are several types of medical treatments, clinicians can compare the effectiveness of different therapy.

### ***17.1.2 Natural History and Clinical Course of Disease***

The natural history of disease refers to the prognosis of disease without medical intervention which can be divided into the following four periods: the biological stage, subclinical stage, clinical stage, and outcome. Changes like DNA alternation are subcellular in the biological stage, and thus often cannot be defined due to the sensitivity of the clinical test. In the subclinical stage, pathologic evidence develops and could be obtained. Later, when noticeable signs and symptoms like pain, disfigurement, or fever occur in the clinical stage, patients may come to seek help from the clinician and then a diagnosis may be made. After treatment, the outcomes may be cure, disease controlled, or even death.

The natural history of different diseases varies greatly. Some diseases with a short latency may present obvious symptoms and outcomes in a short period of time, such as acute infectious diseases. Some chronic noncommunicable diseases may have a relatively long natural history such as cardiovascular diseases and diabetes. Different strategies can be taken in different stages of natural history to improve prognosis.

The clinical course is the progression of disease following medical interventions. Patients receive a variety of treatments that may affect subsequent course of the disease. The clinical course may be altered by medical intervention; however, there is no medical intervention in natural history. The earlier the effective treatment, the better the prognosis. Prognostic researches are about the clinical course and medical treatment that can improve prognosis and alter the outcome.

### 17.1.3 Prognostic Factors

Prognostic factors are conditions that are associated with the outcomes of disease. Prognostic factors help to identify groups of patients with different outcomes.

Prognostic factors are different from risk factors in two ways. Firstly, risk factors are those associated with increased risk of a disease in healthy people, whereas studies of prognostic factors deal with sick people. Secondly, risk and prognosis describe different phenomena of disease. Risk describes the onset of disease, usually predicting low-probability events. The incidence of disease varies from 1/1000 to 1/100,000 or even less. The study of risk factors requires a relatively large amount of population to evaluate or confirm the relationship between exposure and disease. Prognosis describes a variety of disease consequences following its onset. The consequences, including recovery, disability, complications, and death are relatively frequent events.

1. For a given disease, risk factors and prognostic factors are not necessarily the same and are often considerably different in the following three important points.

Factors associated with an increased risk of disease have little to do with prognosis, in other words, risk factors do not necessarily make a worse prognosis. For example, high blood pressure increases the risk of acute myocardial infarction, but it is not related to a worse outcome of the acute event.

Factors associated with a certain outcome of disease do not have an association with increased risk of disease. Those prognostic factors are not risk factors of a certain disease. Infarction location and arrhythmia are prognostic factors of acute myocardial infarction, but they do not increase the chance of having the disease.

Some factors do have a similar effect on both risk and prognosis. Those factors can not only increase the risk of a certain disease, but also lead to a worse outcome. For example, with the increase of age, both the risk of an acute myocardial infarction attack and the risk of death from it may increase.

2. Prognostic factors are complex and variable, which can usually be described by several categories.

Timing of diagnosis and treatment: Early diagnosis and proper treatment for any diseases are important prognostic factors. The 5-year survival rate of gastric cancer can be up to 100% if discovered early, but may fall to less than 20% for advanced gastric cancer discovered through normal diagnosis.

Characteristics of the disease: The spectrum of disease – from mild to severe – is related to outcomes. Patients with mild illness can have a better prognosis than the severe ones. The duration and pathologic types also vary. Different diseases have different natural histories.

Pathogenic factors: The quantity, quality, and invasive manner of pathogen can affect the consequences.

Characteristics of the patients: The demographic characteristics, genetic background, nutrition, immune system function, and psychological state of the patients all have some influence on prognosis.

Social and family aspects: Economic development level, social insurance system, local medical condition, family economic situation, the relationship between family members, and the patient's religious/belief can affect the prognosis of disease.

## **17.2 Design of a Prognosis Study**

In the beginning, qualified patients with a complete description are selected as the study population. Then the patients are observed and followed up at a pointed time. Finally, all the concerned outcomes are measured to describe the prognosis of disease.

### **17.2.1 Research Methods**

Studies of prognosis are like those of risk. Generally, relevant prognostic factors are identified through descriptive study, and verified through case-control and prospective cohort studies. Any method can be chosen, depending on the study purpose, resources, and time.

In descriptive study, patients with diabetes were randomly selected as subjects. And then they were asked about habitual, diet, and treatment to identify whether those factors were possible prognostic factors of diabetes complications.

In case-control study, newly diagnosed diabetes patients with complication, that is, diabetic nephropathy, were selected as patients; diabetes patients without complications were selected as controls. Cases and controls must both meet the inclusion criteria to ensure they were from the same base population. Controls may be matched to cases on age and gender. The frequency of smoke, cyto-factors and other interesting factors were measured and compared within the two groups. It is efficient and indispensable as it does not need to collect data from a large number of people. But selection bias and recall bias are difficult to manage.

The most common design is cohort study. The incidence and relative risk are measured directly. Prognostic factors are measured before the outcome of the disease. It is discussed in detail in the following section.

### **17.2.2 The Patient Sample**

It is best for a prognosis study when the patients are selected based on the population of all people with the disease in a certain region. Under this condition, the patients' sample is representative and there will be no bias in patient selection. However, most

studies are hospital-based and the outcomes for patients with the same disease may be different when selected from different levels of health facilities.

It is important for a prognostic study to thoroughly describe the patients' characteristics, the sampled and assigned method, the criteria of diagnosis, inclusion, and exclusion. Thus, it will be good for others to reference the study results and decide whether the conclusions can be applied to their patients.

### ***17.2.3 Determine the Starting Time Point***

Cohorts are observed from a pointed time in the course of the disease, which is called zero time. The point can be the time when patients are enrolled in the cohort, such as the onset of symptoms, date of diagnosis, or the beginning of medical interventions. Whatever, it is especially important to make a clear definition of zero time. If the patients are recruited near the onset of the disease, the cohort is called the inception cohort.

What if the patients are observed at different points of the disease? When zero time changes for patients, precise description of subsequent events would be much more difficult. Interpretation of the timing of recovery, death, and others would be hard or even misleading. For example, a cohort of women with breast cancer is assembled. However, women in the cohort are at different stages of breast cancer. Those in the early stages can have better survival than those beginning surgical treatment. Moreover, these results of cancer prognosis would be hard to interpret.

### ***17.2.4 Determine the Outcomes***

Outcomes should be clearly defined and a full range of clinical events of the disease should be included. While clinicians tend to focus on the clinical effects, such as normalization of blood chemistries or reduction of tumor size, patients are more concerned about the remission of symptoms. To guide patient care, outcomes should be related to something that patients can perceive.

Outcomes vary a lot, and death is the easiest to determine. Some outcomes, like myocardial infarction, usually need to be valued precisely or technically measured. Others like the health-related quality of life, are difficult to determine directly, and often need a set of variables to measure. The value of the prognosis study is strengthened when blinding is applied to outcome determination, and composite outcomes are reported.

### ***17.2.5 Sample Size***

There are no special points in the estimation of the sample size in prognosis study. Experience and formulas both can be applied to determine sample size. In clinical studies, each prognostic factor requires at least 10 individuals. If there is more than one prognostic factor, the maximum sample size required by the factor is taken.

### ***17.2.6 Follow-Up***

Follow-up is important, and all patients should be followed for an enough long time to observe concerning events, including some rare adverse outcomes. The length of follow-up depends on the course of the disease. For some communicable diseases, the period is several weeks and decades years for the onset of hepatitis B. The interval should be reasonable to obtain various dynamic changes of the disease. The outcome of diseases with a short course changes rapidly, and the interval can be shorter than those with a longer course.

## **17.3 Describing Prognosis**

### ***17.3.1 Case Fatality***

Case fatality, discussed in Chap. 2, is often the first way to describe prognosis. It is defined as the percent of people who die of a disease in people with the disease. Case fatality is usually used to express the prognosis of short-term diseases, such as acute infectious diseases, acute phase of cardio-cerebrovascular diseases, and cancers with short survival. Case fatality is not suitable for chronic diseases as death may occur many years after diagnosis and competitive events occur more likely.

### ***17.3.2 Remission Rate***

Remission rate refers to the percent of patients in clinical undetectable state after treatment.



### 17.3.3 Recurrence Rate

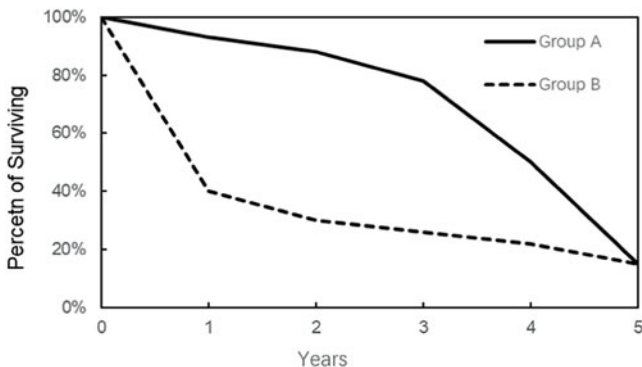
Recurrence rate is the percent of patients who return to disease after a period of remission or recovery.

### 17.3.4 Disability Rate

Disability rate is the percent of patients who are unable to function normally as a result of the disease.

### 17.3.5 Quality of Life

Describing prognosis as a single rate is relatively simple. However, much information is hidden, and in most cases, survival can be quite different with similar rates. Figure 17.1 shows the survival curve of two populations. The 5-year survival is about 15% in both populations, but clinical courses are quite different. In population A, most deaths occurred in the fifth year, while most deaths occurred during the first 2 years in group B. Although 5-year survival is the same in the two groups, survival in group A is clearly better than group B.



**Fig. 17.1** Five-year survival curves in two hypothetical populations

## **17.4 Analysis for Prognosis Study Data**

When we summarize prognosis as a single rate, it does not include the likelihood the patients will experience an outcome at any point during follow-up. Survival analysis is a simple statistical method for estimating the survival of a group over time. It is performed to describe the survival distribution of participants in the cohort. The curves can help to figure out information about the survival event at any point in the course of the disease.

To learn about survival, patients in the cohort are at the same starting point in the disease course and all followed up. Thus, completely data can be achieved when the outcome of interest occurs during follow-up. However, when patients die of the disease other than the outcome event, dropout at any point of time or loss to follow-up, only incomplete data can be obtained which is called censored data. Survival analysis can make efficient use of all data and can be applied to any outcomes from all subjects in the cohort.

Survival time is defined as the period between the starting event and the terminal event of the disease. In survival analysis, the most basic thing is to calculate survival time. Time intervals can be made of interest. In general, life table and Kaplan–Meier analysis are adopted to analyze survival data. Survival curves can be compared by log-rank test. COX proportional hazard model is used to estimate the hazard risk of multiple prognosis factors.

### **17.4.1 Calculating Survival Rate**

#### **17.4.1.1 Five-Year Survival**

Five-year survival is the percent of patients who are alive 5 years from a certain point (usually diagnosis or treatment) in the clinical course of the disease. Actually, 5-year survival is a proportion instead of a rate. It is frequently used to measure the prognosis of cancer after diagnosis because most deaths occur during this period.

#### **17.4.1.2 Life Tables**

Life table is the most commonly adopted approach to measure actual observed survival over time. It is a little more sophisticated method that tries to predict the prognosis of patients. The probability of surviving for one interval following the start of the observation is calculated. Cumulative survival is defined as the proportion who survived from enrollment to the end of the interval. Cumulative survival for the entire follow-up period is the product of each surviving probability of each period. Hazard is usually estimated for a year at a time to estimate survival. Person-years is calculated sometimes and adopted as the denominator. All the data obtained are used

including patients who entered the cohort after the beginning of the observation. Compared to 5-year survival, life table describes survival experience of patients in a more efficient and economical way.

### 17.4.1.3 Kaplan–Meier Analysis

In Kaplan–Meier method, the exact point in time when each outcome occurs is identified. Survival time, including censored data, is arranged from small to large. Survival probability is the ratio of the number of survivals at each point to the number of followed-up till the end of observation (including the alive and death of the disease at that point of time) except for the dropouts. The overall survival is the product of survival probability at each point. The survival curve is the total survival experience during follow-up time presented in a graph. Figure 17.2 shows two simplified survival curves. The horizontal axis is the follow-up period from the beginning of the observation, and the vertical axis is the estimated survival probability. At the beginning of the observation, no one dies and the survival probability is 100%. As time goes on, more and more deaths occur and the probability of surviving decreases. Here, censored data is not used to calculate survival at each point of the time since it is not related to prognosis. If the sample size is small, the survival curve is shown in a stepwise fashion because survival is constant and changes only when the next event happens (Fig. 17.2a). The steps would diminish with the increasing number of patients and for a large cohort, the curve would become a smoothed slope (Fig. 17.2b). Although the follow-up intervals between each new death can be yearly, monthly, or weekly depending on interest or need, the estimated survival is more accurate when the intervals are shorter.

By plotting survival time rather than calendar time on the horizontal axis, Kaplan–Meier analysis deals with censored data and makes considerable use of information on follow-up patients, dropout patients, and deaths. If a patient is censored at 21 months, it is assumed that he/she survived until the next death occurred. The survival curve contains more information besides the observed rates and can be used for any type of time-to-event data. It helps to predict the prognosis of patients with the information of all available data. Although it is used to be applied to

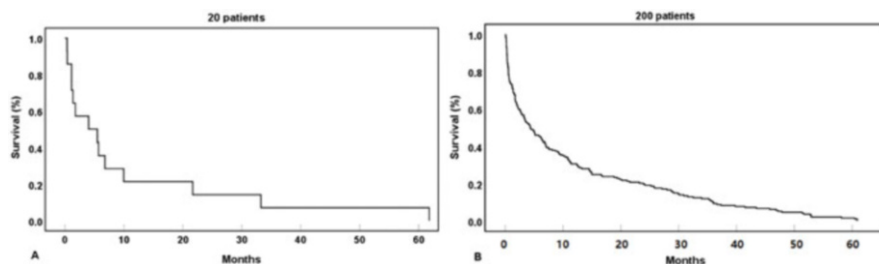


Fig. 17.2 Survival curves of two cohorts with 20 and 200 patients, respectively

studies of small number, more and more large data on survival are now accomplished by Kaplan–Meier analysis with appropriate statistic software.

### ***17.4.2 Several Points About Interpreting Survival Curves***

When interpreting survival curves, several points need to be noticed. The survival rate on the survival curve comes from patients of a hypothetical cohort. It is an estimated probability of the hypothetical population instead of a real population. The survival probability is the best estimation of survival for patients in the cohort. However, the precision of estimated survival depends on the number of cases under observation. On the left side of the curve, in the earlier period of follow-up, there are more individuals at risk. However, on the right side, at the tail of the curve, fewer and fewer cases are at risk because of deaths, dropouts, and late entries. As a result, the estimation would be more reliable on the left side of the curve than the right-hand side.

### ***17.4.3 Comparison of the Survival Curves***

We just discussed the estimation of the survival curve for a single group. If there are two groups, for example, in the clinical trial, survival is estimated separately. Now comes the question, how could we learn the differences between the two survival curves?

In survival analysis, instead of mean survival time, we express average survival time with median survival time since survival time is usually skewed distribution. Median survival time is the length of time that 50% of the study population has had the event. It is the survival experience of one-half individuals and can be easily estimated from the Kaplan–Meier curve. The confidence intervals may also be calculated. In this case, we can compare median survival times for the two groups by easily estimating the ratio of the two medians.

For the comparison of the overall survival experience, Mantel–Haenszel or log-rank chi-square test is applied. The null assumption is that the two survival curves are equal. Essentially, it is used to determine the difference between the observed number of events and the theoretical number at the time of each event. When the null hypothesis turns out to be false, Mantel–Haenszel statistic weights the survival experience on the right side of the curve more than the left side. The log-rank statistic is more powerful when the hazard rates are proportional. When the hazard rates are not proportional, the difference between the two curves will not disappear until the curves cross at a point. The conclusion is still valid but may not be as powerful as when the hazard rates are proportional. If there is a cross in the two curves, the conclusion should be interpreted cautiously.

Life table and Kaplan–Meier analysis can not only be applied to calculate survival, but also useful for any dichotomous or once-occurring outcomes. The end points other than death can be the development of diabetes, the recurrence of acute myocardial infarction, or the cure of infection. When we describe such events other than survival, time-to-event analysis is adopted.

### ***17.4.4 Dealing with Multiple Prognostic Factors***

COX proportional hazards model is an appropriate model that can deal with multiple variables at one time. The dependent variable is the time from the beginning of observation until the concerning outcome. The independent variables are factors that might be associated with the outcome. Hazard risk is the ratio of probability derived from the time-to-event analysis and is a reasonable approximation of relative risk. COX model can be fitted quickly with suitable statistical software.

## **17.5 Common Bias and Controlling**

### ***17.5.1 Bias in Prognosis Study***

#### **17.5.1.1 Assembly Bias**

Assembly bias is one form of selection bias that is also called classification bias. When patients are assembled in the cohort, it is the best that the distribution of the non-studied variables in the groups is similar. However, some important factors, such as the extent of disease progression, the existence of complications, the course of disease, and prior treatment, often differ in the groups. Moreover, those factors may determine the outcome. Therefore, the susceptibility to the outcome would not be equal among the groups being compared.

#### **17.5.1.2 Migration**

Migration is also one form of selection bias. When study subjects leave the original group, drop out of the research altogether or move to another group, migration bias occurs. If the migration in the groups is random, there would be no bias. However, the number and the characteristics of patients dropping out or moving to another group are usually not the same in different groups. As a result, when these migrations are large enough, the groups are no longer comparable, and the validity of the conclusion will be affected.

### **17.5.1.3 Loss to Follow-Up**

Patients in the cohort may drop out at any point due to the long follow-up period of observation; moving away from the place that they are now living, refusing to continue the research, becoming ill, encountering a competitive event, or are suffering from side effects of the treatment. The outcomes of those dropping out of the study cannot be obtained. When the number of losses to follow-ups takes place on a large scale, the validity of the study would be affected. It is generally considered that the proportion of loss to follow-up should be kept under 10%.

### **17.5.1.4 Survival Cohort Bias**

In a cohort study, if patients in the groups are assembled from a hospital because they are receiving treatment there and are just available for the recruitment, the cohort is called a survival cohort or available patient cohort. Survival cohorts are not really cohorts and should be distinguished from true cohorts. Patients in a true inception cohort are observed from the onset of the disease, while patients in survival cohorts are recorded at various points in the course of the disease. Also, survival cohorts do not include patients not in the hospital. Therefore, reports of survival cohorts are lack of precision and are misleading.

### **17.5.1.5 Zero Bias**

In a cohort study, all patients should be followed-up at the same starting point of the disease course. This starting point, also called zero spot, can be the time of diagnosis or the beginning of treatment. Zero spot should be well defined and complied throughout the research. If patients in the cohort were assembled at different points of the disease course, zero bias takes place and could have an effect on study validity.

### **17.5.1.6 Measurement Bias**

Some clear and objective outcomes, such as death, major cancers, and cardiovascular disorders are rarely misdiagnosed. However, other outcomes without a clear definition, such as side effects, subclinical diseases, and special course of disease or disability can be missed due to misclassification or different diagnosis criteria. Measurement bias occurs when researchers try to record those less-clear-cut outcomes, especially when patients in one group have more opportunity to detect their outcomes than patients in another group. The “differences” found among the groups are due to the inconsistencies in the recording of outcomes rather than possible prognostic factors.

## **17.5.2 Control of Bias**

### **17.5.2.1 Randomization**

Randomization is the best way to control confounding. Each subject has an equal chance to be randomly assigned to the experimental group or control group. Not only known factors but those unknown variables that might affect prognosis are balanced in both groups, which means that baseline information in different groups is comparable. As a result, the conclusion will be more precise when differences in prognosis found between groups are caused by certain prognosis factors.

However, the application of randomization is limited. In cohort or case-control studies, it is not possible to adopt this process. It is only possible to use randomization when the study purpose is to evaluate the effects of treatment measures on prognosis.

### **17.5.2.2 Matching**

Patients in the study group can be matched with one or more patients in the comparison group with the same characteristics except for the prognostic factor of interest. The non-studying characters of patients in different groups become similar through matching. Factors such as age and gender are chosen for matching for their relatively strong relationship with most diseases. However, other variables such as race, clinical stage, the severity of the disease, and treatment history may also be considered for matching. Paired matching and frequency matching are two matching choices. For paired matching, one may select no more than four controls at a time. One case and one or two controls are often adopted. In frequency matching, the distribution of the matching factors should be kept consistent between the two groups.

Although matching makes patients similar in different groups as randomization does, it only controls factors that are known to affect the outcome and cannot control the effect of those unknown factors. Another thing is that only a few variables can be matched at a time. Controls will be difficult to find with so many criteria to be met, and overmatching may occur. Finally, the effect of the matched factors on the outcome cannot be evaluated.

### **17.5.2.3 Restriction**

Patients are limited to a narrow range of characteristics to ensure characteristics of the included subjects are consistent among groups under comparison. For example, when evaluating the effect of a new drug on blood pressure, male patients aged 20–60 with a blood pressure of 140–180 mmHg were enrolled. Such restriction certainly increases the homogeneity of patients, but with the sacrifice of

generalizability. The representation of patients in the study may no longer be possible. Moreover, another method is needed to learn the effect of the drug on females, ages, and blood pressure out of that range.

#### **17.5.2.4 Stratification**

Data can be stratified into several subgroups, and results of patients with similar characters are presented. Factors of stratification are those possible confounding factors similar to matching variables such as age, gender, or smoking. The effect of each category on prognosis is then analyzed to discover whether there is confounding in the crude effect. Stratification is a common way to recognize and control confounding. It should be noticed that stratifying many factors at a time may result in having too few patients in some subgroups.

#### **17.5.2.5 Standardization**

Patients from different places or times may have different age or gender distributions. The comparison of two rates would be affected when there are factors that have a strong association with the outcome. Here, the standardized rate can be applied to make the comparison without bias by giving equal weight to the factors. Standardized mortality ratio (SMR) is usually used in prognosis studies. A detailed definition of SMR is discussed in previous chapter.

#### **17.5.2.6 Multivariable Analysis**

In most cases, the prognosis of disease is influenced by many factors. Those factors may be related to one another besides the outcome. Moreover, modifications of effect among factors may also exist. Multivariable analysis is the only way that provides us a comprehensive way to consider the relationship between variables and the outcome, as well as the joint effect of the variables. It deals with many variables simultaneously and can pick up variables that affect prognosis independently from all those in the model. The contribution of each factor to the outcome can also be figured out. In prognosis study, COX proportional hazards model is usually applied in case of time-to-event analysis. Logistic regression is used more often in case-control design in which outcome is dichotomous.

The limitation is that the process of multivariable analysis is just like a “black box.” Multiple comparisons, statistical power, and correlation between variables such as multicollinearity may have an effect on the result. The conclusion may be misleading and, even worse, it is much more difficult to learn where it happens. Therefore, multivariable analysis is usually used after matching or stratification.



**17.5.2.7 Other Methods to Control Bias in Prognosis**

In a prognosis study, it is effective to enroll patients with better compliance to reduce the possibility of loss to follow-up. A clear definition of outcome and use of double-blinding when deciding the occurrence of an outcome will be helpful to minimize measurement bias.

# Chapter 18

## Nosocomial Infections



Zhijiang Zhang

### Key Points

- Nosocomial infections, also termed as “healthcare associated infections”, are infections acquired in health-care facilities. For patients, the infections should not be present or incubating at admission.
- Nosocomial infections can be categorized as endogenous infections and exogenous infections according to types of reservoirs. There are two types of exogenous infections: iatrogenic infections and cross infections.
- The epidemic process of nosocomial infections is usually described with 3 terms: (1) Source of infection; (2) Route of transmission; and (3) Susceptible population.
- Nosocomial infections are related to both patient factors, such as immunocompromise, and medical procedures that is implemented by health professionals at hospitals. It is the responsibility of all health professionals to reduce nosocomial infections.

### 18.1 Introduction

Nosocomial infections constitute a serious threat to the global health. Acquiring infections in health-care settings adds to the patient’s functional disability and emotional stress and in turn affect the patient’s quality of life. Furthermore, nosocomial infections can lead to excessive mortality among hospitalized patients. More than that, nosocomial infections may be disseminated from health-care settings to the general public if not appropriately controlled, and threaten the health of total population.

Nosocomial infections may lead to considerable economic costs. The increased consumption of drugs, the use of additional laboratory and other diagnostic

---

Z. Zhang (✉)  
School of Public Health, Wuhan University, Wuhan, China

procedures, and the need for isolation contribute significantly to the direct costs. The increased length of hospital stay also increases the indirect costs due to work hours lost. Nosocomial infections constitute an economic burden both for the individual patients and for the public health.

## 18.2 Definition and Diagnostic Standards

Nosocomial infections, also termed as “health-care-associated infections,” are infections acquired in health-care facilities, usually during hospital care after admission. To be considered to be nosocomial, the infections cannot be present or incubating at admission. Those infections acquired in the hospital but appearing after discharge, and the occupational infections among staff of the hospital are also considered to be nosocomial.

According to the “Diagnostic Standard for Nosocomial Infections’ issued by the National Health Planning Commission (formerly Ministry of Health of China), the following infections are considered to be nosocomial:

1. For those infections without a clearly defined incubation period, occurrence of infection beyond 48 h since admission; for those infections with a clearly defined incubation period, occurrence of infection beyond the average length of incubation period since admission;
2. The infection is directly related to the last hospital stay;
3. Appearance of new infection in sites other than the original ones (except for metastatic lesions caused by pyemia) or isolation of new pathogens in addition to the known pathogens during hospital stay (excluding the possibility of pollution or previous coexistence of infections);
4. Neonatal infections acquired during or after delivery;
5. Latent infections reactivated by diagnostic or therapeutic procedures, for example, herpes virus and *Mycobacterium tuberculosis*;
6. Infections acquired while working at the hospital as medical staff.

The following cases are not considered to be nosocomial:

1. Bacterial colonization without inflammation in open wounds of skin and mucous membrane;
2. Inflammations caused by trauma or non-biological factors;
3. Infection of the newborn through the placenta (onset of the disease within 48 h after birth), for example, herpes simplex virus, toxoplasmosis, or chicken pox;
4. Acute attack of the original chronic infections during hospital stays.

## **18.3 Nosocomial Infection Sites**

There are many potential body sites for the occurrence of nosocomial infections. The following are the most frequent sites for nosocomial infections.

### ***18.3.1 Surgical Sites***

Surgical sites are subject to nosocomial infections. The infections are usually acquired during the surgical operation. The diagnosis is mainly based on clinical criteria: purulent discharge or spreading cellulitis around the wound or the insertion site of the drain. There are varieties of possible infecting microorganisms for surgical sites nosocomial infections, depending on location and aggressiveness of the surgery, patients' immunity status, and antibiotics use. The level of contamination during the surgical procedure is one of the most important risk factors for surgical site infections. Other risk factors for surgical site infections are the surgical procedures per se, the level of asepsis, as well as the virulence of the infecting microorganisms and concomitant infections at other body sites.

### ***18.3.2 Respiratory System***

Patients with several diseases are at high risk of nosocomial pneumonia while hospitalized. The most frequently reported nosocomial pneumonia is among patients on ventilators. Monitoring of clinical manifestation and using radiological imaging support the diagnosis of pneumonia. Specimen investigation can improve specificity for the diagnosis. Infecting microorganisms may be either endogenous, for example, from nose, throat, or stomach, or exogenous, for example, from contaminated equipment. Patients with decreased level of consciousness are also at higher risk for nosocomial pneumonia. Children are susceptible to viral bronchiolitis, while the elderly are vulnerable to influenza and secondary bacterial pneumonia.

### ***18.3.3 Bacteremia***

The incidence of nosocomial bacteremia is low, but its case fatality rate is high – over 50% for some microorganisms. When infection occurs at the entry site of the device, it may be visible. If bacteremia is caused by the microorganisms colonizing the device within the vessel, it may be invisible. In the case of catheterization, the length of catheter, level of asepsis for the insertion procedures, and duration of catheter care are important factors influencing the risk of nosocomial bacteremia.

### **18.3.4 Urinary Tract**

Urinary infections are the most frequently reported nosocomial infections. Compared with nosocomial infections in surgical sites, pneumonia or bacteremia, urinary infections cause less morbidity but occasionally lead to bacteremia and death. Urinary infections can be diagnosed through quantitative urine culture ( $>10^5/\text{mL}$ ). The microorganisms responsible may be acquired from the patient's gut flora or from the health-care facilities.

## **18.4 Microorganisms**

The infecting microorganisms in nosocomial infections may be bacteria, virus, parasites, or fungi, dependent on the patient populations, medical and surgical interventions, implemented nosocomial infection control programs, and health-care settings.

### **18.4.1 Normal Microorganisms in Nosocomial Infections**

#### **18.4.1.1 Bacteria**

Bacteria are among the most frequently reported pathogens in nosocomial infections. These can be commensal bacteria. Infection occurs when immunity of the host is compromised. These can also be pathogenic bacteria, which lead to nosocomial infections when introduced regardless of the immunity status of the host. Staphylococci, pseudomonads, and *Escherichia coli* are the three pathogens of great concern for nosocomial infections.

#### **18.4.1.2 Viruses**

Many viruses can cause nosocomial infections. For example, the hepatitis B virus can be transmitted through invasive medical procedures, for example, transfusions, dialysis, and injections. Enteroviruses may be transmitted by the fecal-oral route. SARS-CoV-2 may be transmitted by respiratory droplet and aerosol.

#### **18.4.1.3 Parasites and Fungi**

Some fungi and parasites may cause infections among hospitalized patients with compromised immunity or undergoing extended antibiotic treatment. Risks of fungi infection increase for hospitalized patients when renovating aging hospitals. The

infecting pathogens can be *Aspergillus* spp., *Candida albicans*, or *Cryptococcus neoformans*.

### ***18.4.2 Antimicrobial Resistance and Nosocomial Infections***

Due to the inappropriate and uncontrolled use of antimicrobial agents, varieties of microorganisms, including bacteria, viruses, fungi, and parasites that can cause infections in humans, animals, or plants, no longer respond to antimicrobial agents that used to be effective. Antimicrobial resistance constitutes a global concern and is especially a problem for nosocomial infections. In health-care settings, resistant microorganisms have larger capability to spread. Patients undergoing surgery, cancer chemotherapy, and transplantation are at high risk of infections with resistant microorganisms. Genetic mutations of the antimicrobial-resistant microorganisms are thus more likely to spread between hospitalized patients in health-care settings. Restriction on antimicrobial consumption plays a role in the control of nosocomial infections.

## **18.5 Categories of Nosocomial Infections**

Nosocomial infection can be categorized as endogenous infection and exogenous infection according to types of reservoirs.

### ***18.5.1 Endogenous Infections***

Endogenous infections occur when microorganisms that cause nosocomial infections are already present within the body. For example, a patient undergoing chemotherapy has a compromised immunity and the dormant tuberculosis becomes reactivated and infects the patient.

### ***18.5.2 Exogenous Infections***

Microorganisms that cause nosocomial infections are transmitted from outside the patient. There are two types of exogenous infections.

*Iatrogenic infections:* The infections are caused by the contamination of medical instruments, equipment, supplies, and sanitary materials used in health care or by the poor sterilization, for example, microorganisms in water, damp environment, and contaminated devices.

*Cross infections:* Microorganisms are transmitted between patients, member of staff, or visitors through direct or indirect contact.

## **18.6 Epidemic Process of Nosocomial Infection**

The epidemic process refers to the development and spread of nosocomial infections within the health facilities. The three terms frequently used to describe the epidemic process for infectious diseases, that is, source of infection, route of transmission, and susceptible population, are used here to describe the epidemic process of nosocomial infections. In view of the significant difference in epidemic process between endogenous infection and exogenous infection, the following details pertain primarily to the latter.

### ***18.6.1 Source of Infection***

Source of infection refers to the natural habitat of microorganisms which may cause nosocomial infection. Patients are one of the most important sources of nosocomial infections. Other important sources of nosocomial infection may be carriers, wet environment, mouse, arthropods, mosquito, and others. For more details, refer to Chap. 11.

### ***18.6.2 Route of Transmission***

Route of transmission refers to the spread of infecting microorganisms directly or through the environment to another person. The important routes of transmission for nosocomial infections may be direct contact, transfusion or infusion of other medical products, intramuscular injection or other invasive medical procedures, airborne transmission, waterborne transmission, arthropod-borne transmission, vertical transmission, and others. For more details, refer to Chap. 11.

### ***18.6.3 Susceptible Population***

The susceptibility varies for patients according to their age, gender, immunity, as well as medical procedures they are undergoing. Patients with compromised immunity are among the most susceptible populations. For more details, refer to Chap. 11.

## **18.7 Prevention of Nosocomial Infections**

The risk of nosocomial infections is affected by both patient factors, such as immunity status, and health-care settings and medical procedures that elevate the likelihood of infection. It is the duty of all health professionals to prevent nosocomial infections.

### ***18.7.1 Preventing Human-to-Human Transmission***

#### **18.7.1.1 Hand Decontamination**

Maintaining hand hygiene is important for reducing nosocomial infections. For handwashing in hospitals, it requires running water, soap, and drying facilities. For hand disinfection, proper antiseptic is required. The procedures of handwashing/disinfection vary for medical procedures that patients will receive or have undergone.

Adherence to hand decontamination is frequently suboptimal. There may be a variety of reasons, including high frequency of patient contact, allergies to hand decontamination products, low perceived risk of infection, low awareness of hand decontamination procedures, lack of time required to complete the hand decontamination procedures, and lack of accessible equipment. The facilities must have policies to evaluate and manage this problem.

#### **18.7.1.2 Clothing**

An outfit, usually a white coat, is needed for staff. In special areas, uniform trousers and gown are required. An outfit must be changed in the case of being exposed to blood or other body fluid. In aseptic units and operating rooms, dedicated shoes should be used as well as caps or hoods that can cover the hair.

#### **18.7.1.3 Masks**

In operating room, staff wear masks to protect patients. When caring for immune-compromised patients, staff must wear masks. For infections which can be transmitted by the air, patients must wear masks when not isolated. When approaching patients with airborne infections, staff must wear masks to avoid being infected.

#### **18.7.1.4 Gloves**

Staff must wear sterile gloves in surgery or other invasive procedures. To protect the immune-compromised patients, staff must wear sterile gloves. Whenever hands are



likely to be contaminated, non-sterile gloves should be worn before patient contacts. When caring for patients with infections that can be transmitted by direct contact or respiratory droplets, non-sterile gloves should be worn to protect the staff. Wash hands with running water after removing or changing gloves.

### **18.7.1.5 Safe Injection and Other Skin-Piercing Practice**

Injection or other skin-piercing procedures increases the risk of infection transmission between patients. It is required to use sterile needle and syringe, prevent contamination of medications, and eliminate unnecessary injections.

## ***18.7.2 Preventing Transmission from Environment***

The hospital environment can be classified into five types of zones according to the possibility of contamination, required level of asepsis, and risk of infection:

Zone A: It is clean areas without patient contact, for example, administrative office and library;

Zone B: It is areas possibly contaminated by microorganisms, for example, passageway and lab;

Zone C: It is areas with patient contact and microbial contamination, for example patients' room and bathroom;

Zone D: It is passageways for clinicians and patients in the designated zone for the diagnosis of infectious respiratory diseases. The entrance of clinicians' passageway connects to clean zones, while the entrance of patients' passageway connects to contaminated zones.

Zone E: It is buffer areas between clean and contaminated zones.

To minimize the microorganisms from environment, cleaning, disinfecting, and sterilizing must be used appropriately for each type of zone.

### **18.7.2.1 Routine Cleaning**

Routine cleaning is scheduled to make the environment visibly clean. The frequency of routine cleaning and cleaning agents need to be specified for all types of reused equipment/devices used in the health-care settings and all areas in the hospital.

### **18.7.2.2 Disinfection of Equipment**

The purpose of disinfection is to remove microorganisms without complete sterilization. The disinfectants must be nonvolatile, free from irritating smells, and not

harmful to equipment or persons. The disinfection procedures must kill or remove the targeted microorganisms.

### **18.7.2.3 Sterilization**

Sterilization is to destruct all microorganisms on the medical devices. Sterilization is performed for those medical devices used to penetrate sterile body surface, as well as medications and parenteral fluids.

## **18.8 Surveillance of Nosocomial Infections**

Surveillance is a program designed to monitor nosocomial infections in a continuous, systematic, and long-term manner. It plays an important role on identifying the early signs of local problems and the evaluation of the effectiveness of nosocomial infection control policy.

### ***18.8.1 Objectives of Surveillance Programs***

The ultimate aim of surveillance is to reduce the burden of nosocomial infections in the local area and alleviate the costs.

The specific objectives of a nosocomial infections surveillance program usually include the following:

1. To monitor trends in nosocomial infections, including incidence, prevalence, and distribution of nosocomial infections;
2. To evaluate the effectiveness of prevention programs, and to adjust the currently ongoing prevention programs accordingly;
3. To recognize sources of nosocomial infections, particularly in situations of an outbreak, and to take immediate actions to control transmission;
4. To find aspects for improvement in the local nosocomial infections control programs.

### ***18.8.2 Implementation of Surveillance Programs***

Surveillance programs can be implemented at the hospital level. Involving partners include the infection control practitioner, physician, nurse, lab staff, director, and administrator. Before implementing a surveillance program, the partners need to decide the following:

1. Which patients and units to be monitored;
2. What type of infections and relevant information to collect;
3. The time period of the surveillance and frequency of monitoring;
4. Data collection and information retrieve methods;
5. Methods for data management and analysis;
6. Methods for information dissemination and feedback collection;
7. Methods for maintaining confidentiality.

Besides at the hospital level, surveillance of nosocomial infections may also be implemented at the levels of local, regional, national, or international networks. On a confidential basis, hospitals may share data with other facilities in the local, national, or international network for the purpose of improving nosocomial infection control programs.

### ***18.8.3 Evaluation of Surveillance Program***

Evaluation of the surveillance programs is necessary. Maintaining contacts with the surveillance program staff can also help maintain their compliance with the guidance of the surveillance program. An evaluation usually includes the following:

#### **18.8.3.1 Strategy Evaluation**

Evaluate whether the surveillance program has the following quality: simplicity, flexibility, acceptance, sensitiveness, effectiveness, and efficiency.

Evaluation can be undertaken by means of field survey, focus group, or interview.

#### **18.8.3.2 Feedback Evaluation**

Feedback evaluation aims to address the following specific issues:

1. Is confidentiality respected during the implementation of the program? Is maintaining confidentiality compatible with data dissemination required for the purpose of infection control?
2. Are the results of the evaluation widely shared internally within the units and externally between facilities in the network?
3. Is the population under surveillance well representative of the target population?
4. Is risk adjustment/stratification appropriately used?
5. Is the length of the surveillance period sufficient to draw a conclusion?

### **18.8.3.3 Evaluation of Data Quality**

The denominator and nominator used in calculating the incidence or prevalence of nosocomial infections need to be periodically evaluated in terms of exhaustiveness (missing patients), completeness (missing data), and correctness (data error).

# Chapter 19

## Epidemiology Design in Clinical Research



Yi Wang

### Key Points

- The process of clinical research include forming research questions, selecting proper epidemiology design, collecting clinical data and doing statistical analysis, preparing reports to publish. The PICO process could help investigators to form research question. For different types of questions, there are different appropriate epidemiological designs could be selected.
- Some checklist items should be included in clinical research reports. The checklists composed reporting guidelines. The reporting guidelines for clinical research include STROBE for observational studies, STARD for diagnostic/prognostic studies, CONSORT for clinical trials, PRISMA for systematic reviews/Meta-analysis, et al.
- Real-world studies are used widely in clinical research now. Real-world studies are different from randomized control trials in many aspects. RCT provides evidence for clinical practice guideline recommendation and real-world study tests if guideline recommendation is practicable.

In previous chapters, we introduced different types of epidemiological study designs and discussed their strengths and weaknesses. The overall strategy of clinical research is the same as that utilized in other areas of epidemiology: observation of incidences between groups and then extrapolation based on any differences. In clinical research studies, the defining characteristics of groups can be symptoms, signs, diseases, diagnostic procedures, or disease treatment. The discussion that follows in this chapter will consequently summarize and integrate the core epidemiological topics involved in the previous chapters. We will concentrate mainly on observational studies, diagnostic/prognostic studies, clinical trials, and systematic reviews looking for the general principles frequently applied in clinical research.

---

Y. Wang (✉)

School of Public Health and Management, Wenzhou Medical University, Wenzhou, China

e-mail: [wang.yi@wmu.edu.cn](mailto:wang.yi@wmu.edu.cn)

Clinical epidemiological studies prefer randomized groups to epidemiological studies. Firstly, the “exposure” in clinical research is usually a treatment approach that tends to be more randomized than the exposures considered in most epidemiological studies (e.g., tobacco or alcohol consumption, diet, or personal or environmental characteristics). Secondly, the results uncovered in clinical epidemiological studies, such as disease progression, complications, or mortality, are comparatively frequently found in the patient groups being compared, making randomized studies more feasible. Thirdly, the potential for confounding is particularly high in clinical epidemiological studies where there is no randomized grouping. In a large number of nonrandomized treatment studies in which a correlation has been detected, it is unclear whether changes in patients’ risk of disease progression, complications, or death are related to the type of treatment they receive.

## **19.1 Design and Implementation of Clinical Research**

Good epidemiological studies are complicated to design and conduct, and the interpretation of their consequences and findings is not as straightforward as researchers would like it to be. So, what can we do to make the best research design? How can we make the most of the clinical practice information available to us? When we read or write clinical research papers, the central question we need to answer is “Are the findings valid?”. If a relationship between predicted values and results is reported by researchers, is this true? If they come up empty, can we trust them? Or could there be another interpretation of the findings, namely, chance, bias, and/or confusion? When investigators perform the clinical study, they almost certainly must read individual articles and reports, especially the guidelines for clinical research published in professional journals. They may produce some of their scientific papers when they are engaged in clinical research.

The first stage in establishing clinical research is to design the study issue that you aim to answer. Then, you would utilize several epidemiological designs to try to uncover the explanation. Therefore, first of all, investigators need to focus on what clinical questions should be answered. Secondly, researchers should also consider whether the research design was suitable for replying to the questions raised. A highly practical approach is very necessary, which we will outline in the parts that follow.

### ***19.1.1 Forming Research Questions***

Usually, clinical problems could be divided into two categories: background question and foreground question. The background question is about the general knowledge of disease such as “what is tuberculous pericarditis?” and “what are the antituberculosis drugs?” The foreground question is the actual problems that

physicians or surgeons encounter in the process of diagnosis and treatment of patients. For instance, physicians want to know “how the utility of the ascites adenosine deaminase (ADA) in the diagnosis of the tuberculous peritonitis?” and “does tuberculous pericarditis require glucocorticoid treatment?”. The foreground question is the main problem in clinical practice. According to different process of clinical practice, there are four types of foreground questions: treatment question, diagnosis question, etiology question, and prognosis question. When physicians are confronted with clinical foreground questions, they want to design a study to solve these problems, they could use “PICO” process to decompose the research problems into specific research content.

In “PICO” process, “P” is an abbreviation for patients or population. It refers to the clinical features of research patients or population. “I” is an abbreviation for intervention or exposure. It means treatment measures or exposure issues that are concerned. “C” is an abbreviation for comparison. It means the control measure and usually means the “gold standard” if it is a diagnostic study. “O” is an abbreviation for outcome. It is the outcome indicators that the research focused on. In Table 19.1, it listed some examples of how to use “PICO” framework to form research question in four different question types.

### ***19.1.2 Commonly Used Epidemiological Design in Clinical Research***

The most important point in clinical research is to identify the research question. If clinicians have proposed a research question, there are different epidemiological designs that could be selected to help answer these questions. Commonly used epidemiological design in clinical research includes cross-sectional study, case-control study, cohort study, nonrandomized controlled trials, and randomized controlled clinical trials. In general, prospective study design has the most content, the most complex methods, and the most representative of epidemiological data analysis. Figure 19.1 illustrates how to decide which epidemiological design to select.

Previously, we have introduced four types of foreground questions. For different types of questions, different epidemiological designs could be selected. For the evaluation of treatment efficacy, the most appropriate study design is a randomized control trial (RCT). However, it is very difficult to carry out an RCT in real clinical practice, especially conducted it in a multicenter study. Besides RCT, a cohort study, case-control study, case report could be selected for the treatment questions. For prognosis question, the most appropriate study design is a cohort study. In Table 19.2, it listed best design could select for each type of foreground questions.

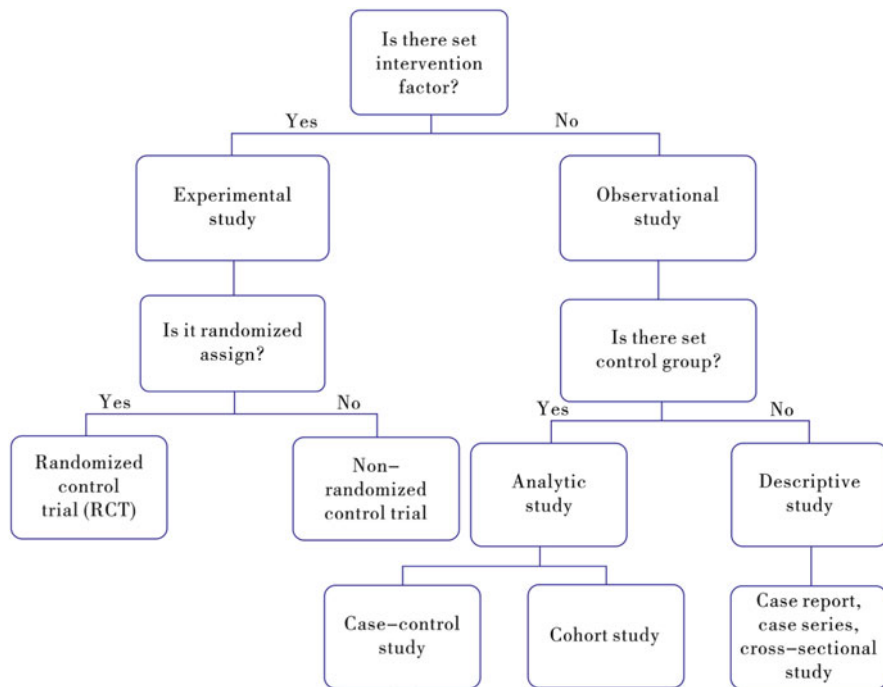
**Table 19.1** Examples of application of the PICO process in clinical research

Question type	Clinical question	PICO	Research content
Treatment question	Do patients with tuberculous pericarditis need to be treated with glucocorticoids?	P: adult patients with tuberculous pericarditis I: antituberculosis + glucocorticoid C: antituberculosis O: death	Can glucocorticoids reduce the risk of death in adult patients with tuberculous pericarditis?
Diagnosis question	What is the utility of the ascites adenosine deaminase (ADA) in the diagnosis of the tuberculous peritonitis?	P: patients with celiac effusion I: ascites adenosine deaminase examination C: gold standard diagnostic method for tuberculous peritonitis O: validity of diagnosis for tuberculous peritonitis	How about the sensitivity and specificity of the ascites adenosine deaminase examination for the diagnosis of tuberculous peritonitis?
Etiology question	How about the risk of a vegetarian suffering from tuberculosis?	P: adults I: vegetarian diet C: common diet O: tuberculosis	Are vegetarians at more risk to develop tuberculosis than nonvegetarians?
Prognosis question	Do patients with tuberculous pericarditis could develop into constrictive pericarditis?	P: tuberculous pericarditis patients O: constrictive pericarditis (there usually have no "I" and "C" in the prognosis question)	What is the probability of patients with tuberculous pericarditis to develop constrictive in the future? What are the prognostic factors to predict patients with coarctation?

### 19.1.3 Collection and Analysis of Clinical Research Data

Routine clinical epidemiological data are primarily those with health, illness, and clinical services that are routinely collected in the population for other uses, such as patient data routinely collected in hospitals. In addition, some are disposable, irregularly collected data, but others are data from specific epidemiological studies (such as prospective studies), such as the purpose of the analysis is not to answer the original questions of the study, but to use the data to explore new non-primary research questions. They are collectively referred to as routine clinical epidemiological data. Routine clinical epidemiological data analysis steps: (1) Analyze the time frame of the data and the characteristics of the variables; (2) Ask questions that can be explored to determine the final research question; (3) Compare with the best





**Fig. 19.1** The flow chart of selection of epidemiological design in clinical research

**Table 19.2** Best study design to select for different clinical questions

Question type	Best study design
Treatment question	RCT
Adverse effect of treatment question	RCT
Diagnosis question	Cross-sectional study
Prognosis question	Cohort study
Etiology question	Cohort study, case-control study

research design, check data for “research design” defects; (4) Estimate the necessary indicators and their confidence intervals; (5) Analyze other possible biases in the data (selection bias, information bias, and confounding bias); (6) Integrated design flaws, biases, and results, and draw conclusions on research issues.

### 19.1.3.1 Data Collection

The collection and management of clinical research data is the main content in the design and implementation phase of clinical research. It involves management techniques and skills and requires researchers to invest a great deal of time and effort. The collection and management of clinical data is a process. Understanding



**Fig. 19.2** Process of clinical research data collection

the content of each link in the process and the relationship between the various links can be a good job in clinical research design and implementation. The process of collecting and organizing clinical data is characterized by a linear process, multi-stage, and multi-link. Figure 19.2 shows this process.

Clinical research is the process of collecting, sorting, storage, analysis, and evaluation of clinical data. It is a linear process and can only be carried out in a sequential manner. The starting point of clinical research is the research object. The researchers have to use various technical methods to obtain clinical data from the research object, then transfer the clinical data to the case report form (CRF), and then transfer the clinical data from the CRF to the database, and prepare for the later statistical analysis and evaluation work. In the process of clinical data collection, the completion of CRF filling and the establishment of the database are treated as a two-phased landmark in the collection of clinical research data. The completion of the CRF design marks important progress in the design of the clinical research implementation plan. The establishment of a database is a key link between the collation and storage of clinical data. There are sophisticated methods and techniques, and the workload is large. The quantity and quality of the input data are guaranteed, and it is organized for later data analysis. The data completed by the CRF, the quality, and the completion of the database are the main evaluation indicators for evaluating the implementation phase of the clinical research organization.

Besides collecting clinical data from practice clinics or hospitals, clinicians could collect clinical data from some open access databases, like SEER (surveillance, epidemiology, and end results). Here, we will introduce an open access database commonly used in clinical oncology research – TCGA. The Cancer Genome Atlas (TCGA) program was launched in 2005 to apply the latest genomic analysis technology. In particular, the whole genome sequencing technology, in-depth understanding of cancer gene changes, and promoting the discovery of new cancer treatment programs, diagnostic methods and prevention strategies, plans to draw a wide range of tumor types and tumor subtypes, multidimensional map of the key genome changes. Moreover, all the data can be shared for free in scientific practice. The TCGA plans to collect sample data for 11,000 patients and 33 cancers (Table 19.3). In 2015, the amount of data collected and generated by the TCGA program had reached 20PB, including 10 million mutations. Investigators could choose interested cancers to download the gene and clinical information and analyze them for particular purpose.

**Table 19.3** TCGA plan cancer sample distribution (33 cancers, 11,000 patients)

Cancer symbol	Type of cancer	Number of samples	Cancer symbol	Type of cancer	Number of samples
BRCA	Breast invasive carcinoma	1097	THYM	Thymoma	124
KIRC	Kidney renal clear cell carcinoma	536	SKCM	Skin cutaneous melanoma	470
LUAD	Lung adenocarcinoma	521	ACC	Adrenocortical carcinoma	80
THCA	Thyroid carcinoma	507	DLBC	Lymphoid neoplasm diffuse Large B-cell lymphoma	48
PRAD	Prostate adenocarcinoma	498	LGG	Brain lower-grade glioma	516
LIHC	Liver hepatocellular carcinoma	377	LAML	Acute myeloid leukemia	200
LUSC	Lung squamous cell carcinoma	504	MESO	Mesothelioma	87
HNSC	Head and neck squamous cell carcinoma	528	OV	Ovarian serous Cystadenocarcinoma	586
COAD	Colon adenocarcinoma	461	TGCT	Testicular germ cell tumors	150
UCEC	Uterine corpus endometrial carcinoma	548	UCS	Uterine carcinosarcoma	57
KIRP	Kidney renal papillary cell carcinoma	291	UVM	Uveal melanoma	80
STAD	Stomach adenocarcinoma	443	CESC	Cervical squamous cell Carcinoma and endocervical adenocarcinoma	307
KICH	Kidney chromophobe	66	PCPG	Pheochromocytoma and paraganglioma	179
BLCA	Bladder urothelial carcinoma	373	SARC	Sarcoma	261
ESCA	Esophageal multiforme	185	CHOL	Cholangiocarcinoma	36
READ	Rectum adenocarcinoma	171	GBM	Glioblastoma multiforme	528
PAAD	Pancreatic adenocarcinoma	185			

The TCGA research team has collected and generated various types of histological and genetic data for these cancers, including gene expression, exon expression, small RNA expression, copy number changes (CNV), single nucleotide polymorphism (SNP), loss of heterozygosity (LOH), gene mutations, DNA methylation, and

protein expression. The clinical information includes patient's basic geographic information, treatment method, historical or clinical stage, survival status, and so on.

By analyzing the cancer genome information to understand the mechanism of cancer development and discover cancer markers and drug effect on gene targets, it can provide support for the accurate diagnosis and treatment of cancer. The TCGA plans to collect data on a large number of cancer genomes and clinical phenotypes. There are potential molecular markers and drug targets for cancer that need to be tapped. Scientific data management programs provide protection for cancer genome research. The practical exploration of cancer genomic map planned in data management can provide a reference to the development and implementation of large-scale scientific programs such as precision medicine and data-driven collaborative research models.

### 19.1.3.2 Data Analysis

Unlike basic medical research, clinical epidemiological research is an applied research conducted in the population to quantitatively explore the general rule of disease, health, and clinical practice, and the results can be directly applied to clinical practice. Clinical epidemiological studies need to be based on specific research questions, selecting designs, controlling bias, collecting data, and then analyzing the data to quantitatively answer research questions. Therefore, data analysis is an important and indispensable part of clinical epidemiology research. Clinical questions generally include etiology questions, diagnosis questions, treatment questions, and prognosis questions. The purpose of data analysis is to scientifically and quantitatively answer these practical questions. Data analysis must have a clear purpose for analysis. The common purpose of clinical epidemiology is shown in Table 19.4. Clearly studied issues are the premise of data analysis. After the question is clarified, it is necessary to put forward a specific and clear analysis purpose. Its content generally includes the following: (1) describe the change in the number of subjects, (2) variable classification and data sorting, (3) describe and compare baseline data between groups, (4) estimate the frequency of outcome events, (5) estimate the magnitude of the effect, (6) the confidence interval of the estimated effect, (7) identify and control the confounding, (8) identify and measure effect modification effects, (9) identify and measure dose-response relationships, (10) other

**Table 19.4** The purpose of clinical epidemiological data analysis

	Research purpose
1.	Estimating relevant statistical indicators such as relative risk and sensitivity
2.	Estimating the confidence interval for the statistical indicator
3.	Controlling for possible confounders
4.	Analysis of dose-response relationships
5.	Analysis of possible effect modification factors
6.	Analysis of other possible biases

analysis. Although the design principles of different studies could variate and the clinical problems, the purpose, contents, and methods of analysis are also different, the analysis of other research data can be regarded as one or more components of the data analysis of prospective research.

In addition, the estimation of this indicator must simultaneously control possible confounding factors. In a randomized control trial, investigators could thoroughly control confounding factors through randomized assign research objects to different groups. However, in observational studies (such as nonrandomized allocation trial studies, cohort studies, and case-control studies), the most effective and feasible method for controlling confounding factors is multivariate regression analysis. The premise of controlling confounding is to recognize possible confounding factors, and the baseline data with confounding factors were collected at the beginning of the study. Other analytical purposes may include identifying and measuring effect modifiers, identifying and describing dose-response relationships, and analyzing and controlling other possible biases.

### ***19.1.4 Preparing Papers for Publication***

The following recommendations provide a guide for investigators to prepare clinical research reports:

#### **19.1.4.1 Choose Target Journal(s)**

Selecting the intended journal category for publication is always the first step for the researchers while drafting the report. When selecting target journals, investigators should take the following issues into account.

##### **How High to Aim**

A question that researchers will face is what height your paper may reach. Many investigators believe that five top general medical journals are particularly attractive carriers for their article: *The Lancet*, *The New England Journal of Medicine (NEJM)*, *The Journal of the American Medical Association (JAMA)*, *Annals of Internal Medicine*, and *British Medicine Journal (BMJ)*. Investigators often have the question of whether their research is suitable for these or other famous journals. It is also challenging to predict success (or lack of success) for experienced researchers, which makes it very difficult to select the most appropriate target journals. In general, if it is adequate for you to seriously consider contributing to a famous journal, it means that your research has been well designed and implemented, and beyond that, you are so courageous. More commonly, the internal debate (within you or your survey team) may be whether you first submitted to a secondary journal (e.g.,

possibly a top journal in your subspecialty field) or are more likely to accept a journal with a lower reputation for your manuscript. One advantage of foresight is that sharp comments may help you improve your article. It is unusual to make substantial improvements to your manuscript based on the opinions of reviewers, but it can happen in some cases. Therefore, if multiple submissions do not make you tired, and receiving too many rejection letters from magazines does not hurt your self-esteem, set a higher goal. If your mental state is irritable and fragile, then choosing a journal with a less high impact factor is more likely to accept your manuscript at the first submission.

### Selecting a Journal with a Fondness for Researcher Topic

Some certain topics or fields are often favored by certain journals. If research in an area that is closely related to your study has previously been published by a journal, it is eligible to be one of your chosen targets. In the meantime, the lack of articles in your field or using your methodology provides the information you should search for elsewhere.

### Tailoring Content to the Target Journal

The majority of the manuscripts you write will be reporting on the clinical investigations you conduct. However, in some cases, investigators may write a paper that focuses more on research methods. These papers explored some issues, such as the best research design, measurement methods, or results interpretation. Researchers can consider publishing their manuscripts in these three types of journals: general medical journals, subspecialty journals, and methodologically oriented journals. Many articles on clinical research are likely to be published in multiple target journals.

### Tailoring Format to the Target Journal

Almost every journal has its own format requirements. Most of these requirements are relatively trivial (e.g., section titles or reference citation styles), and when you are ready to submit your manuscript, you must modify it as required. Of course, there are other more important issues that researchers should address them early on.

#### **19.1.4.2 Choose a Clear Message**

The work may be exceedingly complicated, and a definite result might not be obvious. Until a clear message is determined, investigators must continue to evaluate the essence of the results. Just considering what the reader is going to take away from

a single point, what is that point? After determining the information, the investigator needs to craft the introduction so that readers will be convinced of the significance of the research. Your research should be presented to readers as a narrative. The reader's curiosity should be piqued by the introduction, satisfied by the outcome, and reinforced by the discussion, which should highlight how significant the finding is.

#### **19.1.4.3 Achieve High Quality in Writing**

Here are some tips for creating a high-quality manuscript. (1) Use the active voice: Passive writing is a well-established medical practice. Although the passive voice makes writing more awkward and difficult to understand, adds extra words, and makes the work lose some power, this tradition still exists. The use of active voice is advised in all current publications on writing quality from a variety of nonmedical professions as well as writing suggestions offered by the top medical magazines. (2) Delete unnecessary words: Unnecessary words are utilized by medical writers. Eliminating these terms makes the writing more direct and clearer to read. Journal articles must adhere to strict space restrictions as well. Use as few adverbs and adjectival phrases as you possibly can. (3) Avoid using the verb "to be": The verb "to be" and the passive voice frequently has the same impact. It robs the writing of vigor and energy. (4) Keep paragraphs short: Each paragraph in the article should not exceed five sentences. Clarity will be considered a priority.

#### ***19.1.5 Common Problems in Clinical Research Design***

According to the analysis of clinical papers published, there are six major problems in the design of clinical research programs.

1. Researchers are unclear about the design scheme adopted by their institute  
After investigating the research questions that are of interest to the researcher, the researchers must determine the research design plan based on the results they expect, the strength of the causal connection, and the feasibility.
2. Unclear definitions of primary and secondary study end points  
A study generally has only one primary end point, but there can be several secondary end points. In some studies, it is not correct to write only what the study end point is, but not to distinguish between the primary and secondary end points. In the design plan, the primary and secondary end points of the study need to be clearly written out, which is conducive to the establishment of hypotheses and the calculation of sample size. It is not possible to write all the indicators in parallel and regardless of primary and secondary levels.
3. Have no scientific hypothesis

The entire research process is the process of testing the hypothesis. The hypothesis is based on scientific research problems. Based on the hypothesis, the researcher can determine the sample size, follow-up time, and determine the type of quantitative collection, and statistical methods. Researchers need to establish reasonable assumptions based on the primary end point after the design of the study.

4. Have no controls or unreasonable controls

The four principles of clinical trial design are “random,” “control,” “blinding method,” and “repetition.” The establishment and selection of appropriate controls for the control group is an important part of the research design. Researchers can design blanks, placebo controls, positive standard controls, and other controls based on the purpose of the study. Parallel control is best for the same period.

5. Nominally a randomized control study but not an actual randomized grouping

The stochastic method includes two layers of meanings: One is the generation of random distribution sequences and the other is the concealment of random distribution sequence schemes. If the scheme is not hidden, randomization may be disrupted, resulting in selection bias and measurement bias. The purpose of blinding is to make the research executor not know the specific stochastic method and do not know whether the research object in accordance with the random sequence belongs to the experimental group or belongs to the control group so that complete randomization can be achieved. Researchers should be trained in systematic clinical epidemiology or clinical research methodologies. The significance of clinical research method training is similar to that of standardized training for clinicians. It is an important foundational work and requires the support and efforts of all parties.

6. No sample size was calculated

In addition to exploratory research, because no basic data cannot calculate the sample size, a general clinical study needs to estimate the sample size in advance. A too small sample size will result in large sampling errors, resulting in poor representativeness and poor reproducibility. Researchers should pay attention to the significance of sample size and know the concept of power. As long as there is a consciousness of calculating sample size, the calculation process is not a problem. Now there are many statistical software applications for calculating sample size.

## 19.2 Reporting Guidelines for Clinical Research Reports

### 19.2.1 *Observational Studies Reporting Guidelines*

In September 2004, the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) working group was founded and convened in the United Kingdom to draft the normative meeting of the observational research report. After many revisions, a list containing 22 items (STROBE statement) was released in



2007, which is divided into 6 major aspects including the title, abstract, introduction, method, result, and discussion. Eighteen of these items are applicable to all three major observational research designs (cross-sectional, case-control, and cohort design), and the remaining four are specially used for cohort, case-control, or cross-sectional design, respectively. A new STROBE statement extension was released in 2014 by the Lancet Infectious Diseases. Enhance Molecular Epidemiology Reporting for Infectious Diseases (STROME-ID). The goal is to provide guidelines for effective scientific reporting of molecular epidemiology research to urge authors to take particular hazards to reliable inference into account. The official website (<http://www.strobe-statement.org>) offers free downloads of the STROBE statement and STROME-ID statement.

### ***19.2.2 Diagnostic/Prognostic Studies Reporting Guidelines***

In 2003, Bossuyt PM, an authoritative expert in the field of diagnostic tests, convened a group of experts to establish the STARD group to develop a report on the diagnostic accuracy study – Standards for Reporting of Diagnostic Accuracy (STARD), which was used to standardize diagnostic test studies. In order to solve new problems in diagnostic tests, streamline the reporting process, increase its applicability, and align STARD with CONSORT-2010, Bossuyt PM again convened a group of experts in 2015, including epidemiologists, statisticians, evidence-based medicine experts, doctors, editors, and journalists, and 85 people, based on the STARD 2003, developed a STARD 2015 guide using document research, drafting entries, expert surveys, and group discussions. The STARD statement could be downloaded from <http://www.stard-statement.org>.

### ***19.2.3 Clinical Trials Reporting Guidelines***

The CONSORT (Consolidated Standards of Reporting Trials) declaration was created by a team of scientists and editors to enhance the caliber of RCT reporting. It was revised in 2001 after being initially published in 1996. The statement includes a flow diagram and checklist that researchers can employ to report an RCT. The CONSORT declaration has received support from several top medical publications and influential international editorial organizations. The claim makes it easier to evaluate and understand RCTs critically. The ideas underpinning the CONSORT statement were clarified and expanded upon during the 2001 CONSORT revision to assist researchers and others in writing or evaluating trial reports. In 2001, the CONSORT declaration and an essay explaining and expanding upon it were both published. The CONSORT statement was further amended following an expert meeting in January 2007 and is now available as the CONSORT 2010 Statement. This revision clarifies and updates the prior checklist's language and includes

suggestions for subjects like selective outcome reporting bias which have just recently gained attention. This explanation and elaboration paper, which has also undergone substantial revision, aims to improve the use, comprehension, and diffusion of the CONSORT declaration. Each newly added and revised checklist item is explained along with its purpose, with illustrations of effective reporting and, if available, references to pertinent empirical research. There are several flow diagram examples provided. Resources to aid with randomized trial reporting include the CONSORT 2010 Statement, the updated explanatory and elaboration paper, and the related website (CONSORT, <http://www.consort-statement.org>).

### ***19.2.4 Systematic Reviews Reporting Guidelines***

In 1996, the Quality of Reporting of Meta-Analyses (QUROM) guide was published, which focused on the quality of reporting of randomized controlled trial meta-analysis, which was the earliest reporting specification for systematic review/meta-analysis quality. In the classic monograph “Systematic reviews in health care: meta-analysis in context,” QUROM was recommended as the “gold standard” for evaluating the quality of systematic reviews/meta-analysis reports. The items involved in the QUROM report specification are divided into 6 parts and 18 items, including the title, abstract, introduction, method, result, and discussion. The results section includes the search process and gives the reasons for identifying, including, and excluding randomized controlled trials and exclusions. In 2009, QUROM updated Systematic Reviews and Meta-Analyses for Protocols (PRISMA) in order to improve the quality of systematic review, and meta-analysis article reports. PRISMA is more comprehensive and complete than the QUROM developed in the past. It has a wide range of applications, not only for meta-analysis, but also for systemic evaluation; not only for systematic evaluation of randomized controlled trials but also as a basic specification for evaluation reports of other types of research systems. The PRISMA Reporting Guide consists of a list of 27 items, a four-phase flow chart, and detailed explanations and explanations of relevant items. All these materials could be downloaded from <http://www.prisma-statement.org>.

## **19.3 Real-World Study**

The collection and storage of enormous volumes of health-related data via computers, mobile devices, wearables, and other biosensors have been expanding quickly. This information has the potential to help us design and carry out clinical research in the health-care sector more effectively to provide answers to previously

unanswerable problems. Additionally, we are better equipped to examine these data and apply the findings to the development and approval of medical products as a consequence of the development of sophisticated, new analytical skills. As a result, a growing number of clinical trial designs are being designed which have been derived from real-world data and evidence.

### ***19.3.1 Definition of Real-World Study***

Real-world studies (RWS) originate from effective clinical trials and refer to the nonrandom choice of interventions based on the patient's actual condition and willingness to perform long-term evaluations based on the larger sample size (covering a representatively larger number of subjects). Focus on meaningful outcome indicators to further evaluate the external effectiveness and safety of interventions. The RWS covers a wide range of areas and can be used for diagnosis, prognosis, etiology, in addition to curative studies. The RWS focuses on the effectiveness of research, namely, the size of the evaluation interventions in the real clinical environment. RWS can also be used to evaluate the cost-effectiveness of different health interventions.

### ***19.3.2 The Difference Between RWS and RCT***

Although RWS is very different from RCT (Table 19.5), RCT and RWS are not contradictory or alternative relations of opposition but are complementary and forming a connecting link between the preceding and the following. RCT is the highest level of evidence-based medicine; is the “gold standard” of clinical trial design. It is a recommendation to formulate corresponding treatments guidelines based on RCT, which tells doctors that they can do and should do, rather than have to do it. Therefore, the guideline cannot replace clinical practice. It needs RWS as an effective supplement, and RWS can be used to determine the true benefits, risks, and therapeutic value in clinical practice, so that clinical research conclusions will return to the real world after RCT. Therefore, RWS and RCT are not antagonistic, but complementary to each other.

Overall, RCT provides evidence for clinical practice guideline recommendation. RWS tests if guideline recommendation is practicable, whether answers clinical questions and summarizes treatment recommendations, then returns to clinical practice. Among them, RWS plays a more and more important role nowadays.

**Table 19.5** Differences between RCT and RWS

	RCT	RWS
Research purposes	The outcome of an ideal situation	The outcome of the real situation
Research environment	Strictly controlled conditions	Actual clinical conditions
Research design	Randomized controlled trials	Nonrandomized Control/Effective Randomized Control/Observational Study
Research scheme	Cannot be changed after the program is fixed	Can be adjusted according to clinical practice
Research object	Strict inclusion/exclusion criteria and good homogeneity	Inclusion/exclusion criteria are loose, and diversity is good
Sample size	Minimum sample size	As much as possible
Control group	Standard treatment/placebo	Effective treatment/no placebo
Research data	Designed before the start of the trial, prospectively collecting data	Forward-looking/retrospectively collecting data according to need
Study outcomes	Most recent indicators	Mostly long-term indicators
Follow-up time	Short	Long
Follow-up completion	Better	Uncertain
Ethical review	Need	Need
Clinical registration	Need	Need
Internal effectiveness and safety	Good	Poor
External effectiveness and safety	Poor	Good
Difficulty of work	Relatively small difficulty	Very difficult
Evaluation angle	Evaluating effectiveness from a medical perspective (efficacy)	Evaluate the effect from the patient (effectiveness)

# References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A (2015) Global cancer statistics, 2012 [J]. *CA Cancer J Clin* 65(2):87–108
2. Ministry of Health of the People's Republic of China (2003) *China health statistics yearbook 2003* [M]. Peking Union Medical College Press, Beijing
3. National Health Commission of the People's Republic of China (2021) *China health statistics yearbook 2021* [M]. Peking Union Medical College Press, Beijing
4. American Cancer Society (2017) *Cancer facts and figures 2017* [M]. American Cancer Society, Atlanta
5. Zheng X, Wu K, Song M et al (2019) Yogurt consumption and risk of conventional and serrated precursors of colorectal cancer [J]. *Gut*. pii: gutjnl-2019-318374
6. Monson R (1980) *Occupational epidemiology* [M]. CRC Press, Boca Ration, FL
7. Zheng T, Boffetta P, Boyle P (2011) *Epidemiology and biostatistics* [M]. International Prevention Research Institute, Lyon
8. Du H, Li L, Bennett D, Guo Y et al (2016) Fresh fruit consumption and major cardiovascular disease in China [J]. *N Engl J Med*. 374(14):1332–1343
9. Klemetti A, Saxén L (1967) Prospective versus retrospective approach in the search for environmental causes of malformations. *Am J Public Health Nations Health* 57(12):2071–2075
10. Rothman KJ, Greenland S (2005) *Causation and causal inference in epidemiology* [J]. *Am J Public Health* 95(Suppl 1):S144–S150
11. Shen HB, Qi XY (eds) (2018) *Epidemiology* [M], 9th edn. People' Medical Publishing House, Beijing
12. Strom BL, Kimmel SE, Hennessy S (eds) (2013) *Textbook of pharmacoepidemiology* [M], 2nd edn. Wiley-Blackwell, Hoboken, NJ