Jan Van den Broeck · Jonathan R. Brestoff
Editors

Epidemiology: Principles and Practical Guidelines

Springer

Epidemiology: Principles and Practical Guidelines

Jan Van den Broeck • Jonathan R. Brestoff
Editors

Epidemiology: Principles and Practical Guidelines

Springer

Editors

Jan Van den Broeck
Centre for International Health
University of Bergen
Bergen, Norway

Jonathan R. Brestoff
Perelman School of Medicine
University of Pennsylvania
Philadelphia, PA, USA

ISBN 978-94-007-5988-6 ISBN 978-94-007-5989-3 (eBook)
DOI 10.1007/978-94-007-5989-3
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013936791

Printed on acid-free paper

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

To Yacine and Allison

Foreword

In theory, there should be no difference between theory and practice, but in practice, there is.

–William T. Harbaugh

The essence of the discipline of epidemiology is the application of relatively subtle and abstract concepts to the practical challenges we face in the conception, design, conduct and reporting of research on human health and disease. As a teacher of epidemiology to undergraduate and post-graduate students in public health, medicine, dentistry and an expanding range of other disciplines in the health and social sciences, I have grappled for over two decades with the challenge of helping students link core epidemiological concepts such as bias and confounding with the practical challenges of completing a research project to the standard required for publication. Indeed, the major challenge in research supervision is to bring students to the level where they move seamlessly between theoretical and practical issues in formulating and refining their research questions.

I am not aware of any textbook in epidemiology that bridges this chasm between theoretical and practical issues as effectively and comprehensively as *Epidemiology: Principles and Practical Guidelines*. The authors, Jan Van den Broeck, Jonathan Brestoff and colleagues, take the reader on an excursion over 31 chapters from the conception of research questions to the reporting of study findings, including en route core issues in contemporary practice and topics, such as data cleaning, that are neglected in virtually all textbooks and poorly covered in the literature. This is a book for both students and experienced practitioners.

The world is not currently under-supplied with epidemiology textbooks. Vision and imagination were required to embark on writing this textbook which so effectively fills an important gap in this crowded market. I salute the lead authors, Jan Van den Broeck, a former faculty member of our Department, and Jonathan Brestoff, a recent graduate from our MPH programme, for this achievement.

In particular, I am honoured to acknowledge Jan Van den Broeck's dedication and skill as a teacher and practitioner of epidemiology honed in the class room and in fieldwork over two decades and reflected in the scholarship displayed in this outstanding textbook.

Ivan Perry, MD, M.Sc, Ph.D, FRCP, FRCPI, MFPHM, MFPHMI
Professor and Head of the Department of Epidemiology & Public Health
University College Cork – National University of Ireland, Cork
Cork, Ireland

Preface

Toward integrated learning of epidemiology

The field of epidemiology is growing rapidly and in need of effective practical guides for the development, implementation, and interpretation of research involving human subjects.

There are many epidemiology textbooks covering a range of approaches, but almost all leave the reader asking: "How do I actually conduct a research project in epidemiology?" Our many attempts to address this question in the classroom inspired us to develop a text that supports research practice, and this book is the end product of that inspiration. Unlike conventional textbooks in epidemiology, we break down the research process into discrete stages and steps that help one to develop, conduct, and report epidemiological research.

In doing so, we have adopted a decidedly operational approach and contextualize discussions of research practice with theory and ethics, so that students and professionals from all academic backgrounds may develop a deep appreciation for how to conduct and interpret epidemiological research. Along the way, readers will develop skills to:

- Search for and appraise literature critically
- Develop important research questions
- Design, plan, and implement studies to address those questions
- Develop proposals to obtain funding
- Perform and interpret fundamental statistical estimations, tests, and models
- Consider the ethical implications of all stages of research
- Report findings in publications
- Advocate for change in the public health setting

In our treatment of these topics and others, we integrate discussions of scientific, ethical, and practical aspects of health research. Indeed, at all stages of the research process, each of these aspects directly influences study validity. Consequently, this textbook expands concerns about study validity beyond the usual foci on study design and statistics to include other issues that may also affect the quality and relevance of published findings, examples of which are quality control activities, measurement standardization, data management, and data cleaning. As we discuss each of these topics, we emphasize practical field methods and suggest potential solutions to common problems that tend to arise during study implementation.

The recognition that many different scientific, ethical, and practical aspects interact to affect study quality represents one of the major originalities of the approach taken in this book. As we progress, we discuss a variety of emerging views and innovations in the field that will change the way epidemiology is practiced. We believe that this approach will best situate you, the reader, to conduct epidemiological research. Indeed, epidemiology is a discipline in motion, and this textbook aims to reflect this dynamism and keep pace with its momentum.

As you read, we encourage you to use the text as a step-by-step tool to build your own research project. The experiences of planning and conducting a research study are as important as the underlying epidemiological theory and statistics. As a practicing or future health researcher, you have your own motivations and passions, and we hope this textbook will help you to use your interests to inspire your learning and professional development.

Jan Van den Broeck
Jonathan R. Brestoff

Acknowledgments

This book is the product of 5 years worth of work, and we could not have completed it without the support and contributions of our many talented colleagues, who have written and/or reviewed the content of this book. These individuals come from several academic institutions around the world, and to them we are grateful. We wish to extend a special word of thanks to Rainford Wilks (Epidemiology Research Unit, University of the West Indies, Jamaica), Ivan Perry (Department of Epidemiology and Public Health, University College Cork, Ireland), and Rune Nilsen (Centre for International Health, University of Bergen, Norway), all of whom greatly facilitated and supported this book from its earliest inception. We also thank Olli Miettinen for stimulating discussions and feedback about concepts and terminology. In addition to those who contributed to the content of this book, we thank the US-Ireland Alliance, Trina Vargo (its founder and president), and the George J. Mitchell Scholarship for enabling us to meet at University College Cork and thereby seeding what has become a wonderful collaboration. And most importantly, we are very grateful for the crucial support given to us by our closest family members, and we are happy and proud to mention Yacine, Naomi, Wim, Lucie, and Guido as well as Allison, Malcolm, Nick, and Lois.

Contents

Part III Study Conduct

Part IV Study Analysis

Part V Study Reporting

Contributors

Matthew Baum, D.Phil, M.Sc., M.Sc. Harvard School of Medicine, Boston, USA

Jonathan R. Brestoff, MPH Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Nora Becker Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Jan Van den Broeck, M.D., Ph.D. Centre for International Health, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway

Meera Chhagan, Ph.D., FCPaed Department of Paediatrics, University of KwaZulu-Natal, Durban, South Africa

Jutta Dierkes, Ph.D. Institute of Internal Medicine, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway

Ingunn Engebretsen, Ph.D. Centre for International Health, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway

Department of Child and Adolescent Psychiatry, Haukeland University Hospital, Bergen, Norway

Lars Thore Fadnes, Ph.D. Centre for International Health, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway

Ari Friedman, M.Sc. Department of Health Care Management, Wharton School of Business and Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Ingvild Fossgard Sandøy, M.D., Ph.D. Centre for International Health, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway

Cora Grant, MPH Department of Epidemiology and Public Health, University College Cork, Cork, Ireland

Michael C. Hoaglin Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Shuaib Kauchali, MPhil, FCPaed Department of Paediatrics, University of KwaZulu-Natal, Durban, South Africa

Catherine Kaulfuss, M.Sc. Centre for International Health, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway

Eimear Keane, MPH Department of Epidemiology and Public Health, University College Cork, Cork, Ireland

Emma A. Meagher, M.D. Department of Medicine, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Victoria Nankabirwa, Ph.D. Department of Paediatrics and Child Health, School of Medicine, College of Health Sciences, Makerere University, Kampala, Uganda

Vundli Ramokolo, M.Sc. South African Medical Research Council, Cape Town, South Africa

Bjarne Robberstad, M.Sc., Ph.D. Centre for International Health, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway

Bandit Thinkhamrop, Ph.D. Department of Biostatistics and Demography, Faculty of Public Health, Khon Kaen University, Khon Kaen, Thailand

Thorkild Tylleskär, M.D., Ph.D., MA Centre for International Health, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway

Douladel Willie, M.Sc. Epidemiology Research Unit, and Department of Child Health and Psychiatry, University of the West Indies, Mona, Jamaica

Tracey S. Ziolek, M.Sc., CIP Institutional Review Board, University of Pennsylvania, Philadelphia, PA, USA

Part I

Introduction

Chapter 1: Definition and Scope of Epidemiology
Chapter 2: Basic Concepts in Epidemiology
Chapter 3: Roots and Future of Epidemiology

Definition and Scope of Epidemiology 1

Jan Van den Broeck, Jonathan R. Brestoff,
and Matthew Baum

I'm not sure there is a bottom line…. Continued discussion and dialogue on these important subjects, a whole range of subjects, is important.

John Snow

Abstract

Epidemiology is a methodological discipline offering principles and practical guidelines for the creation of new quantitative evidence about health-related phenomena. Its aim is to contribute to knowledge in support of clinical medicine and community medicine. Epidemiological research uses scientific methods, in which empirical evidence is obtained from a study population to make inferences about a target population. In this chapter we first establish a definition of epidemiology and describe the wide scope of epidemiology in terms of its subject domains, types of research topics, types of study designs, and range of research activities that occur from a study's inception to its publication. Since epidemiology concerns both 'scientific studies' and 'particularistic fact-finding investigations,' we further orient the reader to the scope of epidemiology through a discussion of these. We then introduce general epidemiological principles that health researchers

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

M. Baum, D.Phil., M.Sc., M.Sc.
Harvard School of Medicine, Boston, MA, USA

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_1,

should continuously keep in mind during the planning, gathering and presentation of the empirical evidence. All of these topics are pursued in more depth in the chapters that follow.

1.1 Definitions of Epidemiology

Although the term 'epidemiology' is relatively recent, some roots of modern epidemiology go back to ancient times (*See:* Chap. 3). It has been defined variously, and it may be surprising to learn that consensus on these definitions has not yet been reached. According to the broadest of views (Miettinen 2011a, b), epidemiology is a synonym of *community medicine*. According to this view, one can practice epidemiology by doing epidemiologic research or by practicing public health outside clinical care settings. A community health educator, for example, could be an epidemiologist.

Mostly outside North America there is a competing view that defines epidemiology more narrowly as the methodological discipline that provides quantitative research methods to *public health*, a term that refers to both community and clinical medicine. When epidemiology first became a distinct discipline in the nineteenth century, it focused on the methods of creating quantitative evidence about illnesses encountered in communities at large or in variously defined sub-settings (clinical care settings are one such category of sub-settings). This long-sustained emphasis on methodology is reflected in current public health practice settings (clinical or other), where epidemiologists are hired mostly because they are specialists in quantitative research methods. The editorial view for this text is in line with the latter, more practical view of epidemiology.

1.1.1 Unpacking the Definition of Epidemiology

The definition of epidemiology proposed in this book is:

Epidemiology

The (1) methodological discipline providing (2) principles and practical guidelines for (3) creating new quantitative evidence (4) relevant for clinical and community medicine

(1) Methodological discipline

Epidemiology provides methods for conducting research. The knowledge achieved by the research adds substance to public health, not to epidemiology itself. The exception may be operational research that is done to investigate relative efficiency, validity, and ethics of various research procedures and methods themselves.

(2) Principles and practical guidelines

Methods proposed in epidemiology have three dimensions:

- A scientific dimension relating to validity, reproducibility, and verifiability
- An ethical dimension relating to rights and values
- A practical dimension relating to administration and strategy

Hence epidemiology provides intertwined scientific, ethical, and practical principles and guidelines. Discussions of these three dimensions are integrated throughout this textbook.

(3) Creating new quantitative evidence

In epidemiology evidence is created through research that uses scientific methods. There is disagreement among epidemiologists as to whether epidemiology should concern only evidence produced using quantitative research methods (using statistics as the principle form of evidence) or both quantitative and qualitative research methods. Here, we propose the view that the use of quantitative methods is definitional to epidemiology, but we also recognize that qualitative research methods can support and enhance several aspects of epidemiologic research and often can be codified in a manner that permits quantitative analysis.

(4) Relevant for clinical and community medicine

The new evidence created should have relevance to clinical medicine and/or community medicine, the two of which are typically considered to be non-mutually exclusive elements of public health (Fig. 1.1; also *See:* Textbox 1.1). Clinical medicine concerns the health and well-being of individuals through the direct care of a recognized provider, typically in a clinical setting. Community medicine concerns the health and well-being of a population through intervention at a community level. The distinction between clinical and community medicine is exemplified by a hypothetical clinic established specifically to provide obstetric services to an underserved community and partially funded by a government program aimed at improving public health in such areas. This clinic can be said to practice both clinical and community medicine by caring directly for its patients and by directing their efforts at a community in need of obstetric services, respectively.

Hint

A highly useful exercise, especially in the planning stages of a research project, is to consider how achieving knowledge about the topic under study might have implications for clinical medicine, community medicine, or both. Making theoretical or substantiated arguments about the potential clinical or community health benefits helps to motivate research teams, to increase the likelihood of obtaining funding, and to communicate the importance of one's work to others.

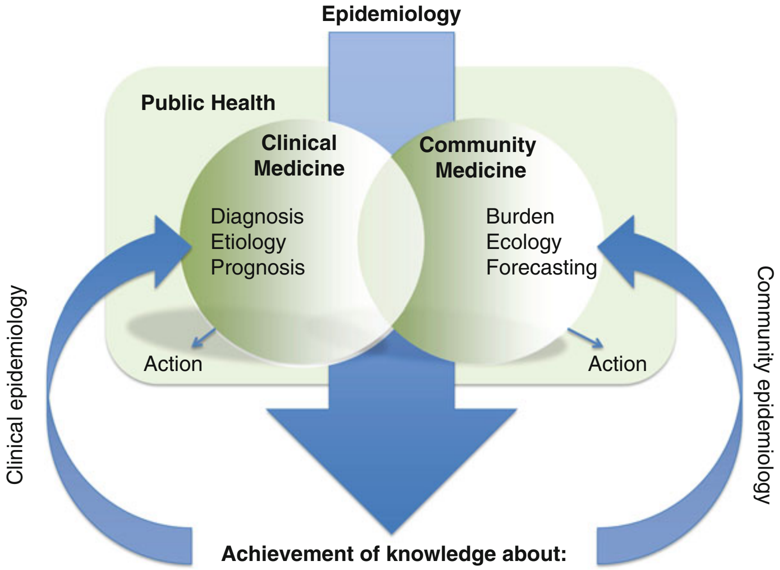


Fig. 1.1 *The place of epidemiology within clinical and community medicine*. The methodological discipline of epidemiology is employed to achieve knowledge about clinical or community medicine (represented by *green spheres*). The spheres of clinical and community medicine are overlapping to illustrate that they interact. Clinical and community medicine interactions often benefit both individuals and society (e.g., obstetric clinic described in the text body), but sometimes they can come into conflict (*See:* Textbox 1.1). Studies directed at typical activities of clinical medicine (diagnosis, etiognosis, and prognosis) or community medicine (burden assessment, ecology, and forecasting) are referred to as clinical epidemiology or community epidemiology, respectively. Knowledge achieved by community and clinical epidemiology is used to inform actions that contribute to the betterment of public health

Textbox 1.1 Resources: When Clinical and Community Medicine Come into Conflict

In clinical medicine, a doctor is expected to act always in the best interests of the specific person seeking medical care. But when there are limited resources, one might need to balance the interests of the individual with that of the community. While also a concern for insurance companies, this conflict becomes salient in publicly funded health care systems. A specific intervention might lead to the best outcome for the individual, for example, but be twice as expensive as another intervention that would lead to a slightly less-good outcome. The public system might decide to offer the latter, less-good treatment to that individual so that more individuals could have access, thereby maximizing community but not individual health. How to and who should do the weighing of individual versus community health interests is a topic of continual debate.

1.1.2 How Similar Are Clinical Medicine, Community Medicine, and Epidemiology?

Many epidemiologists combine research with clinical practice or community health practice. In such activities, one assesses health-related states or risks of individual patients and populations, respectively. These assessments do not necessarily follow the secure, slow path of scientific research. In fact, they rarely do. Instead, they are made mostly using clinical skills and public health skills that are quite different from the skills used in epidemiology.

As pointed out by Miettinen (2011b), there is currently no scientific knowledge base of medical practice in a form that is immediately applicable and useful. Clinical skills, as far as diagnosis is concerned, are a mixture of experience, common sense, intuition, knowledge of differential diagnoses, and ability to find and use literature and decision algorithms. In contrast with a trial, there is not a single hypothesis that is going to be tested using clinical trial methodology. Instead, the diagnostic knowledge is to be created by the clinician-diagnostician by very quickly eliminating thousands of rivaling diagnostic hypotheses, a process which is achieved by quickly proceeding to next questions asked to the patient, examination of a chosen next physical sign, and doing appropriate laboratory tests. This process may seem rather unstructured and unpredictable, but in reality it tends to have a remarkable and useful reproducibility.

Similarly, in community health practice, many of the acquired insights do not come from epidemiological studies and rarely do they come from causally-oriented epidemiological studies. Instead, the public health practitioner often uses assessment methods that do not follow a design prescribed by the current epidemiological paradigms. These methods rather proceed in a way similar to clinical diagnosis, avoiding any formal hypothesis testing and trying to make sense out of a complex and unique situation. Like clinical diagnosis, the reproducibility and speed is often remarkable and useful, and the usefulness strongly depends on intuition and experience mixed with more technical 'qualitative' investigation skills. Examples of such assessment methods are SWOT analysis (Strengths, Weaknesses, Opportunities, and Threats), situation root-cause analysis, in-depth interview, focus group discussions, and rapid assessment procedures. Some of these methods are collectively labeled 'qualitative research methods.' Whilst their usefulness in community health practice is readily apparent, these methods cannot be considered to constitute a type of epidemiological study design because they do not follow a quantitative scientific paradigm.

That having been said, both clinical skills and qualitative assessment skills can be of crucial value in an epidemiological study. The need for persons with clinical skills in clinical epidemiological studies needs no argumentation. Qualitative assessment skills can provide fast and useful information in the design stage and preparation stage of an epidemiological study. Examples include assessments about possible confounders and effect modifiers; likely refusal rates and reasons; local concepts and terminology about diseases; and culturally appropriate wordings of questions and response options (Kauchali et al. 2004).

1.2 The Scope of Epidemiology

The scope of any discipline depends on one's point of view, and epidemiology is no exception. Four frequently used points of view are described below. Although none alone fully elucidates the scope of epidemiology, when taken together they are cornerstones quite useful for the task. The points of view concern:

1. The spectrum of research activities for which epidemiology provides principles and guidelines: study design, conduct, analysis, interpretation, and reporting
2. The range of subject domains within medicine served by epidemiology: infectious diseases, chronic non-communicable diseases, health services, etc.
3. The typology of research questions that are usually addressed by epidemiology: descriptive versus analytical studies
4. The general study design types used in epidemiology: experimental, quasi-experimental, and observational studies

1.2.1 Spectrum of Research Activities in Epidemiology

Research is a process that proceeds in logical, more-or-less pre-determined steps. Stages of all scientific research are study design, conduct, analysis, and reporting. From this point of view, the scope of epidemiology is the spectrum of scientific, ethical, and practical principles and guidelines that are relevant to the design, conduct, analysis, and interpretation/reporting of research on health-related issues in epidemiologic populations. The sequence in which the research process proceeds is approximately reflected in the structure of this book and is summarized in Table 1.1.

1.2.2 Range of Subject Domains Within Medicine Served by Epidemiology

In the mid-nineteenth century epidemiology was mainly concerned with epidemic infectious diseases (*See:* Chap. 2). Today, epidemiology reaches into domains such as normal and pathological morphology and physiology; infectious and non-infectious diseases; preventive and curative medicine; physical, behavioral, mental, and social health; and genotypic and phenotypic aspects of health and disease in life-course and trans-generational perspectives. Epidemiology provides methods to increase knowledge in both clinical medicine and community medicine, which we see as the basis for making a distinction between *clinical epidemiology* and *community epidemiology.*

The main activities of clinical medicine are diagnosis, etiognosis, intervention, and prognostication. Clinical epidemiology supports these activities by providing methodologies for various types of research studies, as illustrated in Table 1.2. The same table illustrates how activities of community medicine are served by community epidemiology.

Table 1.1 Stages of the research process and their common elements

Study stage	Elements
Design	Proposal and protocol development
	Literature review, as part of study rationale development
	Formulation of general and specific aims
	Choice of general type of study design
	Optimal size of a study
	Identification of the study base and planning to access it
	Choice or development of measures, measurements, outcome measures, outcome parameters and analysis methods
	Planning of ethical oversight and data management
	Design of quality assurance and control protocols
	Fundraising and stakeholder involvement
Conduct	Training and study preparation
	Measurement and measurement standardization
	Establishing and maintaining access to study base
	Implementing data management and data cleaning plans
	Data quality assurance and control activities
	Study governance and coordination
	Interaction with stakeholders during study conduct
Analysis	Preliminary, primary, and secondary analyses
	Controlling for bias and confounding
	Subgroup and meta-analysis if relevant
Reporting	Interpretation of results
	Scientific writing
	Reporting data quality
	Dissemination of research findings to relevant stakeholders

Table 1.2 The supporting role of epidemiology for clinical and community medicine

Clinical medicine activities	Clinical epidemiology provides methods for, *inter alia*
Diagnosis and etiognosis	Diagnostic classification; descriptive or analytical studies on disease occurrence
Intervention	Trials and observational studies on treatment effects
Prognostication	Studies on disease outcomes
Community medicine activities	Community epidemiology provides methods for, *inter alia*
Screening, surveillance, health profiling	Surveys, studies on screening and surveillance methods, modeling of disease spread, outbreak investigation
Public health services and interventions	Community intervention studies (including prevention research)
Evaluation of health services and interventions	Health service utilization studies, cost-effectiveness studies

1.2.3 Typology of Research Questions in Epidemiology

The scope of epidemiology is often thought of in terms of the types of research questions addressed. There are many ways of categorizing epidemiological research questions, and a detailed typology is discussed in Chap. 4 (General Study Objectives).

Table 1.3 Frequent objectives of epidemiological research

Classification	Frequent objectives
Descriptive studies (phenomenological orientation)	Estimate the burden of illness
	Describe the natural history of illnesses
	Predict the risk of a health related event
	Derive classification of diseases
Analytical studies (causal orientation)	Identify the causes of illness (or protective factors)
	Evaluate interventions

A broad traditional classification scheme distinguishes between *descriptive* and *analytical* research questions and studies (Table 1.3). Descriptive studies investigate phenomena and their relationships without concern for causality. Analytical studies, on the other hand, aim at demonstrating causal links among phenomena. These types of studies investigate the effects of presumed risk factors, also called *exposures* or *determinants*, on health outcomes with a particular concern for demonstrating that reported relationships are free of potential confounders (*See:* Chap. 2).

Health-related phenomena commonly studied using epidemiology are health states or events in individuals, health-related attributes of populations, or characteristics of functional care units. One's interest in a given health-related phenomenon may be its frequency, severity, causes, natural course, response to intervention, complications, risk factors, protective/preventive factors, and other aspects (*See:* Chap. 4). Epidemiologic studies of individuals tend to focus on normal and abnormal morphology and function. Also of interest may be how illnesses secondarily affect subjective experiences, physical/psychological function, and social function e.g., quality of life and wellbeing, both of which are higher-level, multidimensional attributes (*See:* Chap. 10 for more information about quality of life measures).

Population characteristics studied in epidemiology include burdens and inequalities – differences in morbidity, mortality, burdens, risks, effects, etc. (*See:* Chap. 4). One may be tempted to alternatively define epidemiology as the discipline concerned with investigating health inequalities, thereby hinting to its important social-ethical dimension. Indeed, many inequalities are unfair and unacceptable socially and ethically, and research into their existence, causes, and alleviation needs to be supported by a discipline that renders the investigations scientific and efficient and that ensures studies are carried out in full respect of its participants.

1.2.4 The General Study Design Types Used in Epidemiology

In epidemiology the most frequently used traditional (i.e., mainstream) general study designs are considered to be experimental, quasi-experimental, or observational. Examples of each are listed in Table 1.4. In mainstream typology, e*xperimental studies* are those in which the researcher allocates intervention levels in a randomized fashion and then observes and compares the outcome of interest among each randomized arm. In *quasi-experimental studies* the allocation of intervention levels is

Table 1.4 Mainstream typology of general study designs

Study type	Examples
Experimental	Randomized-controlled trial
Quasi-experimental	Non-randomized trial
Observational	Cross-sectional study
	Cohort study
	Case–control study
	Ecological study

non-randomized but otherwise similar to experimental studies. In *observational studies* the participants may or may not undergo interventions, e.g., as prescribed by their health care providers, but the researcher only observes and does not allocate intervention levels in the research context.

Whenever the interest is in the occurrence of events or change of status over time, a *follow-up study* is usually preferable for validity reasons. In such follow-up studies, one can follow the experience of a *cohort* (a group with fixed membership determined by some admission event) or a *dynamic population* (a group with non-fixed membership, where entries and exits occur) over time. In a *cross-sectional study*, one studies a cohort at a single, fixed individual follow-up time (most frequently, follow-up time zero) or a dynamic population around a fixed point in calendar time (e.g., a survey). This cohort or dynamic population is assessed once for their current health-related states of interest and for determinants of interest. In a *case–control study*, individuals with a particular health-related state of interest ('cases') are identified in a cohort or dynamic population (perhaps in a cross-section thereof) and their antecedent experience in terms of presumed-causal factors and presumed confounders is assessed and compared with the time-equivalent past experience of a sample of the target population from which the cases originated ('controls'). *Ecological studies* typically look at concomitant variation of group statistics (of multiple groups) on outcomes and exposures. A more extensive discussion of general study design, with a partly different typology is found in Chap. 6.

1.3 Particularistic Versus Scientific Studies

In planning a study, when the researcher has to specify the target population (*See:* Textbox 1.2), there is often the choice to define the target population with temporal-spatial constraints ('particularistic study') or without such constraints ('scientific study'). Whatever the choice, a group of study subjects will have to be identified in space and time whose characteristics fit the definition of the target population and whose relevant experiences will be observed and measured. In addition, in both cases scientific methods of investigation (including study design) are followed.

In scientific studies, one chooses to define a highly abstract population, such as newly diagnosed adult patients with type 2 diabetes. This choice implies that the researcher expects the generated evidence to be generalizable to all individuals sharing the defined attributes of the abstract population (e.g., any individual with a

Textbox 1.2 Naming and Defining Populations

The term ‘population’ is commonly understood to be synonymous with the term **demographic population**, defined as the inhabitants of a given area; however, a demographic population is only a particular instance of an **epidemiologic population**. Individuals, communities, or institutions that are the focus of attention in epidemiological research constitute epidemiologic populations. They can be defined theoretically as target populations or directly observed as study populations.

Early in the study planning process, one must define a **target population**, the theoretical epidemiologic population about which one seeks to achieve knowledge. The units whose attributes/experiences are the focus of an epidemiologic study can be individuals or groups (e.g., households, villages, etc.). Taking the common case of ‘individuals’ as an example, the specific type of individuals of interest in a particular research study may be further specified by combinations of temporal, spatial, environmental, biological, and behavioral characteristics. The inclusion of temporal and spatial restriction criteria in this specification is not always necessary. If place and time criteria are not part of the definition of the target population (e.g., patients newly diagnosed with type 2 diabetes mellitus), a study will more often be labeled ‘*scientific*’. If place and time criteria are included in the definition of a target population, a study will often be labeled ‘*particularistic*’ (e.g., the infant with protein-energy malnutrition in Bwamanda in 1992). In both cases, the target population includes an abstract type of people.

A **study population** or study sample refers to the collection of observation units on whom data have been or will be collected to make inferences about the target population. A study population can be but is not always a statistically representative sample of all individuals whose individual characteristics fit the definition of the target population. Although an epidemiologist performs measurements on a study population, the purpose of a study is not strictly to learn something about the study population. Explicitly, the aim of the epidemiologist is to achieve knowledge about the target population. For example, in a clinical trial one is not just interested in how an intervention works in the patients involved in the study; rather, the main interest is in knowing something about how *future patients* will react if they receive the intervention.

new diagnosis of type 2 diabetes), irrespective of whether they participated in the study. This is a bold but risky position given that external validity (generalizability) depends on achieving internal validity during the study and other issues of credibility. But there is another more important reason why the position is risky. Perhaps the greatest possible fallacy in epidemiological thinking, and probably the root of most contemporary controversies about the value of epidemiology, is to think that the statistical results (outcome parameter estimates or test statistics) from a scientific study – whether it be an experimental, quasi-experimental, or observational study – represent estimates of 'universally true occurrence relations'.

We are of the view that causally oriented epidemiological studies do not estimate a single, true, abstract, or universally generalizable relation between a study factor and an outcome. They only provide true generalizable evidence on such relations *conditional on an often complex, always particularistic and variable distribution matrix of measured and unmeasured confounders and effect modifiers.* Failure to appreciate this to the fullest can lead to misguided irritations about, for example, the very normal fact that epidemiological studies on the same topic (including clinical trials) often lead to very different or even contradictory results (*See also:* Textbox 25.1).

The set of covariates that underlie a 'true relationship' cannot be expected to be homogeneous in time and space. Modern science has revealed a staggering diversity within and among individuals and populations with respect to constitutional, environmental, and behavioral-social conditions. Moreover, it is now understood that there are large fluctuations in these conditions over brief periods of time (e.g., within the span of just a 1 day). It is not surprising that health states are greatly influenced by changes in these conditions. An equally staggering number of ever-refined subclassifications of health state characteristics are now appreciated. Thus, it becomes increasingly difficult to define a target population and a distribution matrix in a way that consistently replicates statistical study results. Generalizability only holds until the next paradigm shift, and the current rate of paradigm shifts in exposure and disease classification is so high that a new focus is needed in epidemiology. The scientific task of discovery has become a task of quantifying relationships and now is also a task of exploring and 'taming' heterogeneity.

1.4 General Epidemiological Principles

The scientific, ethical, and practical dimensions of epidemiology have led to the development of principles that have a bearing on all or nearly all stages of the research process, and we therefore refer to them as general principles. Decisions about design, execution, and reporting of the research should be geared towards epidemiology's general principles. While many potential candidates for general principles might be identified, Panel 1.1 aims to highlight what we consider to be the most important ones.

Without exception, these principles ultimately derive from ethical considerations, even those concerning validity and efficiency, as it is unethical to conduct a study that will be invalid or that wastes resources unnecessarily. While it is helpful to think about how our general principles relate to the broad ethical principles of respect for autonomy, non-maleficence, beneficence, and justice (Panel 1.2), we present our general principles in the form and degree of specification that we consider most useful for those designing and carrying out epidemiologic research. These general principles will be frequently referred to and further discussed later in the book. Below we provide a basic orientation.

Panel 1.1 General Principles of Epidemiology

- Minimize risk of avoidable, unacceptable harm
- Respect the autonomy of participants
- Respect the privacy of participants and confidentiality of their data
- Minimize burden, preserve safety, and maximize benefit for participants
- Maximize societal relevance
- Contribute minimally biased evidence to the overall pool of evidence on an issue
- Maximize completeness of data for analysis and archiving
- Guarantee verifiability of study procedures
- Pursue parsimony

Panel 1.2 Broad Ethical Principles Relevant to Epidemiology

Respect for autonomy Respecting the capacity of an individual to make an informed un-coerced decision

Beneficence The concept that researchers should mind the welfare of participants

Justice The concept that researchers should act with moral rightness and maintain fairness and justness

Non-Maleficence The concept that researchers should minimize the exposure to potential harm

Minimize risk of avoidable, unacceptable harm: Sometimes misleading when referred to only as “first do no harm,” this principle reflects the obligation not to expose participants to avoidable or unacceptable harm, even if doing so carries significant costs, and to minimize exposure to avoidable risks of harm. Most studies will expose participants to some amount of physical, psychological, economic, or legal risk. The researcher has a responsibility to foresee and minimize exposure to such risks. It is not always clear at what threshold a harm becomes unacceptable or which types of harms are inherently unacceptable; ethics review boards, however, can be useful resources in discussing this question for a given study.

As a clarifying example, this general principle might translate to an obligation to draw blood with sterile needles (minimize risks) and to not draw blood at all if sterile needles cannot be found (risks cannot be minimized below a decent threshold), even if this means the study cannot be conducted. Or, it might translate to an obligation to terminate the study or to withdraw a patient from the study if doing so might avoid significant harm even though early termination might affect the quality of the data. Historically, this concern was established in response to inhumane and harmful studies of the mechanisms or natural history of a disease, such as the Tuskegee Study of untreated Syphilis (*See:* Textbox 16.1).

Respect for the autonomy of participants: This principle protects participants' self-determination, or the ability to make one's own decisions about one's life (like whether to participate in research after having considered the risks and benefits). Seeking informed consent when enrolling participants is one method of respecting autonomy, as is making sure there is no coercion or undue inducement, either of which would invalidate that consent. At some time during the study, moreover, a participant may competently decide to withdraw consent; building in mechanisms for withdrawing from the study would respect that autonomous decision.

Respect for the privacy of participants and confidentiality of their data: Rigorous measures should be taken to ensure the security of identifiable information obtained from participants and to minimize the intrusiveness of research. Designing a study that uses the method of information gathering that is least intrusive into the private lives of participants while still enabling valid data collection might be a specification of this principle.

Minimize burden, preserve safety, and maximize benefit for participants: This principle combines aspects of beneficence and justice. Namely, studies should maximize the cost/benefit ratio for participants and ensure that the group undertaking the burden/risks of research are also benefiting from the research; one group should not take all the risks while another benefits. In theory, keeping this principle in mind will also help ensure that the study population is generalizable to the target population. Although many epidemiologic studies may not have significant safety concerns, there will at absolute minimum be the burdens of time spent and of possible adverse effects of participation. Execution of this principle, especially in populations that differ in culture from the researchers', may require particularly careful consideration of or perhaps even preliminary research on what is considered burdensome or beneficial to the participants. While this principle contains similar elements to the first principle, we think it helps to keep them conceptually separate.

Maximize societal relevance: Research should address a health issue relevant to the target population and have the realistic possibility of bringing society closer to improving related health outcomes. Research on methods may directly support the maximization of societal relevance. Community engagement (and engagement with other stakeholders) in the design, conduct, and dissemination of research may be a specification of this principle. Because even the best conducted research will have little impact if it is poorly or too narrowly communicated, a specification of this principle might be publishing clear, well written papers in appropriate journals and ensuring that research is disseminated in forms able to be understood by and meaningful to the different types of stakeholders.

Contribute minimally biased evidence to the overall pool of evidence on an issue: All research studies will have limits to their internal validity and generalizability. This principle represents the duty to maximize the benefits of research by working to push back these limits. Avoiding conflicts of interest (e.g., making sure the research is truly independent) might be a specification of this principle. Identifying weaknesses in existing research on a topic and designing a study that does not replicate these weaknesses may be another specification.

Maximize completeness of data for analysis and archiving: A principle again aimed at maximizing the usefulness of research, this one allows the epidemiology community to address multiple questions with the same database and to return to the database in the future to address newly raised questions. Following this principle helps to avoid the new burdens, risks, and costs of collecting new data. Completeness of data also enhances the precision and sometimes the unbiasedness of study findings.

Guarantee verifiability: It is essential to be able to verify past studies, especially if conflicting results emerge. Rigorously detailing methods, documentation of data quality aspects and archiving samples can be thought of as specifications of this principle. If a study cannot be verified, it cannot be trusted, and thus cannot be used to benefit society.

Pursue parsimony: This principle reflects a duty not to expend resources (time, money, personnel, etc.) needlessly or to expose participants to needless risks or burdens. Specifications might be to enroll only as many participants and continue collecting data only as long as necessary to reach a scientifically valid and rigorous answer to the specific research question being investigated.

To illustrate that these principles cut across many stages of the research process, let us consider the principle of maximizing data completeness and the many points in a study at which incomplete data may arise. During a clinical follow-up study there are many opportunities to lose participants and to miss or lose information. When planning for the number of participants to recruit, one must try to

Textbox 1.3 Epidemiology and Its Link to Culture and Politics

There is an overarching social-ethical dimension to epidemiologic research that inevitably links it to culture and politics. This has implications for the choice of research questions and the fair allocation of resources to competing research questions. Investigators, research institutions, and companies have an ethical obligation to mind their potential contributions to society. Indeed, they are often required to adhere to international and national policies aimed at reducing unfairness. Likewise, policy makers must support epidemiological research on inequalities and they must take into account the foreseen effects of any policy decision on health inequalities, at all levels from local to global. To do so, they will need a trustworthy knowledge-base on health inequalities provided by epidemiologic research. Epidemiologists are consequently important stakeholders of the socio-political process.

identify a target number (or range of numbers) desired for analysis, and then account for expected rates of attrition and refusal to determine the number of participants to recruit. Researchers may lack the necessary resources to boost lagging enrollment rates or to prolong the enrollment period. After enrollment there may be a few late exclusions of participants who appear not to be eligible after all, and some subjects may withdraw their participation or be unable or unwilling to accommodate certain measurements. Of the recorded data, some may prove to be outliers or to be the result of contamination, and re-measurement may be impossible. The laborious task of data entry may be incomplete, and source documents may be lost or damaged. In preparations for analysis, data transformations may be incomplete (e.g., some data transformations cannot handle negative data), and finally, some analysis methods (e.g., multiple linear regression) can only use records with complete data on all the variables in the model. As a result of these potential problems and others, discrepancies between the targeted sample size and number of samples analyzed are common; in fact, serious discrepancies may occur. Consequently, the power of analyses and precision of estimates can drop below ‘useful’ levels, and, to the extent that missing information is related to outcomes and their determinants of interest, study validity may be compromised. Epidemiological guidelines on how to respect the principle of data completeness must therefore be taken seriously. Given the high importance of this particular topic, issues associated with data completeness will recur in other chapters of the book. A similar line of reasoning can be developed for the other principles listed.

It is important to realize that these principles form a web-like framework in tension with each other; principles, therefore, will at times come into conflict. To extend the example above, in trying to maximize precision of estimates, one might seek to enroll a very large participant pool. However, this might put participants at needless risk and lead to inefficient use of public funds and time and thus come in conflict with both risk minimization and parsimony. It is by the difficult task of weighing and balancing these principles that we arrive at conventions of acceptable levels of risk, cost, and statistical power. This balancing act is especially evident when dealing with “maximizing” and “minimizing” principles.

Hint

When planning a study, a useful exercise is to consider each of the general principles of epidemiology in a step-by-step manner, much like how problems leading to data incompleteness were charted above. This process may be time consuming but will yield high dividends and ultimately save significant amounts of time.

Having armed ourselves with a definition of epidemiology and heightened our senses to its scope and key general principles, let us proceed to have a close look at basic concepts of epidemiology in Chap. 2.

References

Kauchali S et al (2004) Local beliefs about childhood diarrhoea: importance for healthcare and research. J Trop Pediatr 50:82–89

Miettinen OS (2011a) Epidemiological research: terms and concepts. Springer, Dordrecht, pp 1–175. ISBN 9789400711709

Miettinen OS (2011b) Up from clinical epidemiology & EBM. Springer, Dordrecht, pp 1–175. ISBN 9789048195008

Basic Concepts in Epidemiology

2

Lars Thore Fadnes, Victoria Nankabirwa,
Jonathan R. Brestoff, and Jan Van den Broeck

The theory of probabilities is at bottom nothing but common sense reduced to calculus.

Laplace

Abstract

Basic or core concepts are by no means simple or unimportant. In fact, the true hallmark of an expert is a deeper understanding of basic concepts. In this chapter we will introduce basic epidemiological concepts. Epidemiological research addresses the occurrence of health-relevant characteristics or events in a specified type of people. The characteristic or event of interest is often referred to as the ‘outcome’ and the type of persons in which it occurs is often referred to as the ‘target population’. The frequency of the outcome can be of interest itself, or, the interest may be in the link between the outcome’s frequency and one or more determinants, often called ‘exposures’. Analytical studies address causal links, in contrast to purely descriptive studies. Irrespective of whether a study is descriptive or analytical, empirical evidence is obtained by documenting relevant experiences of a study population, a sampled group of individuals who are intended to represent the target population of interest. To describe such empirical

L.T. Fadnes, Ph.D. (✉) • J. Van den Broeck, M.D., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: lars.fadnes@cih.uib.no; Jan.Broeck@cih.uib.no

V. Nankabirwa, Ph.D.
Department of Paediatrics and Child Health, School of Medicine,
College of Health Sciences, Makerere University, Kampala, Uganda

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_2,

evidence, the frequency concepts of risk, rate, and odds are essential. The frequency of the outcome is often compared among different levels of exposure. In analytical studies, this comparison must strive for freedom from the blurring effects of confounding. In this chapter we explain this phenomenon of confounding. We also discuss the exploration of factors that mediate or modify a causal link. The final section of the chapter discusses types of biases in study findings.

2.1 Occurrence Relations

Core concepts in epidemiology are summarized in Panel 2.1. Perhaps the most basic of those concepts is the *occurrence relation*. In epidemiological studies, one investigates the occurrences of outcomes and/or the relationship between outcome occurrences and exposures. The most *basic occurrence relation* (Fig. 2.1) that can be studied is the relationship between a single exposure and an outcome.

Additional elements may need to be added to the occurrence relation when designing a study. When the study is ‘analytical’ (*See:* next section), showing a causal link between an exposure and outcome usually requires taking into account other factors that might confound (blur) the detection of that link (discussed further below in the section on confounding). Thus, in analytical studies these additional factors, called *confounders* need to be included in the occurrence relation. The diagram representing the occurrence relation is then called a *causal diagram*, of which the most basic form is shown in Fig. 2.2.

Panel 2.1 Summary of Basic Concepts in Epidemiology

Analytical studies Studies seeking to demonstrate a causal link

Bias Deviation from the true value

Causal link A statistical association that is free of the distorting influence of confounding factors

Cohort A fixed group of subjects composed on the basis of a once-off selection criterion and followed to study the frequency of occurrence of the outcome

Confounder A third factor that distorts (away from the true independent effect) the observed association between exposure and outcome

Descriptive studies Studies not seeking to demonstrate a causal link

Dynamic population A group of subjects with varying composition over calendar time because membership, based on a chosen criterion, only lasts for as long as the criterion is fulfilled

Effect modifier A factor by whose level the relation between exposure and outcome changes

Exposure Determinant; factor related (causally or acausally) to the outcome

(continued)

Panel 2.1 (continued)

Hypothesis A scientific idea (*Based on Miettinen* 1985)

Information bias Bias in the statistical study result caused by problems with measurement, data processing or analysis

Measurement Investigation of an attribute of a single observation unit; and the recording of a 'representation' characterizing the attribute under the form of a value on a measurement scale

Mediator A factor by which the exposure exerts its effect on the outcome

Observation unit Person or other entity, member of the study base, whose characteristics or experience is to be measured

Occurrence relation The object of study: the proposed relation among outcome, exposures (and sometimes confounders and effect modifiers)

Odds probability of having (or developing) the outcome divided with the probability of *not* having (or developing) the outcome

Outcome The phenomenon whose frequency of occurrence is studied

Population cross-section A 'snapshot' of a cohort at a particular follow-up time or of a dynamic population at a particular calendar time

Rate Frequency of occurrence

Risk Probability of some state or event developing

Selection bias Bias in the statistical study result caused by problems of selection or retention of study participants

Study base The real-life experience of members of a cohort, dynamic population or population cross-section that will be documented to provide empirical evidence about the occurrence relation

Study population The group of people that will provide for the study base

Target population The type of people about which evidence will be created in the research

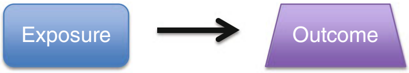


Fig. 2.1 The basic occurrence relation. A single exposure is related to a single outcome

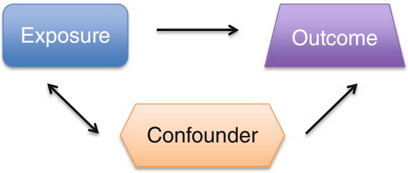


Fig. 2.2 The basic causal diagram. A single exposure is related to a single outcome. A third variable – known as a confounder – is also related to the outcome and is associated with the exposure

2.2 Target Population and Study Population

Occurrence relations are studied for a specified *target population*. As discussed in Chap. 1, the target population is the type of persons the research tries to create evidence about. The target population can be entirely abstract (e.g., adults with a specific illness), or there may be some space or time restrictions (e.g., inhabitants of a specific area). In practice, we study the real-life experiences of a group of persons who represent the target population; this group is called the *study population*. The collective experience of the study population is called the *study base*. Chapter 5 will explain in greater detail the three possible types of study base that can be used: cohorts, dynamic populations, and population cross-sections. In brief, *cohorts* are fixed groups of persons whose exposures and outcomes are documented over a defined period of follow-up time. *Dynamic populations* are non-fixed groups whose attributes of interest are measured in the people fulfilling a set of criteria during a study, with people moving in and out of the study population according to whether they (still) fulfill these criteria. A *population cross-section* is a "snapshot" of a study population at a specific time. In all three cases, attributes and experiences in the study population are recorded either repeatedly or once. Because the study population represents the target population, the empirical evidence and relationships found in it can be used to make *inferences* about the target population.

2.3 Descriptive Versus Analytical Research

All epidemiological studies investigate health phenomena using quantitative methods involving statistical estimation and/or testing. As discussed in Chap. 1, there are two broad types of epidemiological studies: descriptive and analytical studies. But what distinguishes a descriptive study from an analytical one?

The fundamental divide between these two study types is whether or not causality is addressed. In a *descriptive study*, the outcome of interest might be the prevalence of a disease, a correlation, or a shape of a relationship in one or more groups. However, in such studies there is no focus on whether one phenomenon causes or prevents the other. In principle, descriptive research does not address questions regarding causal links between phenomena. The aim is rather to show if the frequency is different between the categories of a determinant, regardless of the reasons for any observed differences.

Analytical studies, on the other hand, are aimed at demonstrating possible causal links among observed health phenomena and are therefore considered to be causally-oriented. The causal links may be associated with an increase or decrease in the frequency of the outcome of interest. Put another way, analytical studies investigate whether determinants (often referred to as *exposures* or presumed *risk factors*) are causally linked with health-relevant outcomes.

To further illustrate descriptive versus analytical studies, consider two different studies, one descriptive and the other analytical, both addressing the relationship between average weekly beer consumption and squamous cell lung carcinoma. In the descriptive study, one might compare beer consumption rates in patients with lung cancer versus the general population, without any attempt to address whether beer consumption is a causal factor for lung cancer. This would yield descriptive information on whether beer consumption is any higher or lower in the patients. In the analytical study, one would attempt to determine whether beer consumption *causes* lung cancer and, if so, to what degree beer consumption increases or decreases the risk of lung cancer. This can only be achieved when it can be convincingly shown that the relationship is free from the effects of *confounding factors*. In other words, it is essential to demonstrate that an observed association is not explained by additional factors (confounders), such as the observation that beer drinkers are more likely to smoke tobacco, a very well-known cause of lung cancer.

2.4 Risks, Odds, and Rates

When describing empirical evidence about occurrences and occurrence relations, the frequency concepts of risk, odds, and rate are essential.

2.4.1 The Distinctions Among Risk, Odds, and Rate

In epidemiology the term '*risk*' is used to denote the probability of some state or event developing (Eq. 2.1) and is expressed as a proportion or percentage. Take, for example, the term 'incidence risk.' An incident case of a disease is a new occurrence in a susceptible individual (e.g., the development of lung cancer in a previously cancer-free individual). 'Incidence risk' is the probability of the outcome (e.g., lung cancer) newly developing over a defined period of time.

$$\textbf{Risk} = \text{probability of a state or event developing} = p \tag{2.1}$$

'*Odds*' is the probability of *having or developing* the outcome divided by the probability of *not* having or developing the outcome (Eq. 2.2). For example, in a cross-sectional study, the odds of cardiac disease is the probability of *having* cardiac disease divided with the probability of *not having* cardiac disease. In a cohort study it would be the probability of *developing* cardiac disease divided by the probability of not developing it.

$$\mathbf{Odds} = \frac{p}{1-p} \tag{2.2}$$

Where:
p = the probability of a state being present or an event occurring

The concept of '*rate*' will be used in this textbook to mean the 'frequency of occurrence' (Miettinen 2011). Rates in this sense can be of a proportion-type or density-type. A *proportion-type rate* is the number of occurrences out of a total number of instances in which the occurrence could have happened. A *density-type rate*, on the other hand, is the number of occurrences out of a total amount of at-risk time (also called 'cumulative person time' or 'population time'). To avoid confusion, one must be aware that many epidemiologists only use 'rate' to denote the latter density-type rates; this restricted use of the term *rate* is still debated (e.g., Miettinen 2011).

2.4.2 Practical Application of Risks and Odds

Risks, odds, and rates are often compared among those who are exposed to a specific factor and those who are not exposed to the same factor. If the outcome is categorical and binary (e.g., healthy or ill, alive or dead, or any characteristic that is present or absent), risk assessment can be made from a two-by-two table (Table 2.1).

To illustrate risk assessment with a two-by-two table, let us consider a theoretical study aimed at assessing whether seat belt use in cars is associated with a decreased risk of death in individuals involved in collisions between two or more cars. The investigators decide to compare the risk of death among those involved in car collisions in areas that have introduced a regulation requiring the use of seat belts versus in similar areas that have not implemented such regulations. The study participants can be categorized according to their exposure (i.e., living in a regulated area or an unregulated area) and outcome status (i.e., death or no death). Table 2.1 is a two-by-two table presenting the study results. This type of table is known among epidemiologists as 'the *basic two-by-two table*'.

Table 2.1 The basic two-by-two table of exposure versus outcome

Exposure level	Outcome: death	Outcome: no death
Exposed (Living in area with regulated seat belt use)	a	b
Unexposed (Living in an area without regulated seat belt use)	c	d

With this information it is possible to calculate the risk and the odds of death based on the exposure status and to compare these values using the relative risk and odds ratio, respectively.

$$\text{Risk among the exposed} = \frac{a}{a+b}$$

$$\text{Risk among the unexposed} = \frac{c}{c+d}$$

The relative risk is the risk among the exposed divided by the risk among the unexposed (Eq. 2.3):

$$\textbf{Relative risk} = \textbf{RR} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \tag{2.3}$$

Similarly, the odds and the odds ratio can be calculated.

$$\text{Odds among the exposed} = \frac{\frac{a}{a+b}}{\frac{b}{a+b}} = \frac{a}{b}$$

$$\text{Odds among the unexposed} = \frac{\frac{c}{c+d}}{\frac{d}{c+d}} = \frac{c}{d}$$

The odds ratio is the odds among the exposed divided by the odds among the unexposed:

$$\textbf{Odds ratio} = \textbf{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} \tag{2.4}$$

A relative risk or odds ratio of 1 suggests equal outcome frequencies in the exposed and unexposed groups. The value 1 is called the *null value,* i.e., the value indicating a *null effect*.

2.5 The Epidemiological Approach to Showing Causal Links

2.5.1 The Basic Temporality Criterion of Causality

For an exposure to be a cause, the exposure must have preceded the outcome, a requirement commonly referred to as the *basic temporality criterion.* The opposite situation is often referred to as *reverse causality,* which is when the outcome has a causal effect on the exposure, e.g., if a disease outcome such as cardiac failure causes an exposure of interest, such as inactivity. This is of particular concern in studies where it is difficult to assess the order of events, such as in many cross-sectional and retrospective studies. In these designs, much of the information regards past events or experiences and is often obtained using patient recall and/or medical records. Take, for example, the known associations between obesity and depression: obesity is associated with increased risk of having a major depressive episode (MDE), and prior history of a MDE increases the risk of developing obesity. Thus, obesity is a cause of MDE, and MDE is a cause of obesity. If attempting to study these two health phenomena, it is therefore necessary to rule out prior exposure to the outcome of interest (either obesity or depression depending on the specific study question) in order to avoid issues of reverse causality.

2.5.2 Types of Causality-Oriented (Analytical) Studies

In epidemiology there are two main types of studies addressing questions of causality: observational etiologic studies and intervention studies. These are also known as observational-etiognostic and intervention-prognostic studies, respectively (Miettinen 2004). They will be discussed amply in Chap. 6 (General Study Designs). Within each of those two broad types of causality-oriented studies, the focus can be on one or more of the following issues:

- Whether a causal link exists
- How strong the causal link is
- Whether other factors can modify the strength of the causal link
- Whether a factor is a mediator in a causal chain

To provide a brief introduction, in observational-etiognostic studies, such as cohort studies and case–control studies, the fundamental question is: to what extent does an exposure cause an outcome? In intervention-prognostic studies, such as randomized controlled trials, the question is rather: to what extent does *imposing an exposure* change the frequency of an outcome.

Let us consider one example from each analytical study type. Both examples will be based on causes of decompression sickness, a serious and potentially-life threatening condition that can affect divers upon ascent. A team of investigators is collaborating with government agencies to develop a deeper understanding of the causal factors contributing to decompression sickness. The researchers hypothesize that the depth of diving and speed of ascent (exposure) are causal factors for the onset of decompression sickness (outcome).

They first address this hypothesis using an observational-etiognostic study in which they monitor 1,000 divers over 10 dives each (10,000 dives total). They use remote electronic devices to observe and record the depth of the dive and many other factors, such as nitrogen pressure in the diver's blood, the rate at which the diver descended, the duration of the dive, and the rate at which the diver ascended to the surface. Each diver phones the research team after their dives to report whether they required clinical assistance for decompression sickness or whether they experienced any hallmark signs or symptoms of decompression sickness. Based on this information, the researchers perform regression analyses to test whether the depth of diving and speed of ascent increase the risk of having experienced decompression sickness or its signs and symptoms, and they adjust for known and potential confounders to be confident that the association will indicate the presence or absence of a true causal link between the depth of diving or speed of ascent and decompression sickness.

Let us presume that the researchers determine that the speed of ascent is a strong causal factor for the onset of decompression sickness. They decide that this association must now be tested using an alternative approach, so they employ an intervention-prognostic study. They enroll 2,000 different divers and randomly assign them to one of two groups: one that will be asked to modify their diving ascent to a slower-than-standard rate and one that will be asked to continue diving as usual (standard ascent rate). They then assess the same parameters as in their observational-etiognostic study over 10 dives per diver. Indeed, they determine that those who were assigned to the slower-than-standard ascent rate experienced a lower risk of decompression sickness than did those who were assigned to the standard diving group. Their study included a rigorous assessment of potential confounders that were accounted for during analysis to be sure that this result was free from the influence of confounders. A deeper discussion of confounding and various examples of confounding will follow later in the chapter.

2.5.3 The Counterfactual Ideal

In the previous example, there is a critical assumption: that the experience of the slower-than-standard ascent rate would have reduced the risk of decompression sickness in the other group had they also slowed their ascents. This assumption refers to what has been called the *counterfactual ideal*. This ideal is a theoretical scenario in which:

- A specific person can be exposed to both levels of exposure at the exact same time (slower ascent and standard ascent) and
- The potential outcome (decompression sickness) can be observed at both levels of exposure in the same person

Essentially, the counterfactual ideal is a theoretical situation in which we suppose that the levels of exposure can be directly compared under exactly identical situations at exactly the same time. In such a scenario, it would be possible to ask what would have happened under a hypothetical change of the exposure level and therefore directly test causality. Unfortunately, this ideal is practically impossible. Instead we attempt to get as close as possible to achieving the counterfactual ideal by making sure that any outcome-determining characteristics and external influences, which can act as confounders, are adjusted for when contrasting the exposure levels.

2.5.4 Cause Versus Causal Mechanism

In analytical studies a statistical association between an exposure and an outcome is potentially a causal link, and the strength of evidence for this causal link is directly related to how well potential or known confounders are taken into consideration or adjusted for. Let us assume that we have shown the existence of a confounding-free association and believe that we have evidence supporting a causal link between an exposure and an outcome. What is the meaning of that association or causal link? This association implies that the exposure direclty or indirectly causes the outcome or, put another way, that the exposure and outcome are in a causal pathway. However, the details of that causal pathway remain unknown. If the causal pathway involves intermediate steps, then those intermediate factors are called *mediators*. For example, imagine that exposure A is causally linked to outcome X, but the causal pathway involves a sequence in which exposure A causes exposure B, and exposure B causes outcome X. In this case, exposure B also has a causal link to exposure X and serves as a mediator of the causal pathway that links exposure A to outcome X. Potential mediators can be measured and their role studied in analytical studies (*See:* Sect. 2.7.1). It is important to realize that an exposure's causative effect could indicate that there is some illness-predisposing mechanism operating in those with the exposure, some illness-protective mechanism in the unexposed, or a mixture of both.

2.5.5 Causal Webs

Traditional epidemiological approaches often involve investigating multiple suspected causes simultaneously in a single etiologic study. The usual analytical approach is to include all of the suspected causal factors as independent variables in multivariate regression analyses. However, more complex networks of causation are increasingly recognized, and more sophisticated causal models are increasingly needed. Pearl (2010) has developed a general theory of structural causal modeling with potential for implementation for the estimation of causal effects, mediation, and effect modification given such complex occurrence relations. Approaches to include hierarchically structured and nested causal factors have also been developed, e.g., multilevel modeling. Discussions of these advanced analytical strategies are outside the scope of this textbook.

2.6 Confounding

Epidemiologists often conduct studies to describe the causal effects of exposures, but in many cases end up with mere associations between exposures and outcomes that are not free from the blurring effects of confounders. Confounding hinders our ability to see the true causal effect of the exposure on the outcome. It can mask associations when they truly exist, or indicate spurious associations when in fact there are no causal relationships.

2.6.1 Confounding: Types and Conditions

Observation of an association between an exposure and an outcome does not necessarily imply causation. In the absence of random error and bias, there are several possible explanations for such associations in nature, including the following:

1. The exposure causes the outcome (Fig. 2.3)
2. The outcome causes the exposure (reverse causation) (Fig. 2.4)
3. The exposure causes the outcome and the outcome causes the exposure (Fig. 2.5)
4. The non-causal exposure and the outcome share a common cause (Fig. 2.6)
5. There is another determinant of the outcome, which is not a cause of the exposure but whose distribution is unequal among exposure levels (Fig. 2.7)
6. The causal exposure and the outcome share a common cause (Fig. 2.8)

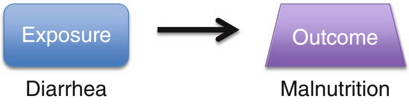


Fig. 2.3 Exposure causes the outcome. For example, diarrhea causes malnutrition

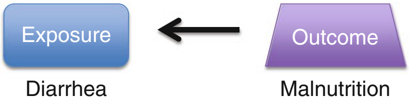


Fig. 2.4 Outcome causes exposure (reverse causation). For example, malnutrition causes diarrhea

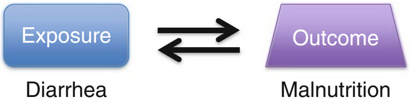


Fig. 2.5 Exposure causes outcome and outcome causes the exposure, creating a 'vicious circle.' For example, diarrhea causes malnutrition, and malnutrition may further worsen diarrhea, and so on

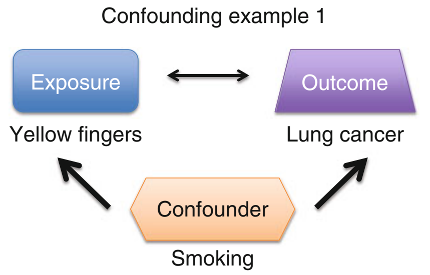


Fig. 2.6 Non-causal exposure and outcome share a common cause. The observed association between exposure and outcome is entirely due to confounding. Causal effects are shown by *thick arrows*, observed non-causal associations with *thin arrows*. For example, smoking causes lung cancer and yellow fingers, which may lead to an apparent causal link between yellow fingers and lung cancer

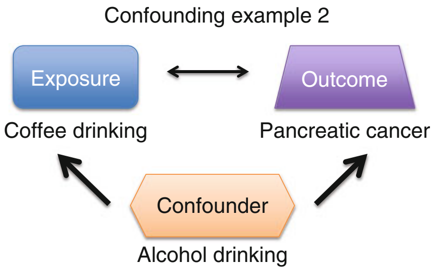


Fig. 2.7 The third factor is a determinant of the outcome and (non-causally) associated with the non-causal exposure. The observed association between exposure and outcome is entirely due to confounding. *Thick arrows* are causal effects; *thin arrows* are observed non-causal associations. For example, alcohol drinking causes pancreatic cancer, but alcohol drinking is also related to coffee drinking. Although it appears that coffee drinking causes pancreatic cancer, that apparent association is due to the confounder only

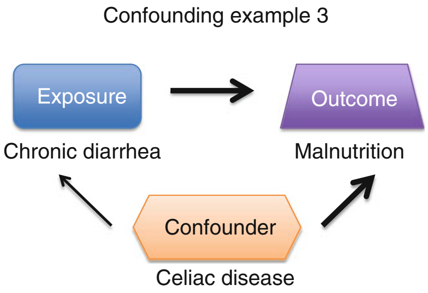


Fig. 2.8 The causal exposure and outcome share a common cause. The observed association between the exposure and outcome is partly causal but overestimated by the confounding influence of the common cause. For example, chronic diarrhea causes malnutrition, but so too does Celiac disease. Some of the association between Celiac disease and malnutrition is due to chronic diarrhea, but there is a diarrhea-independent component to malnutrition in Celiac disease. Thus, if one does not control for Celiac disease when assessing chronic diarrhea as a causal factor in the development of malnutrition, the apparent exposure-outcome relationship will be over-estimated

The first explanation (Fig. 2.3) is what epidemiologists are often searching for and has been discussed at length earlier in this chapter. The second explanation (reverse causation, Fig. 2.4) is raised when it is unclear whether the exposure comes before the outcome. If the exposure always comes before the outcome – such as some genetic exposures and their associated diseases, or such as prospective studies in which the exposure is assessed before the outcome occurs – reverse causality is a non-issue. Figure 2.5 shows a scenario in which the exposure and outcome cause each other in a vicious circle, as is known to be the case with infection causing malnutrition and also malnutrition causing infection.

The explanations presented in Figs. 2.6, 2.7 and 2.8 are cases of what is referred to as confounding. One of the features common to the scenarios in Figs. 2.6, 2.7 and 2.8 is that there is an imbalanced distribution – between the exposed and unexposed groups – of determinants of the outcome other than the exposure of interest (i.e., non-comparability between the exposed and unexposed groups with respect to other determinants of the outcome). Thus, the observed risk/rate in the unexposed does not equal the counterfactual risk of the exposed (i.e., the risk/rate of the exposed had they not been exposed). Common to all confounding are the 'criteria' listed in Panel 2.2.

Uncontrolled confounding can cause an effect estimate to be either more positive or more negative than the true effect. Confounding variables that are positively associated with both the exposure and outcome or negatively associated with both the exposure and outcome make the observed association more positive than the truth (Fig. 2.9). On the other hand, variables which are negatively associated with the

Panel 2.2 The Classical Confounding Criteria

To cause confounding, a variable should:

- Be unequally distributed among exposure levels (because of a causal *or* non-causal association between the confounder and exposure)
- Be a cause of the outcome or be strongly associated with a cause of the outcome
- Be outside the causal pathway between the exposure and outcome, i.e., it should not be a mediator

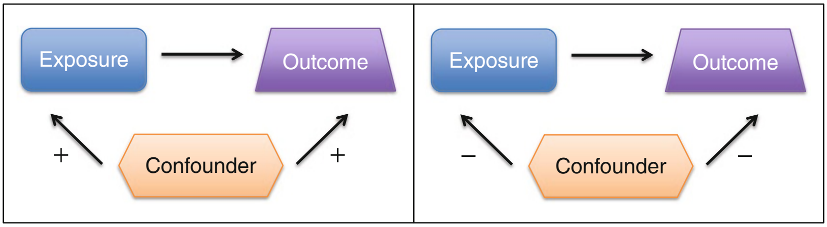


Fig. 2.9 Confounding in a positive direction. In both cases, the confounder is related to the exposure and the outcome in the same directions. The confounder will increase the apparent relationship between the exposure and outcome

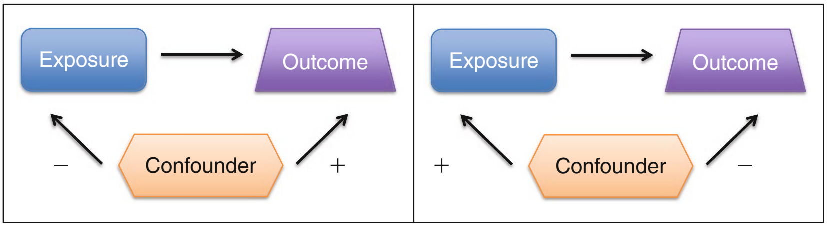


Fig. 2.10 Confounding in a negative direction. In both cases, the confounder is related to the exposure and outcome in different directions. The confounder will decrease the apparent relationship between the exposure and outcome

exposure and positively associated with the outcome, or vice versa, make the observed association more negative than the true association (Fig. 2.10). This direction of confounding will be true regardless of whether the main effect is protective or harmful.

2.6.2 Management of Confounding

Confounding may be prevented in the design of the study or adjusted for in the analysis. Methods used in the design stage include randomization, matching and restriction (e.g., by use of exclusion criteria making the groups more homogenous). Commonly used methods in the analysis stage include stratification, standardization, and multivariable analysis. Each of these methods is briefly introduced below. More information is found in Chaps. 6, 22 and 24.

Randomization is used in experimental studies and consists of randomly allocating participants to intervention arms. When successful, randomization will result in groups with equal distributions of the other factors associated with the outcome other than the intervention, and thus it breaks the links between the common causes of the exposure and outcome. When a study sample is sufficiently large, on average, randomization will result in equal distributions of common causes of both the exposures and outcome. However, randomization is unfeasible or unethical in many instances, for example when an exposure is clearly harmful or beneficial.

Matching is sometimes used in observational studies. Subjects are deliberately selected such that (potential) confounders are distributed in an equal manner between the exposed and unexposed groups. Matching does not come without limitations, though. Perhaps most notably, matching can be expensive as it makes it more difficult to recruit participants and achieve the required sample size. In addition, the effects of matched variables cannot be studied.

In restriction, the study is limited to respondents with the same value of the confounding variable of interest. Thus, the study population is more homogenous than it would be without restriction. For example, if biological sex is a known potential confounder, the study can be restricted to only studying either males or females (although this would raise ethical concerns). Restriction is often simple, convenient, and effective. And it is particularly useful when confounding from a single variable is known to be strong. However, restriction may make it difficult to find enough study subjects, and it can limit generalizability of the findings (a problem of limited external validity).

Methods of managing confounding during data analysis are discussed in Chaps. 22 and 24. In brief, stratification is a commonly used method to control for confounding in which data analysis is stratified on the potential confounding variable. Separate analyses are conducted for those with and those without the confounding characteristic. Stratification is cumbersome when there are multiple potential confounders, as the data would have to be split into several strata. This is problematic as it may result in severe losses in statistical power and reduce the likelihood that a conclusion can be made. Another approach to managing confounding is to employ multivariable analyses using regression methods to control for multiple confounders at the same time. Such analyses can also be used to control for continuous variables without categorizing them, unlike stratification. Irrespective of which approach is chosen, ultimately theory should always guide the selection of variables considered as confounders, and careful reasoning is necessary because confounding is context-dependent: a variable may be a confounder in one context but not in another.

When assessing confounding in an observational design, it is essential to measure factors that could be causally related to the outcome. Poorly accounting for known, potential, or plausible confounders that are not measurable or poorly measurable can obscure true causal links or indicate false links. Any previously unsuspected or unknown confounder, newly shown to be important would constitute a potential paradigmatic shift in the causal thinking about a disease or other health outcome. If a new risk factor is identified, then previous causes (including previous confounders) thought to be genuine before may become ‘weaker’ or even disappear. Consequently, as small paradigmatic shifts succeed each other, the causal webs tend to re-shape, and the strength of the links tends to change.

2.7 Mediation and Effect Modification

2.7.1 Mediation

Mediators or intermediate factors are those factors that are in the direct causal chain between the investigated exposure and the outcome (*See:* Fig. 2.11). When investigating causal links, adjusting for these factors might remove true associations or reduce their magnitude. For example, in a study assessing the association between cardiac disease (outcome) and nutrition (exposure), adjustment for nutritional variables such as plasma lipids and cholesterols is likely to reduce the measured effect size. This is because changes in the lipids and cholesterol might be triggered by the nutritional exposure. That is, changes in lipids and cholesterol are part of the mechanism through which the nutritional exposure causes cardiac disease (Fig. 2.11). When selecting confounders for adjustment, it is important to make sure that the selected confounders are not in fact partly or entirely mediators. To the extent that

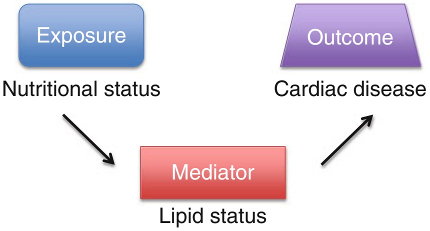


Fig. 2.11 A mediator defined as a variable in the casual pathway between the exposure and outcome. For example, nutritional status causes cardiac disease by affecting lipid status

Table 2.2 Risk traffic deaths as outcome from traffic accident among persons not having used helmet and having used helmets

Exposure level	Died during the accident	Survived the accident	Case fatality rate (%)
Helmet used	200	800	20
No helmet used	200	1800	10

Table 2.3 Risk of death from traffic accidents with and without the use of a helmet, stratified into those driving motorcycles and those driving vehicles. Only crude point estimates presented

Exposure level	Died during the accident	Survived the accident	Case fatality rate (%)
Stratum-1: Motorcyclists			
Helmet used	199	791	20
No helmet used	100	100	50
Stratum-2: Vehicle drivers			
Helmet used	1	9	10
No helmet used	100	1700	6

they are, the observed effect will tend to be diluted. Statistical methods of mediation analysis exist to assess the mediating role of variables. These methods are beyond the scope of this book.

2.7.2 Effect Modification

In some cases, the initial conclusions after first analysis are incorrect. An example could be an investigation of traffic casualties among people using helmets and those not using helmets. One might initially find that traffic casualties are more common among those using helmets (Table 2.2).

Does this mean that helmet use is a risk factor? Not necessarily. What if, for example, helmets were used nearly exclusively by motorcyclists and rarely by those driving cars? Would it still be reasonable to compare the risks without taking this difference between the groups into account? Table 2.3 explores this question by presenting results of a stratified analysis among motorcyclists and people driving vehicles:

This example shows that using a helmet is a preventive factor rather than a risk factor among the motorcyclists. This is an example of effect modification (also called 'interaction'), which exists when the effect of the exposure on an outcome differs by levels of a third variable. In the helmet example, the effect of wearing a helmet in a traffic accident depends on whether one was riding a motorcycle or driving a car.

2.8 Bias in Epidemiological Research

Bias refers to systematic deviation of results or inferences from truth (Porta et al. 2008). It results from erroneous trends in the collection, analysis, interpretation, publication, or review of data (Last 2001). Bias may result in the overestimation or underestimation of measures of frequency or effect. The cause of a biased statistical result may be in the selection of information sources, in the gathering of information (measurement and data management) or in the analysis of gathered information. The role of measurement error is often crucial. Both random and systematic measurement error can lead to biased estimates of effect (*See:* Chaps. 11 and 27). It is not feasible to completely eliminate measurement errors, but minimizing them and estimating their influence is a priority in epidemiological research. Bias is often categorized, according to the source of the problem, into selection bias and information bias. A special type, publication bias, will be discussed in Chap. 31.

2.8.1 Selection Bias

Selection bias is a form of bias resulting from (i) procedures used to select subjects or (ii) factors that influence loss to follow-up. At the core of the various selection biases is the fact that the relationship between the exposure and the outcome for those participating in the study is different than for those who theoretically should have been included in the study. Selection bias due to sampling and enrollment procedures will be discussed further in Chap. 9.

2.8.2 Information Bias (Measurement or Analysis Bias)

Information bias is a form of bias resulting from problems with the measurement of study variables or the processing of data. This can have various reasons including challenges with recall of information, social desirability, use of sub-optimal measurement tools, and unfortunate phrasing of questions and answer alternatives. Chapter 27 gives multiple examples of information bias resulting from measurement error. In Chap. 18 we will further discuss recall bias, social desirability bias (Zerbe 1987) and bias resulting from poor formulation of questions (Schwarz 1999).

In this chapter we discussed some core concepts and terms in epidemiology. These ideas are the result of a constant evolution in the theoretical framework of epidemiology, with progressive conceptual developments and sometimes conflicting uses of terms. The emergence, refinements, and re-definitions of concepts in quantitative health research can be traced back to long before epidemiology became a discipline, even before formal quantitative statistics-based comparisons became used. Thus, in the next chapter we discuss historical roots of epidemiology and then contemplate some of the emerging issues in the field that will very likely change the future of our discipline.

References

Last JM (2001) A dictionary of epidemiology, 4th edn. Oxford University Press, Oxford, pp 1–196. ISBN 0195141687

Miettinen OS (1985) Theoretical epidemiology. Delmar, New York, pp 1–359. ISBN 0827343132

Miettinen OS (2004) Knowledge base of scientific gnosis: II. Gnostic occurrence relations: elements and temporal structures. J Eval Clin Pract 10:357–359

Miettinen OS (2011) Epidemiological research: terms and concepts. Springer, Dordrecht, pp 1–175. ISBN 9789400711709

Pearl J (2010) An introduction to causal inference. Int J Biostat 6(7):1–59

Porta M, Greenland S, Last JM (2008) A dictionary of epidemiology. A handbook sponsored by the I.E.A, 5th edn. Oxford University Press, New York, pp 1–289. ISBN 9780195314496

Schwarz N (1999) How the questions shape the answers. Am Psychol 54:93

Zerbe WJ, Paulhus DL (1987) Socially desirable responding in organizational behavior – a reconception. Acad Manag Rev 12:250–264

Roots and Future of Epidemiology 3

Jan Van den Broeck and Jonathan R. Brestoff

Study the past if you want to define the future.

Confucius

Abstract

The first purpose of this chapter is to outline the roots of epidemiology as a methodological discipline, using a multiple-threads historical approach. We unravel what we see as the main historical threads relevant to the development of current health research methods involving human subjects, giving attention to the ethical, scientific-theoretical, and practical aspects. Roots of epidemiological concepts and methods go back a long time, to before epidemiology became a named discipline and before formal statistical comparisons of occurrence frequencies started being made. We take the stance that ancient thinkers, dating back at least as far back as Aristotle, formed early concepts that have been essential to the development of modern epidemiology as we know it. We therefore treat such critical developments as directly relevant to the history of epidemiology. As an introduction, we begin with a discussion of belief systems. We then discuss a series of historical threads, starting from health research topics, over ways of causal thinking about health, to the design of empirical information, research ethics and stakeholder participation. Other threads relevant to epidemiology such as history of data management, analysis, and study reporting, are not covered. Finally, we explore some possible and desirable future developments in epidemiological research.

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_3,

3.1 Belief Systems

Points of view established through belief systems can lead to the relief of human suffering caused by illness and ignorance. Belief systems represent a continuum, on one end of which is faith and the other science. Perhaps driven by the raging debates on religion versus science, there is a common misconception that science is independent of belief. Although science does, by definition, rely on empirical evidence to support the existence of an occurrence or entity, scientists must always decide to what degree they believe in the evidence provided and in the theoretical ideas used to contextualize that evidence. The primary distinction between faith- and science-based belief systems is in their requirement for supporting empirical evidence. In faith-based belief systems, believing in the existence of an occurrence or entity requires no evidence and relies mainly on revelation, authority, anecdotal accounts, and tradition. In science-based belief systems, belief in a phenomenon only occurs when sufficient empirical evidence is available to support its existence (Morabia 2011).

The relative importance of faith-based and science-based belief systems as alternative and (sometimes) competing means of achieving knowledge has changed throughout history (Seife 2004). As the self-consciousness and confidence of mankind increased, so too did trust in research (which involves empirical data collection) as a means for achieving valid knowledge. Obtaining evidence through research requires many skills, including theory-based reasoning and hypothesis generation; thus, a discussion of the roots of epidemiology appropriately starts with acknowledging the ancient Oriental and Greek philosophers for their contributions to the awakening of human reason and later philosophers, such as Kant, for exploring and describing human reason's boundaries.

The *scientific method* is the systematic method of empirical investigation believed by most scientists to yield the most valid, traceable, and reproducible evidence about a theoretical research question. The scientific method – defined as such – has evolved considerably over calendar-time. Historians are divided about whether it is justifiable to trace the history of a discipline back to periods before it become known under its current name. Some epidemiologists, like Morabia (2004), take the view that the defining period for epidemiology is the seventeenth century, when formal comparisons of occurrence frequencies started being used. For others, the history of epidemiology starts in the nineteenth century when epidemiology became a recognized discipline in Britain. We take the view that there were researchers and scientific methods (plural) long before the words "researcher" and "scientific method" were used. Similarly, we believe that epidemiologists and epidemiology existed well before the terms came into use. After concepts, theories, empirical methods, and statistical approaches are introduced, they are refined and formalized both *before* and after a new discipline acquires a name. Theory-based learning from empirical observation has existed from ancient times and so has an interest in learning about health-related phenomena. Thus, although many will argue that Hippocrates and Fracastoro, for example, are not real epidemiologists, one can see that these figures have been crucial to the development towards current epidemiological thinking.

Given the definition of epidemiology proposed in Chap. 1, we suggest that the history of epidemiology should not be confused with the history of medicine or public health in Western civilization. The history of epidemiology deals specifically with *the roots of* principles and methods used in comparative population based health-related *research* throughout history. Several historic threads can be followed looking at how various steps of the research process have been carried out over the centuries in several civilizations. When tracing these threads, the historian describes milestone events and new ideas, explains them by putting them in context, and indicates how the events and ideas have influenced the subsequent practice and conceptualization of health research. Unfortunately, only fragments of selected historic threads, mainly relating to Western civilization, can briefly be touched upon in the next sections. These threads are listed in Panel 3.1. In our discussions of threads below, we draw mainly on publications by epidemiologists with an interest in the history of our discipline.

Historic developments in each of the aspects of epidemiology have not always run in parallel. Thus, each thread is discussed separately with some cross-referencing where relevant. Due to space restrictions we will not cover the important threads of data management, statistical analysis and study reporting.

Before we uncover some roots of epidemiology, we must introduce some key concepts and terms. Panel 3.2 highlights a selection of concepts and terms and explains their meanings as used in the sections below.

Panel 3.1 Historic Threads Pertinent to Epidemiology as a Methodological Discipline

- History of research topics
- History of causal thinking about health
- History of epidemiological study designs
- History of research ethics and stakeholder participation
- History of research data management and data analysis
  History of study reporting

Panel 3.2 Selection of Key Concepts and Terms Relevant to History of Epidemiology

Empirical Based on measurement

Health research Systematic activity aimed at achieving knowledge about health – related states and events

History of epidemiology The study of calendar time – dependent changes in how medicine has used research with human subjects to increase its knowledge base

(continued)

Panel 3.2 (continued)

Phenomenon A state or event (*Miettinen* 1985)

Research Systematic activity aimed at achieving knowledge

Research ethics Discipline providing ethical principles and guidelines for the design, conduct, analysis, and dissemination of research involving human subjects

Research methods Methods of designing, conducting, analyzing, and reporting a research study

Scientific method Method contemporarily believed by most scientists to yield the most valid evidence about a research question

Scientific research Research using the scientific method

3.2 History of Health Research Topics

This historical thread concerns when and why some research questions about health are asked and others seem to be ignored (Susser and Stein 2009; Krieger 2011), and is intimately linked to the next thread about causal reasoning. Hypotheses and presumed causal factors are never independent of the conceptual frameworks of the time. Naturally, humans possess a fundamental curiosity about and hunger for knowledge about circumstances or behaviors that lead to or protect against illness. Such knowledge first came about by intuition and by experiential learning from trial and error. But at what point did humans start to use *empirical research* (in the sense of systematic evidence collection to answer some theory-derived question) to achieve desired health-related knowledge?

There is no clear-cut answer to this question (in part because the answer depends on one's opinion of what constitutes research; *See:* Textbox 3.1), but McMahon et al. (1960) have drawn the attention of epidemiologists to *Air, Water and Places,* a Hippocratic text (ca 400 BC) in which are found several ideas still relevant to public health research. This text points out what are now known as "risk factors" at different nested levels of observation, like country, area, city, and individual behavior. It also emphasizes the need to study food hygiene, diet, clean water, and exercise for health, giving multiple examples. Although this text may not be considered to be epidemiology by all, one may perceive some roots of epidemiological thinking. Indeed, empirical information from observing nature and patients had an important role in Hippocrates' thinking, and he used this information to make generalizations into the abstract about nature. But Hippocrates was not the only ancient thinker to ask health-related questions. For example, attempts of diagnostic and etiognostic classifications were important concerns of, for example, the Greek philosopher Aristotle. This ancient call for a classification of disease has been echoed over the millennia, and one such call by Thomas Sydenham (1624–1689) was particularly poignant: "All diseases should be reduced to definite and certain species with the same care which we see exhibited by botanic writers in their phytologies."

Textbox 3.1 When Did Population-Based Health Research and Experimentation Start?

When population-based health research started is unclear. According to Hetzel (1989), “One of the oldest references to goiter is attributed to the legendary Chinese Shen-Nung Emperor (2838–2698 BC) who, in his book **‘Pen-Ts’ao Tsing’** (‘A treatise on herbs and roots’) is said to have mentioned the seaweed Sargasso as an effective remedy for goiter.” Hetzel also states that “…the Wei dynasty (AD 200–264) attribute deep emotions and ‘certain conditions of life in the mountain regions’ as causes of goiter.” Clearly correct insights had been gained regarding the therapeutic effect of (iodine-rich) seaweed for goiter and the higher prevalence of endemic goiter in mountainous areas. It is unclear how exactly these precise insights were gained but it seems possible that ancient healers – proto-scientists endowed with particularly passionate interests in health issues – may have used systematically repeated trial and error runs or systematic series of observations to arrive at recommendations and conclusions similar to the above. Whether such investigations were attempts to answer theory-based questions is also plausible, no matter how primitive or ‘wrong’ the theory might appear today.

Meinert (1986) cites an example of a planned research experiment that can be found in the **Book of Daniel**. It consisted of a comparison of persons put on a 10-day diet of pulses (a legume) with persons eating another diet and found that those eating the former came to appear ‘fairer and fatter in flesh’ in comparison with the latter. However, controversy exists and doubt remains about whether this can be considered an example of research (e.g., Morabia 2004).

At the end of the eighteenth and beginning of the nineteenth century, rapid developments in science and philosophy (in what is known as the ‘Age of Reason’) were accompanied by the industrial revolution. Health inequalities were great and epidemics frequent. Hence there was heightened interest in public health and preventive medicine in that period and an increasing recognition of environmental and social causes of disease. There was also a strong impetus to conduct analytical research on epidemic diseases.

In the twentieth century, along with increasing success in combating infectious disease, issues of non-communicable disease became prominent as research questions, especially those regarding cardiovascular diseases and cancers. The interests in environmental, social, and heritable determinants of ill health were developed throughout the twentieth century, leading ultimately to a modern understanding of illness, in which consideration is given to:

- Multiple interacting risk factors rather than single factors as the causes of disease
- Lifestyle factors that might cause or prevent disease
- New modern paradigms, including the ‘Barker hypothesis’, about the early-life origins of adult disease
- Other complex health phenomena

The twentieth century also witnessed a worldwide explosion of research into the effects of pharmacological preparations, surgical interventions, behavioral therapies, and various types of community level interventions, each in terms of effectiveness, safety, cost, and acceptability.

3.3 History of Causal Thinking About Health

Along with the investigational interests discussed above, there have been notable shifts in causal theories about health.

3.3.1 Early Paradigms About Causes of Disease

Since pre-history a prevailing paradigm has been that divine anger causes illness and divine grace cures. For example, in the Iliad (Homer, ca. 800 BC) an epidemic of plague is sent by Apollo, the god of healing. Disease was thought to have supernatural origins, an idea that has never fully disappeared (Irgens 2010). In apparent contrast to supernatural causes of disease, the Hippocratic texts provided a conceptual framework in which disease was caused by environmental and behavioral factors that led to imbalances among four body fluids: blood, phlegm, black bile, and yellow bile. Fever, for example, was thought to be caused by excess bile in the body (Krieger 2011). In ancient China, illness was considered to be the outward manifestation of internal disturbances in complex natural systems that were subject to environmental and social-behavioral factors (Krieger 2011).

By the Middle Ages in Europe, the ancient works of the Hippocratic authors, Galen, and others had been forgotten, and disease was again mostly considered to have supernatural causes. The works of these ancient writers, however, had been preserved via the Islamic tradition and were gradually reintroduced to Europe as the Renaissance period began to unfold. Physicians versed in these texts took important roles in the medical schools emerging in European Mediterranean countries during the thirteenth century, thereby helping to infuse ancient ideas of disease causality across Europe and, eventually, much of the world. In other words, with the Renaissance came renewed study of ancient medical texts, and the long-forgotten theories on natural causes of disease re-emerged.

3.3.2 Religion Versus the Scientific Method

Throughout the Renaissance, faith- and science-based belief systems co-existed mostly without conflict (Seife 2004). During the era of Galileo Galilei (1564–1642) a few individuals and organizations, fearful of the potential of science (particularly cosmology) to disprove the existence of God, deployed propaganda campaigns that effectively created conflict between religion and science as approaches to achieving valid knowledge. Science was portrayed as heretical (Seife 2004). Simultaneously, Francis Bacon (1561–1621) proposed the inductive or ‘scientific method,’ and the scientific community had increasingly come to accept this approach as a valid way of achieving knowledge. Bacon stated that scientific reasoning depends on making generalizations (inductions) from empirical observations to develop general laws of nature. He suggested that scientists carefully observe persons and nature rather than only resort to explanations based on faith, classic texts, or authority. Bacon’s description of the scientific method brought a modern conceptual framework to Hippocratic texts that proposed observing environmental and behavioral factors to explain illness.

> **Discussion Point** Belief in a supernatural cause that occurs prior to any form of natural cause is perfectly compatible with the theories and practice of modern science

3.3.3 Contagion Versus Miasma as Causal Paradigms

During the Renaissance, a controversy arose between proponents of the theory of contagion and those of the theory of miasma or ‘bad air’ as main causes of disease. Saracci (2010) has drawn the attention of epidemiologists to the fascinating scientist Gerolamo Fracastoro (1478–1553) from Padua, Italy, who claimed in ‘*De Contagione et Contagionis Morbis et Eorum Curatione’* (1546) that diseases are caused by transmissible, self-propagating material entities. Initially, there was no idea that these entities could be living; the contagions were thought of more as substances than as germs. Fracastoro claimed that the contagions can be transmitted directly from person to person, or, indirectly from a distance. He also theorized about strategies to combat contagions that are still relevant today:

- Destruction by cold or heat
- Evacuation from the body
- Putrefaction
- Neutralization by antagonistic substances

Fracastoro also suggested that syphilis was spread through sexual intercourse, based on observations that the spread of the disease followed the movement of army regiments (Irgens 2010). During the nineteenth century the miasma-contagion debate would reach a high and the contagion theory (also known as the germ theory) eventually prevailed in no small part due to the strong experimental work of Louis Pasteur (1822–1896).

3.3.3.1 Recognition of Specific Non-infectious Causes of Disease

As important as the contagion-miasma controversy and the concluding contributions of Louis Pasteur have been, this debate concerned only vague influences of the environment on health. While the debate raged, the causal role of several more specific non-infectious environmental hazards had become recognized. For example, in 1700, Bernardo Ramazzini, called ‘the father of occupational medicine,’ produced an influential work ‘*De Morbis Artificum Diatriba*’ dealing with a wide range of occupational hazards (Franco and Franco 2001). And in 1775, Percivall Pott recognized that chimneysweepers' exposure to soot was carcinogenic (Susser and Stein 2009).

3.3.4 Philosophical Contributions to Causal Reasoning

Several philosophers, such as Immanuel Kant and John Stuart Mill, have influenced the way scientists thought about causality. In Mill's '*Canons*' (1862), he describes some analytical approaches – general strategies to prove causality – that are still used today:

- 'Method of difference.' This method recognizes that if the frequency of disease is markedly different in two sets of circumstances, then the disease may be caused by some particular circumstantial factor differing between them. This method is akin to the basic analytical approach now taken in trials and cohort studies, i.e., showing that disease outcome is more or less frequent in the presence of a particular exposure
- 'Method of agreement.' This method refers to situations where a single factor is common to several circumstances in which a disease occurs with high frequency. This method is akin to the approach taken in traditional case–control studies, i.e., showing that an exposure is more common in cases
- 'Method of concomitant variation.' This method refers to situations where the frequency of a factor varies in proportion to the frequency of disease. This kind of reasoning is often used in ecological studies, i.e., showing that exposure and disease outcome vary together

3.3.5 Causal Interpretation Criteria

Koch (1843–1910) and Henle described a sequence of studies and results needed for proving that a single infectious agent causes a disease (these are known as the Henle-Koch postulates). These causal criteria have been very helpful in identifying the infectious causes of a number of diseases. Evans (1976) proposed a revision of the Henle-Koch postulates describing the sequence of studies and results needed for proving the causal role of an exposure in general. The mainstream modern approach to showing causality actually involves two steps. Step one is showing an association between the determinant and the outcome phenomenon free of bias, confounding, or reverse causality. Step two is further evaluation of credibility, perhaps also using some of the Evans criteria or Hill criteria (Hill 1965), which will be discussed in Chap. 27. The modern Bayesian approach rests upon modification of prior beliefs about causal links by evidence in research data.

3.4 History of Epidemiological Study Designs

Study design has two main aspects: general design (*See:* Chap. 6) and planning of measurements. These two aspects will be discussed separately below.

3.4.1 Roots of Approaches to General Study Design

Learning from trial and error may be seen as the first quasi-experimental approach. The Hippocratic approach to gathering empirical evidence could be considered 'qualitative' as there was no quantitative formal hypothesis testing or effect estimation, nor were there formal comparisons of quantified outcomes among determinant levels. Comparisons between determinant levels were made but only informally by using the 'computing power of the mind.' The first clear types of more formal designs are the early case series studies. An often cited example was published by Lancisi (1654–1720) in *'De Subitaneis Mortibus'* (1707), where he described a detailed pathological investigation of a case series of sudden deaths in Rome, probably due to myocardial infarctions (Saracci 2010). This was an early case series study of a non-communicable disease. Case series studies are the prototype of observational study designs.

3.4.1.1 Experimentation

As to experimental study designs, one of the earliest known clinical trials – on scurvy, a major problem for British sailors – was performed by James Lind (1716–1791) (*See:* Textbox 3.2).

The precursor of randomization in clinical trials was presented at the congress of scientists in Pisa, 1838. At that meeting, the idea was brought forward of alternating allocation to treatment alternatives as a means to better show superiority of new treatments. But the first modern randomized controlled trial would not occur until 1946 with the MRC trial of Streptomycin on tuberculosis (MRC 1948).

3.4.1.2 The Idea of Formally Contrasting Determinant Levels

Early clinical trials contrasted outcomes frequencies among treatment levels. Such quantitative comparative approaches had been taken earlier for observational studies, most notably by demographer John Graunt (1620–1674) who performed formal subgroup comparisons with observational data. However, the most famous example of the importance of contrasting determinant levels comes from the work of John Snow (1813–1858), in which he performed an outbreak investigation that ultimately led to the elimination of an exposure to a pathogenic source. During the cholera epidemics in London in 1849 and 1854, Snow postulated a water-borne cause of cholera. He noted that the disease was more frequent in those areas of the city that

Textbox 3.2 The Early Trial of James Lind

Twelve sailors with scurvy took part in a trial aboard the ship HMS Salisbury (20 May – 16 June, 1747). James Lind assigned six treatments, presumably in a random way, to two men each: cider; vitriol; vinegar; seawater; oranges plus lemons; and a concoction of garlic, mustard, radish, Peru balsam, and myrrh. Within 6 days those receiving citrus fruits were fit for duty. The others remained sick.

received their water supply from a particular water company that used water from a 'dirty' part of the river. He then went on to close the water pumps of that company to show that the disease rate dropped dramatically after the closure.

3.4.1.3 Sample Size Considerations

In the abovementioned trial by James Lind, there were only two subjects in each treatment arm. There must have been some expectation on the part of Lind that two per arm would be more reliable than one per arm. However, a deep appreciation of the importance of sample size was not achieved until the contributions of William Farr (1807–1883), who is known to have made several contributions to study design. He pointed out the need for sample size considerations and formally introduced the concept of retrospective and prospective studies.

3.4.2 Modern Epidemiological Study Designs

3.4.2.1 Ecological, Cohort, and Case–Control Studies

Formal ecological studies have been very popular for exploring possible causal links since the nineteenth century. They are still used today as evidence of an association between an exposure and outcome, but this study design comes with serious limitations that are often difficult or impossible to address (*See:* Chaps. 5, 6 and 27 for more details on ecologic studies), so they are not considered to be a popular approach.

Today, more popular than ecological studies are cohort studies. The Framingham Heart Study, which was started in 1948, is often considered a landmark cohort study (Dawber et al. 1957). Approximately 5,200 men and women aged 30–62 years in Framingham, Massachusetts, were followed long-term. This research program identified major risk factors for heart disease, described the natural history of cardiovascular disease, and set the standard for modern cohort studies, which have long been the paradigm for observational etiognostic research. Only relatively recently has the at-least-equivalent usefulness of the case–control approach become fully clear. Examples of case–control studies are available from the first half of the twentieth century. Doll and Hill (1950) are often credited with popularizing the case–control design with a landmark study showing an association between smoking and lung cancer. Even after the Doll and Hill paper, however, case–control and case-base approaches have long been considered inferior to the cohort approach and became only very progressively recognized as alternatives. A *Lancet* editorial in 1990 discussed rankings of methodological strength (as found in contemporary methodological books) and stated "The case–control study…falls behind the randomized controlled trial and the prospective and retrospective follow-up study and barely overtakes the humble anecdote." This point of view is now considered antiquated, as case–control studies with density sampling are quite robust. Olli Miettinen (1976, 1985, 1999) has been perhaps most influential in promoting the proper use of secondary study bases in study design, a process that is still ongoing.

3.4.2.2 Modern Developments in Study Design

In the second half of the twentieth century, many important epidemiological concepts around *object design* became firmly established (*See:* Chaps. 2 and 5). Olli Miettinen has been a main driving force in developing object design concepts (*See also:* Morabia 2004). In modern study design, the case–control approach and the primary cohort-based approaches are generally seen as equivalent for observational etiognostic studies. Miettinen (2010) has proposed a new approach, called 'the single etiologic study' that is an improvement of the traditional designs, but it has not yet trickled down into common epidemiological thinking and practice. Recently, Mendelian randomization and other designs using instrumental variables have come to be added to the armamentarium of the observational epidemiologist. Clinical trial design has evolved into various types, including stepped and cross-over designs, and improved randomization and minimization methods have been gradually developed. The serious limitations of classical diagnostic performance studies are also becoming clearer and constitute an important challenge for traditional clinical epidemiology and evidence-based medicine (Miettinen 2011).

3.4.3 History of Measurements of Health-Related States and Events

Developments in measurement methods are driven by and run in parallel to the changing interests in particular research questions and, consequently, with changing conceptual paradigms of objects under study. For example, the development of microscopy can hardly be imagined without a theoretical interest in objects (e.g., microbes) that cannot be visualized with the naked eye. Anthropometry is one of the oldest of types of measurements (so too are autopsy and the counting of deaths and survivors). In an old Hindu textbook on surgery, the '*Sushruta Samhita*' (c. 600 BC), it is stated, "Adult stature is 120 times a man's finger width." In Hellenistic times it was known that total height is 7.5 times the height of the head. Hippocratic texts recognized that climate influences body size and shape, and it was recognized by Galen (130–200 AD) that body proportions are linked to health.

Patient observation, interview, and physical exam have been and will likely always remain important for assessments in clinical care and research. Various forms of highly technical measurement instruments and questionnaire-based scales for latent attributes now often aid physical examination and interview-based measurements. In the past decades, these methods of assessment have rapidly been supplemented with more sophisticated measurements and more advanced methods of data extraction from administrative or health records. Moreover, routine objects of measurement now include molecular analyses of biological samples and complex physiologic measurements as well as physical and biochemical assessments of the environment.

3.5 History of Research Ethics and Stakeholder Participation

3.5.1 History of Research Ethics

Subject protection and Good Clinical Practice guidelines are relatively recent phenomena in research history and were developed mainly after World War II (WWII). Before the war and until some time thereafter, it was usually assumed that the high ethical standards of patient care, as advocated by Hippocrates and Sydenham, would guarantee subject protection in research. History has proven that assumption very wrong. For example, highly unethical research has been conducted in the United States before, during, and after WWII (Beecher 1966; White 2000; Kahn and Semba 2005; Horner and Minifie 2010). The same has happened in several other countries but most notably in Nazi-Germany and Japan (Tsuchiya 2008). It is the particular atrocity and scale of the Nazi medical experiments that eventually awoke spirits and led to important post-war milestone events, starting with the Nuremberg Doctors Trial in 1946 (McGuire-Dunn and Chadwick 1999). The judgment pronounced in this trial of Nazi doctors included a set of ethical guidelines known as the Nuremberg Code. This document started the modern era of human subject protection in research. As pointed out by McGuire-Dunn and Chadwick (1999), the Nuremberg Code stated, among other important points, that:

- There should be no expectation of death or disabling injury from the experiment
- Informed consent must be obtained
- Only qualified scientists should conduct medical research
- Physical and mental suffering and injury should be avoided

In the decennia after the dissemination of the Nuremberg Code, the international medical community gradually developed more elaborate codes of ethical conduct in research, most notably the successive versions of the Declaration of Helsinki (World Medical Association 2010) and the guidelines of the Council for the International Organization of Medical Sciences (CIOMS 2010), the latter with increased relevance for research in low- and middle-income countries. CIOMS has recently produced international ethical guidelines for epidemiological studies (CIOMS 2010). Along with the response from the international medical community, there have been important milestones in legislation, mainly spearheaded by the United States. One such milestone was the publication in the U.S. Federal Register of the Belmont Report in 1979. A reprint of this important document can be found in McGuire-Dunn and Chadwick (1999). The Belmont report outlined three ethical principles upon which regulations for protection of human subjects in research should be based. These three principles are now widely known as:

- Respect for persons,
- Beneficence, and
- Justice/fairness in the selection of research subjects.

These have been the guiding principles for the U.S. Code of Federal Regulations (also reprinted in McGuire-Dunn and Chadwick 1999), and they have inspired similar legislation in other countries. The translation of these principles into guidelines and laws has been slow and progressive. It is worth noting, for example, that even in 1986 there were debates in major medical journals about whether fully informed consent was the appropriate thing to do (Simes et al. 1986). At that point the arguments against fully informed consent were still based on the abovementioned fallacious idea that highly ethical patient-doctor relationships were sufficient to protect research subjects. In that period it was also still possible to engage in trial participant dropout recovery programs without disclosing alternatives for similar-quality health care outside of the trial (Probstfield et al. 1986).

A very important recent process has been the development of Good Clinical Practice (GCP) guidelines for investigators of pharmaceutical products and medical devices. High-income countries with important stakes in the pharmaceutical industry initiated this process. The most important milestone publication is recognized to be the ICH-6 Guidelines (the International Conference on Harmonization 6, 1997), as this document provided a reference for clinical research in the European Union, Japan, and the USA. Since the ICH-6 Guidelines were released, the concept and practice of GCP have been more widely adopted, adapted, and expanded, and some have now been incorporated into legislation. Some countries have designed their own GCP guidelines (e.g., South Africa) adapted to local contexts.

3.5.2 History of Stakeholder Participation

Governments have always been important stakeholders of health research. The processes involved in research funding were relatively informal before WWII, but after the war the need for ethics review and for national and international funding agencies became clearer. Other important stakeholders include potential manufacturers and providers of remedies for illnesses. The dangers surrounding the relationship between physicians and pharmacists have been long recognized. In the earliest medical schools in Europe, for example in the School of Salernum (thirteenth century), there were strict prohibitions around any incentives given by ‘pharmacies’ to doctors. The twentieth century has seen the explosion of a huge pharmaceutical industry. This industry is now an important initiator of pharmacological research, a scenario that has led to great concerns about the validity of industry-funded studies, and indeed, problematic industrial incentives to doctors continue to exist. In modern times, the role of public-private partnerships in public health research is becoming increasingly important (Textbox 3.3).

Textbox 3.3 The Increasing Importance of Public-Private Partnerships

On one level, **government agencies** are now frequently involved in determining research priorities of the private sector. As an example, the USA Federal Drug Administration (FDA) directly influences pharmaceutical development projects by advising the sponsoring company on safety concerns that will need to be addressed.

(continued)

Textbox 3.3 (continued)

On another level, **private public health organizations**, such as The Gates Foundation, often partner with governments and organizations around the world to develop research priorities, implement necessary studies, and deploy demonstrably effective public health measures. Organizations such as these highlight the importance of international and global communities as stakeholders in health-related research.

On the ground level, members of the community are now frequently involved in reviewing study proposals and in establishing local research priorities (as in community-based **participatory research**). Consequently, public-private partnerships in health-related research have simultaneously become more globalized and more localized.

3.6 The Future of Epidemiology

3.6.1 Epi-Optimism

Reigning in some epidemiological circles over the past decades has been pessimism about the field. Part of this pessimism seems to be rooted in the observation that so many analytical studies on the same topic produce very different and sometimes contradictory results. We do not hold this view and wish to invoke a sense of optimism about epidemiology (epi-optimism). Indeed, the mere existence of inevitable inter- and intra-subject differences and the various types of study designs with many different approaches to dealing with effect modification and confounding predict that effect estimates will be highly different across studies, including clinical trials (Maldonado and Greenland 2002). As it appears, epidemiological thinking has yet to come to grips with the phenomenon of heterogeneity, which should no longer be seen as chaos but as the essence itself of theoretical occurrence relations.

We argue that the understanding of dogmatic concepts such as a ‘true relationship’ or ‘true effect size’ should become more nuanced. Scientific generalizability is a valid concept, but it is, in epidemiology especially, heavily ‘conditioned’ by heterogeneity in distribution matrices of confounders and effect modifiers. Another way of viewing heterogeneity is as an opportunity for achieving a deeper understanding of a disease process (*See*: Sect. 3.6.2.1). A greater ‘heterogeneity tolerance’ may positively influence the way epidemiology and epidemiologic study results are perceived by the wider public and, indeed, by future generations of epidemiologists.

3.6.2 The Focus of Future Epidemiological Research

3.6.2.1 Effect-Modification Research

Given the heterogeneity just described, epidemiology must shift its focus from searching for universal true relationships to documenting effect modification.

'Epi-pessimism' will hopefully give way to enthusiasm for more comprehensively studied effect modification, more uniformly reported effect modification in single studies, and better modeling of effect size differences in meta-analyses. Such a shift in thinking may have substantial consequences for the way studies are designed and results reported and interpreted. Sample size concerns, for example, will have to focus on the need to create credible evidence about a range of potential modifiers. These should include individual susceptibility factors as well as contextual factors. For intervention research, the contextual factors to be studied as effect modifiers include intervention delivery aspects and background factors. Part of the future may lie in collaborative multicenter studies involving diverse, well-documented distribution matrices of covariates. In scientific reports, recommendations such as "This relationship needs to be explored in other settings" could become more specific as to what effect modifiers should be better examined.

Greater study of effect modification will ultimately pave the road towards better-personalized care and better-adapted delivery of community interventions. The dogmatic concept of a single best treatment modality for all patients with a given condition will, through the study of heterogeneity, give way to the realization of individually-oriented interventions (i.e., 'personalized medicine'). As we advance towards personalized care, important questions will arise regarding research methods and their development.

3.6.2.2 New Diagnostic Research

As pointed out by Miettinen (2001), a vast area of diagnostic research remains virtually unexplored. This includes diagnostic prevalence studies, or in other words, diagnostic research that documents the probability of certain illnesses given a specific individual profile of antecedents, signs, symptoms, and diagnostic test results. The implementation of these ideas will be a huge but exciting challenge ahead and will rely partially on the development of methods for risk prediction modeling and more serious investigation of diagnostic performance tests (Miettinen 2011).

3.6.2.3 More Research on Research

The problem of publication bias reveals one of the weaknesses of the contemporary research process (*See also:* last section of this chapter). It would seem that more operational research is needed on research itself: where is research most likely to go wrong in individual studies, a collection of studies on a given topic, or even an entire field? When? Why? With the growing importance of Good Clinical Practice guidelines and regulations, data cleaning and other aspects of data handling should emerge from being mainly gray literature subjects to become the focus of comparative methodological studies and of process evaluations. Such types of studies should focus on the optimal procedures (balancing validity and cost-effectiveness) given local resources and cultural factors. Better understanding of processes in research will require epidemiologists to learn more from process analysts, psychologists, and social scientists.

3.6.3 Research Tools of the Future

The future will no doubt bring many paradigm shifts, changes in the use of terminology, new ethical challenges, new tools, and tools adapted from other scientific disciplines.

3.6.3.1 New Approaches to Study Design

To some extent, study design developments have tended to follow the identification of needs in research, and this is likely to continue. For example the need to study rare diseases quickly must have contributed to the refinement of the case–control design. Structural Causal Modeling is an example of a newly evolving area in etiognostic study design (*See:* Pearl 2010). Another is the single etiologic study design proposed by Miettinen (2010, 2011). In etiognostic research, the distinction between experimental and observational cohorts could become blurred: for example, mixed observational-experimental multinational cohorts may include long observational run-in periods to extensively document relevant effect modifiers before any experimental perturbation of determinants. After the intervention, continued observational follow-up of the cohort will become the rule, to determine long-term outcomes and to look at how responses to earlier interventions modify responses to later interventions.

3.6.3.2 New Research Databases

We are currently witnessing the emergence of large bio-banks of prospectively collected biological samples with addition of varying amounts of clinical, environmental, and behavioral information. These could give a boost to research and help to advance research methods, but the ethical and legal issues around making bio-banks internationally and easily accessible are not fully resolved (Kaye 2011; Zika et al. 2011).

There is a wider problem of public accessibility of research data in general. Epidemiology has yet to develop global, publically accessible banks of anonymized research databases. In other words, before deciding on setting up a new study involving the collection of new data, it should become possible for epidemiologists to find an answer to the question: where can I find an existing dataset that I could use to address the research question I have in mind? Perhaps one day most analytical studies will make individual participant data available for meta-analyses. Perhaps we should also expect more intelligent electronic libraries, semi-automated systematic reviews, global libraries of validated questionnaires or questions, and libraries of research methods for specific types of research questions.

3.6.3.3 New Assessment Technologies

New technologies will have a substantial impact on the development of epidemiology and of epidemiological research (Hofman 2010). The search for better and more objective measurement instruments will continue in medicine and outside of it; these innovations will continue to improve measurements in epidemiological research. To deal with confounding, mediation, and effect modification, a continuing challenge will be to measure the hitherto unmeasured. As measurement innovations come into play, scientific concerns will continue to prompt scientists to focus on measurement standardization. Although greater capabilities come with improved measurement tools, new technologies in epidemiologic research will raise newly encountered ethical challenges, both in health care and in health research.

Mobile phone technologies in particular are expected to have huge potential to improve measurements in epidemiologic research. The use of mobile phones for health purposes (irrespective of whether it is for personal, clinical, or research uses) is known by the generic term *m-health* (Vital Wave Consulting 2009; OpenXdata 2010). The interactive user interface may facilitate data collection, and thereby enable the large-scale diagnostic prevalence studies that are currently lacking. Phones are also easily adapted with other technologies, such as cameras, that allow imaging in the field and photograph-, video-, or audio-based data collection for analysis later.

Another challenge ahead in the near future is how to make optimal use of metabolomics, genomics, and proteomics. Integrating the "-omic" technologies and epidemiologic research are very challenging but not outside the realm of possibility (*See:* caBIG, as discussed in Textbox 3.4). There are currently still some problems with the validity of these approaches as methods for diagnosis and prognosis, but the "-omics" hold great promise for gaining an understanding of human health and illness and will therefore continue to be an important area for research in the future.

Textbox 3.4 The Future of Turning to Already Existing Databases

Many great questions are left unaddressed not because someone failed to think of the questions but because the researcher was unable to realize that evidence was at their fingertips. Substantial resources have been invested in the creation of **large databases**, such as the National Health and Nutrition Examination Survey (NHANES), and many of these are available to the research community-at-large. Data from many more studies are privately held by investigators worldwide. Among all of these public and private databases, one might be suitable to answer a research question raised by non-affiliated epidemiologists. Gaining access to that database would reduce the need to repeat the study; enable preliminary analyses that might be necessary to justify larger, more expensive studies; be useful for the design of other experiments (e.g., by estimating the variance of a factor under investigation); and facilitate meta-analyses using original data.

One could imagine the existence of a **database of databases** (DOD), where an investigator can search for variables and retrieve a list of all logged studies that contain them (or sub-components thereof). Such a DOD would address many of the issues addressed above in this textbox. Such a DOD

(continued)

Textbox 3.4 (continued)

raises issues of data exchange regulation and privacy to participants (among other things) for privately held databases, although these issues don't exist as prominently for databases already publicly available to the research community. Investigators might also find useful and productive collaborations around the world.

Currently, one of the best-developed systems to identify potential collaborators is **caBIG** (www.cabig.nih.gov), run by the NIH National Cancer Institute. caBIG has some data-sharing capabilities and provides excellent practical advice on how to make your databases shareable.

Other areas of research on measurement technologies that will be important to epidemiology in the future include:

- The development and application of nanotechnologies
- Three-dimensional imaging
- Safety assessment and monitoring of test products during research
- The assessment of human resources, including needs-based planning
- Qualitative methods (e.g., qualitative pilot studies on health among culturally and socioeconomically diverse countries; one current approach is the use of the Rapid Epidemiologic Assessment, promoted by the World Health Organization)

New analysis methods – The statistical analysis methods, as they are currently used in epidemiological practice, are nearly restricted by the easier options available in standard statistical packages. This situation has had some unfortunate consequences:

- It has contributed to a dominance of statistical testing over statistical estimation
- Within statistical testing, it has led to a nearly complete attention on null hypothesis testing
- It has lead to the failure of or delays in incorporating important new methods into standard software

Several eminent epidemiologists have warned against improper and excessive use of statistical testing. Some have even argued that statistical testing should be abandoned altogether (Rothman 2010) in favor of the use of statistical estimation. What we are likely to see, though, is a shift in balance towards more estimation than testing, not a complete disappearance of testing.

More and more causal effects are being demonstrated, more causal pathways have been progressively unraveled, and the complexity of causal networks leading to health-related outcomes has become better appreciated. Surveillance systems and health databases of the future will gather an increasingly wider array of longitudinal data on health determinants. Analysis methods will need to keep up with this evolution. For example, statistical methods to adjust for time-varying confounders have not yet found broad application, but this may change. Along with this evolution, applications of structural causal modeling, data mining, multilevel analyses, and related methods may gain prominence as the methods-of-choice for arriving at useful simplifications.

As to analysis aids, we can expect improved friendliness of statistical packages and an increased range of analyses included in them. Analysis tools in support of new study designs, as the ones proposed by Miettinen (2011), will hopefully be included.

3.6.4 The Future Architecture of Health Research

The continuing problems of publication bias (*See:* discussion point and Chap. 31) and of limited measuring and reporting of data quality unfortunately indicate that, after centuries of progressive sophistication of scientific methods, epidemiology is still too often defeated by subjectivity. It would seem, therefore, that behavioral sciences and epidemiology have a joint mission that promises many battles. The health research community and the International Committee of Medical Journal Editors seem rather slow in responding to the publication bias problem. And the registration of clinical trials has been insufficient to curb publication bias. A significant response is becoming a pressing need. Such a response will require the joint efforts of various stakeholders in research and will undoubtedly give an enormous boost to epidemiology.

3.6.4.1 Globalization in Health Research

Today, health care is considered to be a global public good and international and global initiatives to boost health research in specific domains are becoming more common (Keush et al. 2010). More and re-enforced consortia on broad topics of interest are needed (Nwaka et al. 2010) to provide better opportunities for, among others:

- Access to each other's cohorts, tools, data, publications, and expertise
- Multidisciplinary work
- Collaborative research grants
- Training and sharing of research management and ethics expertise
- Laboratory capacity building (Wertheim et al. 2010)
- Research-based partnerships with private sector, including the development and delivery of new health products (Keush et al. 2010)
- Communication between researchers themselves and policy makers (*See:* Chap. 30)

Publication bias is the skewed representation of the overall available evidence on a research topic in a body of published literature resulting mainly from the tendency of:

- Researchers to submit for publication only studies with positive findings (i.e., showing a statistically significant difference) and to withhold negative study findings (i.e., statistical results supporting the absence of effects)
- Journal reviewers to recommend acceptance of articles with positive findings and rejection of articles with negative findings
- Journal editors to preferentially send to peer-review and accept for publication articles with positive findings

Discussion Point What could be ways to combat publication bias?

Structural changes are needed to improve global fairness in access to research, research tools, and educational materials (Van den Broeck and Robinson 2007). Low- and middle-income countries have not been given enough support to build research capacity. Assertions have been made that one cannot adequately manage the clinical research process in resource-poor settings. It is important that this misapprehension, which contributes to perpetuating poverty, be resolved, and that all countries are given a chance to be involved. Although there are many challenges to high-standard clinical trial research in resource-poor settings, solutions are not far-fetched. There are many good examples of high-standard clinical research performed in low-income countries (*See:* Doumbo 2005). Research infrastructure – including staff, facilities, equipment, and training – can be developed in any setting provided appropriate funding is made available. The capability to perform clinical research does exist in most countries but needs more recognition by sponsors through stable, continued funding support and assistance in building centers of excellence (Van den Broeck and Robinson 2007). International and global networks and partnerships with the private sector will be crucial for this purpose, as will be an enhanced focus on research on 'neglected diseases' (Keush et al. 2010; Moon et al. 2010).

This chapter was the third in a series of chapters introducing epidemiology (Part I). Here we have touched on some of the roots of epidemiology and, to a lesser and more speculative extent, how these roots are expected to nurture fruits of the future. Among current epidemiologists different opinions exist about what a proper scientific epidemiological approach should be. Epidemiology is in motion. Yet, there is enough commonality in views and practices for the next chapters to contain a general description of modern study designs and implementation methods.

References

Beecher HK (1966) Ethics and clinical research. New Engl J Med 274:1354–1360

Council for International Organizations of Medical Sciences (2010) International ethical guidelines for biomedical research involving human subjects. CIOMS, Geneva. http://www.cioms.ch. Accessed Sept 2012

Dawber TR et al (1957) Coronary heart disease in the Framingham study. Am J Public Health 47:4–24

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

Doumbo O (2005) It takes a village: medical research and ethics in Mali. Science 307:679–681

Evans AS (1976) Causation and disease: the Henle-Koch postulates revisited. Yale J Biol Med 49:175–195

Franco G, Franco F (2001) Bernardino Ramazzini: the father of occupational medicine. Am J Public Health 91:1382

International Conference on Harmonization (1997) ICH-E6 guidelines. http://ichgcp.net/. Accessed Sept 2012

Hetzel BS (1989) The story of iodine deficiency. Oxford Medical Publications, Oxford, pp 1–236. ISBN 0192618660

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

Hofman A (2010) New studies, technology, and the progress of epidemiology. Eur J Epidemiol 25:851–854

Horner J, Minifie FD (2010) Research ethics I: historical and contemporary issues pertaining to human and animal experimentation. J Speech Lang Hear R 54:S303–S329

Irgens LM (2010) History of epidemiology. In: Killewo JZJ, Heggenhougen K, Quah SR (eds) Epidemiology and demography in public health. Elsevier, San Diego, pp 2–20. ISBN 9780123822000

Kahn LM, Semba RD (2005) They starved so that others be better fed: remember Ancel keys and the Minnesota experiment. J Nutr 135:1347–1352

Kaye J (2011) From single biobanks to international networks: developing e-governance. Hum Genet 130:377–382

Keush GT et al (2010) The global health system: linking knowledge with action – learning from malaria. PLoS Med 7(1):e1000179

Krieger N (2011) Epidemiology and the people’s health. Theory and context. Oxford University Press, Oxford, pp 1–381. ISBN 9780195383874

Maldonado G, Greenland S (2002) Estimating causal effects. Int J Epidemiol 31:422–429

McGuire Dunn C, Chadwick G (1999) Protecting study volunteers in research. A manual for investigative sites. Center Watch, Boston, pp 1–238. ISBN 0-9673029-1-9

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

Medical Research Council (1948) Streptomycin treatment of tuberculous meningitis. Lancet 1:582–596

Meinert CL (1986) Clinical trials. Design, conduct and analysis. Oxford University Press, Oxford, pp 1–469. ISBN 0195035682

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

Miettinen OS (1985) Theoretical epidemiology. Delmar, New York, pp 1–359. ISBN 0827343132

Miettinen OS (1999) Etiologic research: needed revisions of concepts and principles. Scand J Work Environ Health 25:484–490

Miettinen OS (2001) The modern scientific physician: 3. Scientific diagnosis. Can Med Assoc J 165:781–782

Miettinen OS (2010) Etiologic study vis-à-vis intervention study. Eur J Epidemiol 25:671–675

Miettinen OS (2011) Up from clinical epidemiology & EBM. Springer, Dordrecht, pp 1–175. ISBN 9789048195008

Moon S et al (2010) The global health system: lessons for a stronger institutional framework. PLoS Med 7(1):e1000193

Morabia A (2004) History of epidemiologic methods. Birkhaeuser, Basel, pp 1–405. ISBN 3764368187

Morabia A (2011) Santé: distinguer croyances et connaissance. Éditions Odile Jacob, Paris, pp 1–320. ISBN 9782738126283

Nwaka S et al (2010) Developing ANDI: a novel approach to health product R&D in Africa. PLoS Med 7(6):e1000293

OpenXdata (2010) Open-source software for data collection. www.openxdata.org. Accessed Sept 2012

Pearl J (2010) An introduction to causal inference. Int J Biostat 6(7):1–59

Probstfield JL et al (1986) Successful program for recovery of dropouts to a clinical trial. Am J Med 80:777–784

Rothman KJ (2010) Teaching a first course in epidemiologic principles and methods. In: Olsen J, Saracci R, Trichopoulos D (eds) Teaching epidemiology. A guide for teachers in epidemiology, public health and clinical medicine. Oxford University Press, Oxford, pp 77–89. ISBN 9780199239474

Saracci R (2010) Introducing the history of epidemiology. In: Olsen J, Saracci R, Trichopoulos D (eds) Teaching epidemiology. A guide for teachers in epidemiology, public health and clinical medicine. Oxford University Press, Oxford, pp 3–23. ISBN 9780199239474

Seife C (2004) Alpha and omega: the search for the beginning and end of the universe. Penguin Books, New York, pp 1–294. ISBN 0142004464

Simes RJ et al (1986) Randomised comparison of procedures for obtaining informed consent in clinical trials of treatment for cancer. Brit Med J 293:1065–1068

Susser M, Stein Z (2009) Eras in epidemiology. The evolution of ideas. Oxford University Press, Oxford, pp 1–352. ISBN 9780195300666

Tsuchiya T (2008) The imperial Japanese experiments in China. In: Emanuel EJ et al (eds) The Oxford textbook of clinical research ethics. Oxford University Press, Oxford, pp 31–45. ISBN 9780195168655

Van den Broeck J, Robinson AKL (2007) Towards research equity – challenges of safety monitoring during clinical trials in resource-limited settings. West Indian Med J 56:163–165

Vital Wave Consulting (2009) mHealth for development. The opportunity of mobile technology for healthcare in the developing world. Washington DC, and Berkshire,: UN Foundation and Vodaphone Foundation Partnership

Wertheim HFL et al (2010) Laboratory capacity building in Asia for infectious disease research: experiences from the South East Asia Infectious Disease Clinical Research Network (SEAICRN). PLoS Med 7(4):e1000231

White RM (2000) Unraveling the Tuskegee study of untreated syphilis. Arch Intern Med 160:585–598

World Medical Association (2010) The declaration of Helsinki. http://www.wma.net/en/10home/index.html. Accessed Sept 2012

Zika E et al (2011) A European survey on biobanks: trends and issues. Public Health Genomics 14:96–103

Part II

Study Design

Chapter 4: General Study Objectives
Chapter 5: The Specific Aims
Chapter 6: General Study Designs
Chapter 7: Study Size Planning
Chapter 8: Funding and Stakeholder Involvement
Chapter 9: The Recruitment, Sampling, and Enrollment Plan
Chapter 10: The Measurement Plan
Chapter 11: The Quality Assurance and Control Plan
Chapter 12: The Data Management Plan
Chapter 13: The Analysis Plan
Chapter 14: Ethics Support

General Study Objectives

4

Jan Van den Broeck and Meera Chhagan

All men by nature desire knowledge.

Aristotle

Abstract

This chapter provides advice on the identification, justification, and formulation of general study objectives. There are five major types of research topics that can be addressed: diagnostic, etiognostic, intervention-prognostic, descriptive-prognostic, and methods-oriented topics. Within each major type we discuss topics in clinical medicine separately from topics in community medicine. Commonly, the researcher has many research questions, perhaps as a result of previously conducted research, but needs to include into the study rationale the interests of stakeholders, the virtual importance for public health, and the availability of resources. Decisions to do a study may require an updated insight into existing evidence on the topic with the aim of identifying knowledge gaps. We therefore briefly discuss methods of the literature review. One considers at this earliest stage of planning that not all research requires new data collection; other potential sources of new evidence include existing research databases, and the joining of ongoing studies.

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

M. Chhagan, Ph.D., FCPaed
Department of Paediatrics, University of KwaZulu-Natal, Durban, South Africa
e-mail: Chhagan@ukzn.ac.za

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_4,

4.1 Types of Research Topics

Epidemiological research aims to augment the knowledge-base of clinical medicine and community medicine, and the discipline of epidemiology provides the methodology to achieve this aim. General scientific knowledge in support of diagnosis, etiognosis, and prognosis (with and without intervention) of illnesses, considered together with the particular profile of the patient, allow the scientifically educated health care worker to propose clinical decisions to the patient.

Epidemiology also provides methods to produce estimates of the burden of illness (past, present, and projected future with or without intervention) and capacity of health care in populations, thus contributing to informed public health intervention decisions.

This chapter deals with the types of issues that are often found compelling as topics for investigation. In Chap. 1, we noted that the typology of research questions includes descriptive and analytical studies. Here, we refine that typology by describing the five major types of research questions:

- Diagnostic
- Etiognostic
- Intervention-prognostic
- Descriptive-prognostic
- Methods-oriented

This typology builds on work by Miettinen (2002). In addition to the four types proposed by Miettinen we included a fifth type, the methods-oriented research question. The justification is that not all studies aim directly at creating medical knowledge. Some studies aim to contribute to this only indirectly by creating or improving the methodology to be employed in other epidemiological studies that more directly aim at creating medical knowledge. This latter type of research is sometimes referred to as ‘design research.’ Table 4.1 lists these five types with brief examples of research questions and annotations regarding whether or not they address causality.

For the researcher conceiving a new study it is crucial to be able to place the general aims correctly in one of the five types because the consequences in terms of study design options are important. To enable that critical task, we describe the five types in detail in the next sections (using Panel 4.1 terminology), distinguishing questions asked in clinical settings from questions asked in community medicine, as this distinction also has important implications for study design (*See:* Chap. 6).

4.2 Diagnostic Research

4.2.1 Diagnostic Research Questions in Clinical Medicine

Most illnesses (diseases, defects, states and processes resulting from injuries) are readily classifiable according to the International Classification of Diseases ICD-10 (WHO 2010). They are known to have and are often defined on the basis of having

Table 4.1 Types of research questions in epidemiology

Type of general aim	Causal orientation	Example of a research question (abbreviated in the form of a title)
Diagnostic	No	Community medicine: Gender inequality in the incidence of H1N1 infection
		Clinical: Signs and symptoms of patients presenting with H1N1 infection
Etiognostic	Yes	Community medicine: Effect of infrastructural factors on H1N1 attack rate
		Clinical: Effect of hand hygiene practice on the risk of H1N1 infection
Descriptive-prognostic	No	Community medicine: Prediction model for resurgence of small area H1N1 epidemics
		Clinical: Risk prediction model of bacterial pneumonia in H1N1 infection
Intervention-prognostic	Yes	Community medicine: Effect of hand hygiene promotion campaigns on H1N1 incidence
		Clinical: Effect of antiviral treatment on the duration of illness from H1N1
Methods-oriented	No/yes	Validation of a simplified tool for measuring nutrition knowledge in children
		Causes of observer error in anthropometric measurements

combinations of signs, symptoms, and lab results, depending on the severity and stage of illness. Most have known risk factors and medical antecedents. The traditional clinical diagnostician, keeping in mind that unusual combinations do occur and that patients often present with more than one illness, uses this knowledge for comparison with a presenting patient's profile of antecedents, risk factors, signs, symptoms, and lab results. Based on this mental (subjective) comparison, a set of differential diagnoses emerges in the mind of the clinician, as well as an idea of what sequence of additional signs and lab tests could be useful to progressively restrict the differential diagnostic set until a classification ('the diagnosis') is reached. Parallel decisions may further relate to the desired degree of illness subclassification. In settings with very restricted resources, for example, one may even forego classical diagnosis and conclude to limit the diagnostic assessment to a mere combination of signs and symptoms compatible with various illnesses but considered detailed enough (usually based on a perceived high probability for one illness) to usefully base an intervention upon it (Van den Broeck et al. 1993). There are several ways in which epidemiological research can assist with the diagnostic process roughly described above.

The traditional types of *clinical diagnostic research* listed in Panel 4.2 have proven their usefulness, but they also have an important limitation: such research tends to have a design with a backward directionality. The approach of traditional diagnostic research studies is usually to look back at certain antecedent features and test results in cases of a specific illness and non-cases. The practical problem of the

Panel 4.1 Selected Terms and Concepts Relevant to Conceiving and Formulating General Study Objectives in Epidemiology

Adverse effects Unintended undesirable effects of intervention, foreseeable or unforeseeable

Course of illness Temporal changes in presence, severity, and sickness associated with illness and illness complications

Defect Structural somatic deficit

Diagnosis (Process of gaining) probabilistic knowledge about the presence of a defined illness (*based on* Miettinen 2001a, b)

Diagnostic profile The set of signs, symptoms, antecedents and test results, present at some stage of the diagnostic process, taken into consideration at that point by the diagnostician as information relevant for decision making about next steps towards diagnosis

Differential diagnostic set Set of illnesses still under consideration at the current stage of the process of diagnosis

Disease Pathological somatic process

Effectiveness (1) Compliance- or coverage-dependent efficacy (2) Balance of the modifying effects of negative and positive modifiers of the compliance- or coverage-dependent efficacy

Efficacy Whether or not, or, degree to which, the intended effect of an intervention is achieved

Efficiency The reciprocal of the resources spent to achieve a defined goal

Epidemic Pattern of illness occurrence in which the incidence of the illness exceeds expectation

Health Freedom from illness (Miettinen 1985)

Illness Presence of disease, injury or defect

Injury Infliction on the body causing a defect and/or a pathological process

Latent illness Illness undiagnosed on behalf of a lack of illness manifestations

Literature review A summary and interpretation of the body of evidence existing around a research question

Morbidity The distribution of illnesses in a population (Miettinen and Flegel 2003)

Mortality The occurrence of death in a population

Placebo Mock intervention

Prognosis Expected future course (Miettinen 1985)

Prognostic profile Set of attributes or experience indicative of the future course of illness or morbidity pattern

Screening regimen Scheme of successive assessments/tests leading to early diagnosis and treatment of asymptomatic patients with a defined illness

Secular trend Currently refers to a trend over a very long calendar period of at least 15 years; Formerly referred to a trend over a century (*Latin: S*aeculum, Century)

Test intervention An intervention willingly introduced to study its effects on individual health or morbidity/mortality

Test product Substance of which the effects are assessed in a trial

Panel 4.2 Traditional Types of Diagnostic Research Questions in Clinical Medicine

- Frequency of **illness manifestations** by severity, natural history, and medical antecedents
- Description of **normal development,** e.g., growth standards
- Usefulness of (sequences of) **diagnostics tests** mostly as judged by their so-called 'predictive' value or by their likelihood ratio

clinician, however, is not how to arrive at the diagnosis of one single pre-specified illness. The problem is of a totally different, forward-oriented nature. Specifically, it is about knowing what is the differential diagnostic set associated with a presenting patient's profile and which sequence of questions, signs, examinations and tests leads to the fastest and most efficient narrowing of that differential diagnostic set. Miettinen has pointed out a potentially more useful (and relatively neglected) type of diagnostic research study design (Miettinen 2001a, b), one in which there is forward directionality – the diagnostic prevalence study, also called the diagnostic probability study (Miettinen 2011 and Panel 4.3).

Diagnostic prevalence studies can produce tools, such as *diagnostic probability functions*, that may have useful applications in clinical practice (Miettinen 2011) (*See also:* Chap. 24). Miettinen also notes that, with such applications, the usefulness of doing an additional diagnostic test should be determined based on how much the post-test probability for the illness will increase compared to the prior probability of the illness given the patient's profile (Miettinen 2001, 2011). This new paradigm, although very compelling, has not yet been widely accepted in epidemiology. In the future, this type of diagnostic research may use artificial intelligence based tools.

As pointed out above, the diagnostician's main goal is to arrive at a diagnosis for ill patients presenting with a certain diagnostic profile. But not all people with a disease display overt clinical signs or symptoms; that is, some people with a disease are in a latent phase. Thus, another concern in clinical medicine is the diagnosis of *latent* cases of illness through screening, especially for illnesses that tend to have a better prognosis when diagnosed and treated earlier rather than later.

4.2.1.1 Screening

Screening is the application of a screening regimen aimed at diagnosing latent cases of illness. The screening regimen always starts with an initial test to identify individuals with a high enough probability of latent illness to warrant one or more further tests, and so on, until a final diagnosis of latent illness is reached. Development and evaluation of a screening regimen involves answering intervention-prognostic research questions. Relevant research questions about the diagnostic productivity of a screening regimen are listed in Panel 4.4 (*See also*: Miettinen 2008):

Panel 4.3 Research Questions Addressed by Diagnostic Probability Studies

- Given a particular individual profile of antecedents, risk factors, signs, symptoms and test results, what is the probability (prevalence) of having a defined illness?
- Given such a profile, which illness out of a differential diagnostic set is most likely?
- Which sequence of tests, by adding the test results to the individual's diagnostic profile, has the greatest and fastest potential of singling out the 'true' illness or illnesses?

Panel 4.4 Research Questions Around Screening

- In what proportion of people does applying the screening regimen produce a diagnosis at a latent stage, and is this proportion higher than for diagnoses made outside the screening regimen?
- How frequently are healthy people unnecessarily subjected to the further diagnostic work-up after the initial screening test?
- How frequently does the initial or follow-up test lead to complications?
- What proportion of cases of latent illness remains undiagnosed and perhaps falsely re-assured in spite of participating in the screening regimen?
- What is the probability of diagnosing of latent illness in a screened person as a function of age and other personal characteristics?

4.2.2 'Illness Burden' and 'Response Capacity' Questions in Community Medicine

Whereas clinical diagnosis focuses on overt or latent illnesses of individual patients, community health 'diagnosis' focuses on the burdens of illnesses in *populations*. Clinical diagnosis eventually informs proper treatment. Likewise, knowledge of the burdens of illnesses and response capacity in a community can help in making proper decisions about how best to respond, e.g., through health education, the organization of health care, etc. Community health workers often engage in surveillance, assessment, improving response capacities, health education, vaccination, and other community services. In a way, community epidemiology allows for the achievement of knowledge needed for the fair allocation of public resources to activities that will best enhance health of the population. This includes knowledge of monetary costs of illnesses and interventions. The decision of what constitutes fair allocations of public resources is not straightforward and needs to be based at least partly on knowledge about burdens and response capacities. Decisions also need to be based on the acceptability and preferences of the population concerned and to be brought about in a participatory manner.

Panel 4.5 Diagnostic Research Questions in Community Medicine

- What is the current burden of illnesses and risk factors in the community, in terms of prevalence, incidence, severity, relative frequency distribution, or clustering within individuals?
- How do illnesses cluster in time and space (Some of the most ancient types of research questions in epidemiology are those relating to short-term temporal-spatial clustering of illnesses: epidemics)?
- How do illness burden patterns evolve over longer calendar time periods (sometimes over very long periods: secular trends)?
- What resources are available to tackle illnesses in the community? What is the availability, accessibility and functionality of health services e.g. human resources in health? What are the monetary costs of possible interventions?
- What are the inequalities (gaps, disparities) in health, health education, health information and health care among subpopulations defined by sex, age strata, ethnicity, built environment, socio-economic status, country regions, countries and world regions?

Whereas most of this burden and response capacity research is particularistic (i.e., aiming at characterizing the burden or response capacity in a particular population), there is also research that aims at generalization 'into the abstract' beyond the particular study population. For example, epidemics of specific illnesses often have shared characteristics, and their natural evolutions seem to follow certain patterns that are amenable to scientific investigation. Indeed, illness burden and response capacity research poses a variety of types of research questions that are shown in Panel 4.5. This can be called '*community-diagnostic research*'.

4.3 Etiognostic Research

4.3.1 Etiognostic Research Questions in Clinical Medicine

The clinician has a natural interest in knowledge about causes of illness and in knowing to what extent these causes have played out in a particular patient. Knowledge of the causes of illness aids in diagnosis by allowing identification of antecedents of disease in future patients. But such knowledge has other important uses as well. For instance, it may allow targeted actions to prevent the worsening of the patient's condition or sometimes even to cure the patient. And by extension, it may also help to define actions that might prevent the illness in that patient's family members. Even on a much broader level, knowledge of causes of disease often is the basis of general health advice to patients who do not yet have an illness in question (e.g., protection from heart attack by eating a diet rich in soluble fiber).

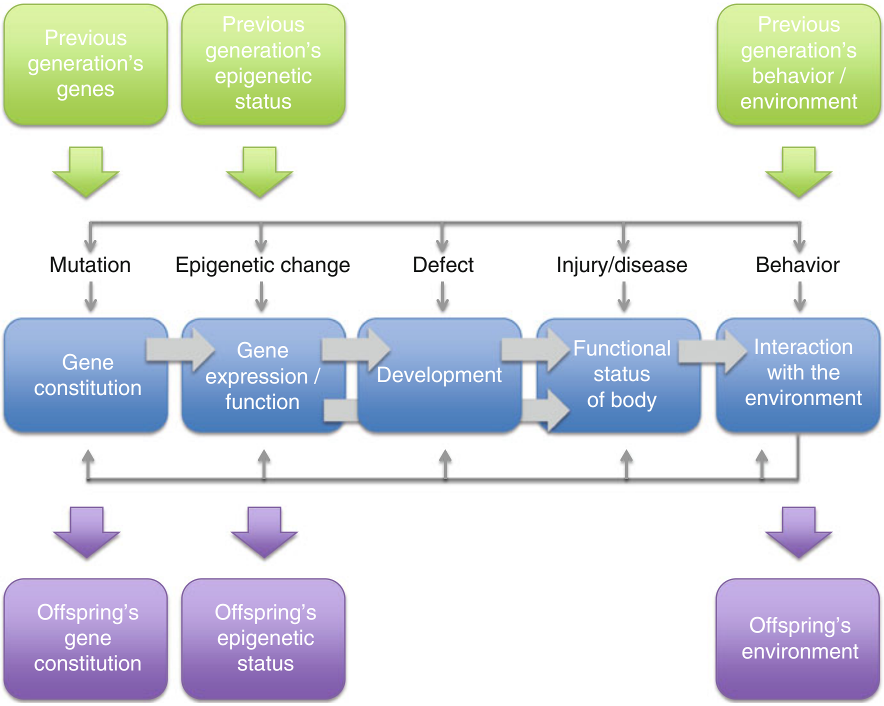

Previous generation's genes
Previous generation's epigenetic status
Previous generation's behavior / environment
Mutation
Epigenetic change
Defect
Injury/disease
Behavior
Gene constitution
Gene expression / function
Development
Functional status of body
Interaction with the environment
Offspring's gene constitution
Offspring's epigenetic status
Offspring's environment

Fig. 4.1 Simplified diagram of categories of etiologic factors affecting health

All diseases and developmental defects have genetic *and* environmental causes. Figure 4.1 shows a diagram of categories of etiologic factors affecting individual health. Gene expression continuously requires interaction with the environment e.g. nutrients. It is this interaction that allows the formation and maintenance of a functioning human body. This interactive process starts in utero with little more than a collection of recombined genetic codes embedded in a mainly maternal environment. From that moment somatic-functional development proceeds but can be delayed, accelerated, disharmonized, or arrested prematurely, locally or entirely. No matter how this development has worked out in the particular presenting patient, ultimately the process will end, either after slow degenerative processes, or after bursts of decay caused by injury and disease, or very suddenly by a fatal event. In the meantime, the whole process will have supported a unique human experience, always worth living and an end-on-itself. All individuals transmit knowledge, environment and sometimes genes to next generations.

From this it follows that there are three broad classes of factors causally related to individual health i.e. three types of factors related to the success of the constructive and maintenance stages of the described interactive process:

- Genetic and constitutional
- Environmental
- Behavioral

Clinical etiognostic research focuses on the extent to which particular individual exposure experiences (broadly speaking episodes of gene/constitution – environment interaction through behavior) affect measurable aspects of somatic-functional integrity in a causative or preventive way. Whenever an undesirable health-related event has occurred or an undesirable state developed in a patient, a multitude of such experiences obviously has preceded throughout the patient's lifetime up till that point. Always, previous generations have contributed, individual susceptibilities have developed, societal factors have played out, physical-chemical and biological factors have had their influence. The question is thus not whether, in this patient, trans-generational, behavioral, societal, constitutional or environmental factors are causally linked to the outcome. They are!

The question in clinical etiognostic research is, rather: are there any specific types of experiences that, if they would or would not have happened or if they would have been made more intense or less intense, through some purely hypothetical modifying action, could have prevented the outcome or made it less severe in at least a proportion of patients, and, in what proportion of patients?

To answer such questions researchers have often addressed one or very few potential risk factors at the time in their studies. This approach has been labeled 'single risk factor epidemiology'. As risk factor epidemiology unveils the importance of increasing numbers of related causal factors to the same health related states and events, there is an increasing need for a form of integration linking the various causes in complex hierarchical models. Interestingly, single risk factor epidemiology and complex multilevel modeling of causal pathways have erroneously been presented as very different paradigms. We rather see that one complements and reinforces the other in several useful ways. For example, data mining exercises sometimes come up with best models that do not seem to make any intuitive sense, sometimes because part of the variables considered are unrelated to the outcome as known through single risk factor epidemiology. The selection of variables for consideration in complex models should be based at least partly on evidence from single risk factor epidemiology and common sense.

Occupational medicine has an interest in the causal role of exposures in workplaces on the occurrence of illnesses (Panel 4.6). In occupational epidemiology the exposures suspected to influence health are often obvious from the kind of work being carried out. For example, in agricultural workers, the health consequences of exposure to pesticides are a topic of interest. Among hospital personnel it is needle-sticks and hospital pathogens that are of special concern. However, particular situations may arise in epidemiology when it is not clear from the outset what the exposures of interest actually are. The task may simply be to investigate 'the causes' of a worrying increase in number of cancer cases in the hospital or workplace or to discover 'the source(s)' of contamination in some localized infectious epidemic. To address this kind of question, qualitative or semi-quantitative preliminary investigations may need to be carried out to identify and specify the potential causes worthy of including in the main study's object design. This exercise requires, from the part of the researcher, scientific knowledge of the etiology of the outcome and particularistic knowledge of research settings and areas. Small qualitative research projects or 'rapid assessments' may help to refine this knowledge.

Panel 4.6 Etiognostic Research Questions in Occupational Medicine

Research questions may concern the potential causal role of:

- **Ergonomic hazards**, e.g., lifting heavy loads, high-risk situations for injury, straining body postures, long working hours with computers
- **Psychosocial hazards** at work, causes of job-related stress
- Undesirable **environmental exposures**: dust, dirt, noise, toxic chemicals, biological substances; the interest here may be in specific agents or in mixtures, or even in the effects of broad, incompletely characterized exposure situations

Note that the result of such preliminary situation assessments may be so convincing in pinpointing a cause that further scientific epidemiologic study is considered unnecessary.

4.3.2 Etiognostic Research Questions in Community Medicine

Community health questions arise about factors causally linked to health burdens, disparities in burdens, and changes in burdens in populations. Research addressing such topics is called *community-etiognostic research*. Observation units in this type of research may be individuals, 'geographical areas,' or other groups. The exposures of interest may be the same as in clinical etiognostic research, comprising the whole spectrum of constitutional, environmental, and behavioral factors. Again, it may not always be clear from the beginning of such a research project what the exposures of interest are (for example, when one starts investigating the causes of an unexplained rise in incidence of cancer in a particular sub-area revealed by surveillance). Community-etiognostic research may also concern the *impact* of policy interventions that were implemented non-experimentally outside research contexts. Ecological variables are also frequently of interest as exposure variables.

4.4 Intervention-Prognostic Research

Etiognostic studies are not the only type of studies that address cause-and-effect relationships. Other types that equally have such an 'analytical' aim include some methods-oriented studies (*See:* below) and intervention-prognostic studies. With the latter, the issue of interest is whether a change in outcome would be brought about by introducing a particular *test intervention* compared to *no intervention* or *another intervention.* Among the latter can be a 'mock intervention' or placebo. The issue addressed is fundamentally different depending on what type of reference intervention level will be used. Comparison with 'no intervention' addresses the *full effect* of the intervention on the outcome, whereas comparison with another intervention, addresses the *difference in effect* between the two interventions (whether or not one of them is a ‘mock intervention’). Addressing full effects is often unethical as it tends to mean leaving part of the patients or communities suffering without help. Most research questions are thus geared towards comparing alternative intervention strategies, notably in situations where there is equipoise as to the possible superiority of a test intervention.

Several aspects of the interventions need to be compared:

- Firstly, interventions can have multiple intended effects, and (compliance-dependent) efficacy in achieving the effects may need to be compared. The interest may be in the existence of an effect, it’s size, or its modifiers or mediation
- Secondly, interventions can have unintended beneficial and adverse effects. As to the latter, one is interested in studying the incidence, timing, and severity of those that are foreseeable and of those that are not foreseeable. Undue effects may be associated with elements of the intervention strategy, or there may be special risks associated with poor compliance or with stopping an intervention once it has been started
- Thirdly, all interventions have various types of costs both to participating individuals concerned and to communities, and these costs have a level of acceptability attached to them for the individuals concerned, for society, and for policy makers. This latter type of issue, however, belongs to the diagnostic domain (as described above), not the intervention-prognostic domain

In both clinical and community health research the comparisons of these different useful properties can usually best be made separately. Indeed, it will be up to the individual patient to weigh knowledge on effectiveness, safety risks, likely individual/family costs and thus (s)he must be informed about these aspects separately. Likewise, for community health, the expected effectiveness at the expected degree of coverage must be weighed against expected public costs and acceptability issues, but each community and group of policy makers has different problems and priorities.

4.4.1 Intervention-Prognostic Research Questions in Clinical Medicine

Intervention may be needed, even in the absence of or before a refined diagnosis, to stabilize vital functions and relieve pain. Thus, in emergency medicine and nursing there are important research questions about how to achieve this in the most efficient and safe way, and, if possible, in a way that will not make subsequent refined diagnosis impossible. When resources are available and the patient, guardian, or close relatives permit, a refined diagnosis and a detailed individual profile of prognostic indicators (including contra-indications, markers of responsiveness, etc.) can be made to inform and propose a specific intervention strategy to the patient. The stated intended effects of that intervention strategy may be (in order of preference ignoring safety, cost, and preference issues) of the types listed in Panel 4.7.

Intervention-prognostic clinical research (also called *clinical intervention research*) usually compares intervention strategies that have the same type of

Panel 4.7 Possible Intended Effects of a Clinical Intervention

- To cure or to speed up cure
- To improve the health state
- To stabilize the health state
- To slow worsening of the health state
- To diminish suffering without an intended effect on a health state itself
- To prevent future illnesses in that individual or in others, e.g., through genetic counseling or prophylactic (preventive) measures for communicable disease

intended effect, as listed above. Occasionally, however, it makes sense to compare a strategy to cure, at considerable foreseen safety risk, with a strategy to improve only but with fewer foreseen safety risks. In such instances the choice of an appropriate effectiveness outcome parameter may be more challenging. Types of clinical intervention strategies that are often studied include new drugs, drug dosing regimens, and routes of administration as well as technical health care interventions (such as surgical operations) and composite therapeutic strategies/regimens.

Discussion Point The purpose of intervention-prognostic research *cannot* be to document the harm caused by an intervention known to have (a high chance of) a harmful effect. This limitation, imposed by the general ethical principle of *non maleficence,* has not always been taken seriously by medical researchers.

Importantly, clinical intervention research should not consider all patients with the same diagnosis as equal and simply study average outcomes of intervention strategies in large groups or in a few disease severity and age/sex categories, as has too often been done in the past, without much concern for the modifying role of the individual patient profile. Knowledge on intervention strategies is incomplete, also in the common sense view of the patient him/herself, without a focus on how the individual patient profile, including the stage of the illness at the start of treatment, influences the intervention-prognosis (*See also:* Chap. 24, Sect. 24.4). In the context of screening for latent illness, the modifying effect of illness stage on the treatment effect is one of the issues to investigate.

In clinical intervention research, one is often faced with the difficulty that not all aspects of an intervention strategy can be studied simultaneously. Thus, the corresponding research questions are often addressed in different phases (Panel 4.8).

Panel 4.8 Types of Trials According to Drug Development Phase

- **Phase-1 trials** – Pharmacologic studies on a limited number of healthy volunteers, after animal experiments have shown acceptable results. The purposes are short-term safety profiling, tolerability assessment, and pharmacologic profiling (absorption, blood levels, elimination) depending on the dose and route of administration
- **Phase-2 trials** – Small-scale trials, done after phase-1 trials have shown acceptable results. The purposes are to further assess safety and sometimes efficacy in a limited number of patients, usually 30–300. Phase-2 trials can provide proof of principle that the treatment works or works at least as well as the reference treatment, though effect sizes are not usually possible to estimate reliably.
- **Phase-3 trials** – Large-scale trials done after phase-2 trials have proven acceptable safety. These studies are always randomized and involve large numbers of patients. Detailed efficacy profiling is done, including estimates of *effect size* and the identification of *effect modifiers*, i.e., the role of individual intervention-prognostic profiles. Medium-term safety profiling is also assessed, usually in a more rigorous manner than in previous phases
- **Phase-4 trials** – Post-marketing studies done after licensing and marketing. The main purpose is surveillance of long-term safety and efficacy as well as survival. Sometimes new studies are done after marketing to look at specific pharmacologic effects and specific risk profiles

4.4.2 Intervention-Prognostic Research Questions in Community Medicine

Public health professionals intervene in communities, such as during infectious epidemic outbreaks, or propose structural interventions to policy makers. They can also propose changes to clinical intervention strategies. The knowledge-base for these types of activities partly rests on intervention-prognostic research, although, as we have mentioned, observational-etiognostic research can also provide evidence about the impact of policies and interventions, specifically those that were implemented non-experimentally outside research contexts. Intervention-prognostic research questions addressed in community medicine often concern primary prevention methods (e.g., vaccines, health care delivery strategies, health education, and infrastructural interventions). The stated interest may be, among others, in the potential outcomes listed in Panel 4.9.

Intervention-prognostic research in community medicine (*community intervention research*) may address the potential for (1) coverage-dependent effectiveness after wider scale implementation, and (2) unintended 'collateral' effects, for example effects on (inequalities in) other disease burdens. To make a parallel with the clinical

Panel 4.9 Possible Intended Effects of a Community Intervention

- The disappearance of an illness from a community, e.g., elimination of polio
- To decrease the total burden of an illness in current or next generations
- To slow an ongoing increase in the size of an illness burden
- To decrease disparities and inequalities in an illness burden
- To prevent a burden of zero form becoming non-zero, or to prevent the development of an inequality
- To develop intervention strategies for specific illnesses or groups of illnesses (including making them more efficient and less costly, so as to free up resources to combat other burdens)

research, this type of research needs more attention to particular profiles of prognostic factors that modify the relationships and predict that the outcomes will be different for different community strata and communities.

4.5 Descriptive-Prognostic Research

Etiognostic and intervention-prognostic research questions have an 'analytical' aim: they address cause-and-effect relationships. This is in contrast with descriptive-prognostic research questions. This study type aims at predicting future changes in health states or health state distributions. Indeed, prediction can sometimes be made without knowledge of causation, and knowledge of causation does not necessarily allow for efficient prediction. A risk factor can be strongly associated with an outcome yet poorly predictive of it (Ware 2006). For example, smoking is strongly related to lung cancer but poorly predictive of it. It is true, however, that strong causative or preventive factors tend to be better predictors than a-causal factors. Of interest in descriptive-prognostic research can be single predictors of an outcome of interest, or how several prognostic indicators jointly predict an outcome of interest.

4.5.1 Descriptive-Prognostic Research Questions in Clinical Medicine

The interests of both the clinical health worker and the patient are, naturally, the probabilities of possible future courses of the patient's illness(es), including possible complications; the probability of newly acquiring another or the same illness; and sometimes the duration of the patient's life. Naturally also, the interest of both the clinical health worker and the patient lies in knowledge of how the patient's individual profile of prognostic indicators (age, sex, interventions, etc.) affects all of these probabilities. Such research topics are addressed in what is commonly known as *clinical prediction research*.

4.5.2 Descriptive-Prognostic Research Questions in Community Medicine

In community medicine descriptive-prognostic research questions are analogous to the ones posed in clinical medicine. Health policy makers, community health workers, and the general public are interested in knowing how current health burdens and their inequalities (as well as the relevant response capacities to address them), are likely to change in the future. There are also questions about the probability of new epidemics and of the recurrences of epidemics. Finally, community health workers and public are interested in how particular prognostic indicators of (sections of) communities modify all of these probabilities. The type of research addressing these topics is often referred to as *forecasting research*.

4.6 Methods-Oriented Research

All of the types of research questions discussed so far are addressed in studies that employ epidemiological tools: abstract design tools as well as instruments and other equipment. All these tools have a certain degree of validity, efficiency and ethical acceptability. Among alternatives, what tool is most valid, efficient, or acceptable is not always clear or easy to determine. Many investigators resort to using traditional study designs, ‘accepted tools,’ and ‘well-trained observers’ without too much further concerns about any limitations and their potential consequences. Some investigators will do pilot studies to learn more about validity, efficiency, and acceptability issues before starting the actual study. And some will collect special data on accuracy and precision during the course of an ongoing study. However, it also happens that studies are specifically set up to investigate methodological issues alone, separate from any ‘mother study’. Such studies tend to aim at verifying, expanding or refining the epidemiological toolkit or at ‘locally adapting’ an existing tool for later use by others.

Methods-oriented studies can address aspects of performance and usefulness of new or potentially improved observer selection and training schemes, sampling schemes, instruments, tests, quality control methods, data management methods, analysis methods and other aspects of methodology. They can focus on accuracy, precision, cost and other efficiency aspects, and on acceptability issues. Such studies are sometimes referred to as ‘operational studies’ (examples of questions are given in Panel 4.10).

Epidemiologists have traditionally focused on measurement methods, and there has been much less interest in potential improvements to methods of recruitment and data management/handling.

Finally, it is worth noting that methods-oriented studies can have descriptive or analytical aims. As an example of the latter, one may investigate factors causally related to measurement error, data handling error, or analysis error. Ultimately, the goal of such studies is also to improve the epidemiological toolkit albeit more indirectly.

Panel 4.10 Examples of Important Questions in Methods-Oriented Research

- What is the validity of this **new measurement method** in comparison with a more invasive gold-standard method?
- Can we replace this traditional measurement method with a new one that is **cheaper and simpler?**
- How can this measurement method be optimized for use in **another setting** or in **another type of patient**?
- Can observer error be reduced by better **standardization** of some aspect of the measurement technique?

4.7 Choice of Topic and Source of Evidence

The possible range of topics to study is probably infinite. Some form of prioritization is thus required (Viergever et al. 2010; *See also:* Chap. 8: Funding and Stakeholder Involvement). A compelling study rationale from a public health perspective precedes concerns about feasibility and study design, meaning that the choice of a topic is one of the first issues that must be considered in detail. After having identified a topic and its rationale, it is an ethical imperative to carefully investigate all possible sources of valid evidence, including prior studies as well as already-established databases before collecting data on new participants.

4.7.1 Multiple Research Questions in the Same Project

Addressing multiple research questions in the same study seems to be the logical thing to do from an efficiency point of view. And, indeed, this is becoming the rule: studies with a single research question are nowadays rare. Health surveys, for example, involve a large number of outcomes, and clinical trials always have efficacy *and* safety outcomes and may also address prediction issues (Miettinen 2010).

A problem with multiple outcomes can arise when one wishes to make a clear distinction between primary and secondary objectives of a study. The primary objective demands the best information and gathering that information is of prime importance; this is one reason that estimates of optimal study size tend to be geared towards the achievement of the primary objective. Gathering ample information on an array of secondary research questions can constitute a distraction from the primary objectives and dilute the precision and decrease the accuracy of information collected on the main outcome.

Multiple outcomes may also be planned with the goal of analyzing them together in a single multivariate analysis. This approach can be useful when the researcher suspects and intends to examine whether all these outcomes are related to a same set of determinants (and with what strength). In addition, multiple outcomes may be targeted in a study not only because there is a separate interest in the occurrence of each, but also because they are conceived as belonging to a same single construct, e.g., the construct of intelligence. In that case the aims may be to study each outcome separately and to combine them into a single score, e.g., an intelligence quotient. Another example is a study of the effect of a treatment on preventing malignant neoplasm. The desired information is the occurrence of various specific types or classes of cancers in addition to the overall occurrence of all classes combined into the construct ‘malignant neoplasm.’

4.7.2 Existing Summarized Evidence

Having developed an interest in a certain topic and a set of related research questions, the epidemiological researcher is often faced with the task of updating her/his knowledge of any relevant evidence. Experienced researchers tend to be broadly knowledgeable about past and current research in their area of expertise and have their preferred ways of remaining up-to-date with the literature in their field. They may have subscribed to automated content alerts and other modern web services, read open access literature online, visit libraries and/or have personal subscriptions to some of the specialist literature relevant to their domain of research. In addition they may be used to keeping an eye on methodology-oriented papers in epidemiological journals. This situation may be very different for students faced with literature review and critical appraisal assignments and with dissertation requirements.

4.7.2.1 Strategies for Assessing Existing Summarized Evidence

When trying to find out more about existing evidence on a research question, one cannot trust brief summaries of evidence commonly found in introduction and discussion sections of papers that have addressed the topic or a very similar topic. For example, Fergusson et al. (2005) describe an instance of how inadequate citing of previous trials by investigators has led to an excess of unnecessary trials of a specific product. What are generally needed are recent systematic literature reviews as well as sources of expert opinion, such as narrative literature reviews, editorials, and commentaries, though these types of publications cannot substitute for reading the most relevant original research on a topic.

In some instances there are systematic literature reviews and expert opinion pieces available on a topic. In many other instances they are not available at all, or only on a tangentially related topic. In these latter cases there is a need for the epidemiological researcher to personally identify, assess, and summarize all relevant studies in a systematic literature review. In the former case there may be a need to update or improve existing literature review(s), depending on the results of a critical evaluation of the existing review(s) and opinion articles and a search for recent evidence.

4.7.2.2 Appraising Literature Reviews and Expert Opinion Articles

Critically reading recent expert reviews and opinion papers has become a key skill to gain insight into existing evidence. The scientific spirit demands this critical

Panel 4.11 Some Key Questions When Evaluating the Quality of Review Articles

- Is the research question specific enough?
- How systematic is the review? Is it a 'Cochrane type' review? How old is it? Was the search strategy comprehensive?
- Was the quality of the selected papers assessed systematically? If yes, how? How were strengths and limitations of papers taken into account in the overall summary of evidence?
- Does the review give due attention to sources of heterogeneity in study results in addition to attention to central tendency in the findings?
- Was there any evidence of publication bias? How was this issue examined?

approach because 'authority and fame', often perceived as signs of high expertise, on themselves do not provide for meaningfully summarized evidence. It would be a mistake to think that all systematic reviews are conducted by true experts in the field or that all experts meticulously apply guidelines of systematic literature review.

When critically reading reviews, one must take into account that reviews can be outdated. Depending on how frequently the topic is researched, the 'deadline' for considering a review outdated may be as short as a few months. A simple electronic search may give an indication as to recent papers. If several reviews exist, checking whether there is overlap of cited papers from defined periods may reveal that they were all incomplete. Another possible problem is that reviews may not be systematic enough. In nearly all cases the evidence presented by reviewers tends to be biased to an unknown degree by publication bias. Finally, the evidence presented, even if unbiased and about the broad topic of interest, may be partly irrelevant to the currently considered project, for example because it does not give enough detail about how the determinant – outcome relationships depend on modifiers. Some important questions in the review of reviews are listed in Panel 4.11.

Some organizations have specialized in setting up databases of systematic reviews. Pioneering work on systematic reviews was done by the Cochrane Collaboration (http://www.cochrane.org) and the United States Preventive Services Task Force (http://www.ahrq.gov/clinic/uspstfix.htm). The Campbell Collaboration focuses on social and educational policies and interventions (http://www.campbell-collaboration.org/)

4.7.2.3 First a Literature Review?

The researcher planning a new study will have to decide whether some form of new or updated systematic literature review is needed for the planned study. The spectrum of existing types of literature review is listed in Panel 4.12.

Narrative reviews are inherently more subjective than systematic reviews and may not be very reproducible in their approach. They are not without importance as they tend to describe valuable insights of experts, even if they are usually backed up by a more or less ad hoc selection of referenced materials. An important difference

Panel 4.12 Types of Literature Reviews

- Narrative
- Semi-systematic
- Systematic
- Systematic with meta-analysis

between the narrative and (semi-)systematic literature reviews is that the latter has a detailed methods section describing search strategies, quality assessment methods, and methods of synthesizing the evidence of the selected papers. Semi-systematic reviews, often performed by students, do use such a detailed methods but the search strategy is not as comprehensive as in a real systematic review (a task that often involves a committee and hired staff). Methods of systematic review will be further discussed in Chap. 25, which also deals with meta-analysis.

4.7.3 Is a New Study Warranted?

Within the context of research groups focusing on specific domains of medicine, the need for a specific new study is often a simple conclusion reached by a previous piece of research carried out. Even when that is the case, it is good to do a new check of evidence available in the literature before engaging with the new plans. Whenever a topic is relatively new to the student, investigator, or group, preparatory literature review is even more essential. However, identifying gaps in knowledge, usually through critical reading of reviews or doing or updating reviews, is only one of the considerations in the decision to embark on a new study. There are many additional questions and, ultimately, the opinions of stakeholders (especially the sponsors), may be decisive, as may be the opinion of the ethics committee.

Important questions include whether existing datasets can be used to answer the proposed research question and whether there is any ongoing research on the same topic. Epidemiology has yet to design a comprehensive and user-friendly system of identifying existing publicly available research databases and whether the available ones are fit for a particular new research question. For information on ongoing clinical trials one can consult registries of trials or consult research sponsors. Most often, however, the only way to find out if similar initiatives are under way is to remain up-to-date in the particular field of research, e.g., through conference attendance.

Efficiency questions may arise as to whether it will be possible to piggyback the new study as an add-on to an ongoing cohort study, or upon any planned and possibly already funded study. One should consider any adverse effects resulting from the supplementary and secondary status of the prospective new project component. Another concern is quality of any data that will be borrowed from the host study and the effect of the ancillary study on the quality of the host study. The necessary data for answering the proposed research question may also be available from registries or non-research datasets with similar concerns about validity and completeness.

Ultimately, it may appear that an independent study with new data collection is desirable, especially if there seems to be sufficient potential access to observation units, excellent measurement tools, and if (in analytical studies) all potential confounders can be identified and measured reliably. Studies that are too small may fail to detect important effects or produce estimates that are too imprecise to be useful. No health authority is interested in or will immediately act upon statements such as "the prevalence of the disease in the area is 10 % (95 %CI: 1–19 %)". Misinterpretation of results of small studies frequently happens and may do more harm than good. On the other hand, results of small scientific studies, if well designed, may contribute to later meta-analyses. In the short term, however, sponsors and other stakeholders have outspoken preferences for studies that are expected to produce strong high-precision evidence.

4.7.3.1 Stakeholder Opinions on Whether a Research Question Should be Pursued

In the present era research sponsors are becoming the main decision makers about what research questions will be addressed. Sponsors often advertise their preferred research areas or even very specific research questions they are interested in. Research institutions like to ensure that research questions addressed within the institute fit well within the larger research programs and strategies and that they have great potential for attracting external funding. Finally, patients, health authorities, hospital management and communities may have their opinion on how useful and acceptable a planned study is. Health authorities may also define research priority areas. As a basis for interaction with the sponsors and other stakeholders it is advisable to write a pre-proposal.

4.8 Developing a Pre-proposal

Pre-proposals usually are no longer than three to five pages. Key content includes:

- An informative title
- A summary of relevant evidence in the literature
- Aims and objectives accompanied by a rationale for why they are relevant, feasible, and potentially important
- Brief description of methods, including study size
- List of key papers
- Timeline and preliminary budget estimate

The pre-proposal must be refined and improved through discussions with scientific collaborators and stakeholders. If all indicate interest and potential support, a more comprehensive proposal must be developed. The necessary elements for inclusion into a full detailed proposal are discussed in detail in Chaps. 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14. Full proposals form the basis of development of ethics proposals, grant proposals and, eventually, the final and official study protocol. Each of these will have to comply with the specific requirements of the institutions or committees concerned.

In this chapter we discussed broad study objectives, presented a classification of research topics, and showed that this classification system is applicable to both clinical and community medicine. In the next chapter we introduce concepts and terms used to pinpoint the more specific aims of research studies.

References

Fergusson D et al (2005) Randomized controlled trials of aprotinin in cardiac surgery: could clinical equipoise have stopped the bleeding? Clin Trials 2:218–232

Miettinen OS (1985) Theoretical epidemiology. Delmar, New York, pp 1–359. ISBN 0827343132

Miettinen OS (2001a) The modern scientific physician: 3. Scientific diagnosis. CMAJ 165:781–782

Miettinen OS (2001b) The modern scientific physician: 4. The useful property of a diagnostic. CMAJ 165:910–911

Miettinen OS (2002) Feinstein and study design. J Clin Epidemiol 55:1167–1172

Miettinen OS (2008) Screening for a cancer: a sad chapter in today's epidemiology. Eur J Epidemiol 23:647–653

Miettinen OS (2010) Etiologic study vis-à-vis intervention study. Eur J Epidemiol 25:671–675

Miettinen OS (2011) Up from clinical epidemiology & EBM. Springer, Dordrecht, pp 1–175. ISBN 9789048195008

Miettinen OS, Flegel KM (2003) Elementary concepts of medicine: XI. Illness in a community: morbidity, epidemiology. J Eval Clin Pract 9:345–348

The Campbell Collaboration (2013) http://www.campbellcollaboration.org/. Accessed Feb 2013

The Cochrane Collaboration (2013) http://www.cochrane.org. Accessed Feb 2013

Unites States Preventive Services task Force (2013) http://www.ahrq.gov/clinic/uspstfix.htm. Accessed Feb 2013

Van den Broeck J et al (1993) Child morbidity patterns in two tropical seasons and associated mortality rates. Int J Epidemiol 22:1104–1110

Viergever RF et al (2010) A checklist for health research priority setting: nine common themes of good practice. Health Res Policy Syst 8:36

Ware JH (2006) The limitations of risk factors as prognostic tools. N Engl J Med 355:2615–2617

World Health Organization (2010) International classification of diseases (ICD-10). http://www.who.int/classifications/icd/en/index.html. Accessed Sept 2012

The Specific Aims

5

Jan Van den Broeck, Jonathan R. Brestoff,
and Meera Chhagan

Judge a man by his questions rather than by his answers.

Voltaire

Abstract

When proposing a study, one first briefly formulates the 'general study objectives' and then describes the 'specific aims' to clearly articulate the essence of the design used to generate empirical evidence about the research question(s) at hand. This is a crucial step in the development of the research plan. Indeed, reviewers of study proposals often consider the 'specific aims section' as the most important section of the proposal, as this section provides them a first insight into the validity and efficiency of the design and methods to be used. This chapter explains that the essence of a study design lies in specifications of the study domain, occurrence relation(s), study base, study variables, and outcome parameters. This chapter also offers practical advice for investigators in pinpointing and describing the specific aims of a research project.

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

M. Chhagan, Ph.D., FCPaed
Department of Paediatrics, University of KwaZulu-Natal,
Durban, South Africa
e-mail: Chhagan@ukzn.ac.za

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_5,

5.1 What Are Specific Aims?

The specific aims describe the essence of the design of a study by briefly describing for each research question to be addressed:

- The *study domain* (the type of persons and situations about which the evidence obtained in the study will be applicable, e.g., children with type 1 diabetes)
- The *occurrence relation* (the phenomena that will be studied and related, e.g., age as a determinant of body adiposity)
- The *study base* (the cohort, dynamic population, or population cross-section that will be used, e.g., a cross-section of type 1 diabetic children from a national patient registry in 2006–2010)
- The *study variables* (the statistical variates that will express the attributes/experiences of the study base representatives, e.g., a body mass index variable representing the level of adiposity)
- The *outcome parameter* (the statistic that will summarize the empirical evidence about the occurrence relation, e.g., a t-test statistic comparing mean body mass index between boys and girls)

These essential elements of the study design should be briefly presented in the specific aims section of study proposals (one will provide more in-depth treatments of each element in other sections of the proposal) and in official protocols. Here, we will expand on each of these elements (using Panel 5.1 terminology) and conclude the chapter with an example of a specific aims section.

Panel 5.1 Selected Terms Relevant to the Formulation of Specific Aims of Epidemiological Studies

Cohort A fixed group of subjects composed on the basis of a once-off selection criterion and followed to study the frequency of occurrence of the outcome

Confounder A third factor that distorts the observed association between exposure and outcome (away from the true independent effect)

Confounding variable Variable representing a confounder in a statistical model

Determinant Factor related (causally or acausally) to the outcome

Determinant variable Variable representing a determinant in a statistical model

Dynamic population A group of subjects with varying composition over calendar time because membership, based on a chosen criterion, only lasts for as long as the criterion is fulfilled

Effect modifier A factor by whose level the relation between exposure and outcome changes

Exposure Determinant; factor related (causally or acausally) to the outcome

Exposure variable Variable representing an exposure in a statistical model

(continued)

Panel 5.1 (continued)

Occurrence relation The object of study: the proposed relation among outcome, exposures (and sometimes confounders and effect modifiers)

Outcome The phenomenon of which the frequency of occurrence is studied

Outcome parameter Type of statistic used to summarize the evidence about the occurrence relation (e.g., a prevalence or an incidence rate ratio or a P value)

Outcome variable Variable representing the outcome in a statistical model

Population cross-section A 'snapshot' of a cohort at a particular follow-up time or of a dynamic population at a particular calendar time

Primary analysis Analysis carried out to produce evidence about the most important specific aim

Study base The real-life experience of members of a cohort, dynamic population or population cross-section that will be documented to provide empirical evidence about the occurrence relation

Study domain The type of persons and situations about which the evidence obtained in the study will be applicable

Study variable A variable representing an outcome, exposure, effect modifier, confounder, or mediator

5.2 The Study Domain

The study domain is the type of persons or situations about which the empirical evidence will be applicable. This concept is roughly equivalent to the concept of 'target population' (*See:* Chaps. 1 and 2). The latter concept tends to be used only when the observation units are individuals.

A study domain is usually well-characterized by three main elements (Panel 5.2). Firstly, one needs to specify whether the *observation units* are individuals or groups (e.g., children). Secondly, one must specify whether *time-space restrictions* apply (e.g., children residing in Zululand in 2010). The choice to include space and calendar-time restrictions in the description of the study domain implies that there is no ambition to generalize beyond that particular chosen place and period. In contrast, a choice for a type of individuals or group without space or time restriction implies that, through the study, one expects to make scientific inferences about this type of individuals or group *in general*. Thirdly, *other domain restrictions* may apply. If the study concerns the course of an illness, then it is natural to limit the study domain to subjects with that illness (e.g., children in Zululand in 2010 with type 1 diabetes).

Panel 5.2 Elements in the Description of the Study Domain

1. **Type of observation unit**: individual or group; type of group in the latter case (e.g., schools, villages, …)
2. **Time and space restrictions**: none for a scientific study; one or both for a particularistic study
3. **Other restrictions**:
   - Restrictions relating to an illness or morbidity profile
   - Restrictions regarding age, sex, or other characteristics that are not direct constituent parts of an illness or morbidity profile

To adhere to the study domain (and to convince reviewers that one will adhere to the study domain), the investigator requires strict definitions for all elements in the description of the study domain. For illnesses, case definitions are mostly based on clinical characteristics, laboratory values, scoring systems, and/or statistical diagnostic cut-offs. Case definitions may be simple or complex and may depend on accepted international classification systems (International Classification of Diseases-10). One sometimes prefers to study existing prevalent cases of an illness. For example, consider a study among prevalent cases of hypertension. There may be untreated as well as treated cases of hypertension, and some of the treated individuals might be non-responsive to their medications or other therapies. The study domain may include all or some of those types, and the description of the study domain must be clear about this.

It is an ethical requirement to include in a study only subjects whose potential data will be informative about the research question. This means that one will need to exclude non-informative observation units from the study domain. It is also an ethical requirement to exclude persons with contra-indications for particular study interventions. Further restriction of the study domain may be needed to exclude rare categories of confounders or effect modifiers (*See:* Chap. 9).

Note that the description of the study domain will be a basis for making the list of inclusion- exclusion criteria for the enrollment phase of the study (*See:* Chap. 9).

5.3 The Occurrence Relation

The concept of occurrence relation is a basic concept in epidemiology (*See:* Chap. 2). The basic elements of an occurrence relation are:

- The outcome (always)
- Determinants (sometimes)
- Effect modifiers (sometimes)
- Confounders (sometimes)

More complex occurrence relations can be of interest in observational-etiognostic research, where causal webs may further include instrumental variables and mediators among other factors. In this section we only discuss the listed basic elements.

5.3.1 Outcomes and Determinants/Exposures

There can be several research questions in one study and, correspondingly, several specific aims and occurrence relations. Each specific aim concerns the occurrence of a health-related state or event, or 'the outcome' (e.g., *level of adiposity*), usually within the same study domain. The outcome is often studied in relation to one or several determinants (e.g., *age*). The concept 'determinant' is used in a broad sense of 'a factor related to the outcome', without any connotation to whether this relation may be causal (causative or preventive) or non-causal. The term can thus be used in the context of research on possible causal effects but also in purely descriptive research aimed only at demonstrating associations. An alternative term for determinant, equally popular in epidemiology, is 'exposure.' A distinction is made between past exposure episodes and current exposure states. When the temporal relationship between two phenomena is considered, the one that occurs after the other is to be termed the outcome, and the other is said to be the determinant/exposure. This distinction is an extension of the basic temporality criterion discussed in Chap. 2. Only cross-sectional state relationships and relationships of outcomes with past exposures are allowable in epidemiology.

5.3.2 Effect Modifiers and Confounders

Sometimes the interest is also in how the determinant-outcome relationship changes by levels of other attributes. An effect modifier is an attribute that influences the (degree of a) relationship between a determinant and an outcome (e.g., sex may be a modifier of the relation between age and adiposity) (*See also:* Chap. 2). Here again, 'effect' and 'effect modifier' are terms that can be used in a broad sense, without connotations to the possible causal or non-causal nature of relationships.

Only when there is an explicit interest in possible *causal effects,* will potential or known confounders become elements of the occurrence relation. As pointed out in Chap. 2, a confounder is an extraneous factor that distorts the estimated causal effect of a determinant on an outcome. In studies of possible causal effects the occurrence relations can involve several confounders and their interrelationships. Complex occurrence relations can nowadays often be formally specified and analyzed through graphical theory and structural causal modeling (Pearl 2010). In such instances the description of the occurrence relation may usefully include a causal graph (Greenland et al. 1999).

5.3.3 Clarifying the Attributes

There is a need for clear definitions of all attributes that will be part of the occurrence relation. Height, for example, could be defined as 'the linear dimension of a person standing maximally erect and looking straight forward, from the soles to top of the head.' Not all attributes can have such specific definitions, however.

An example of a less clearly defined construct (a 'latent construct') is intelligence. We don't know exactly what intelligence is but we think that we can measure some manifestations of it. When specifying an occurrence relation there should always be a preference (to the extent possible) for attributes with clear definitions that can be measured using validated measurement tools with acceptable reproducibility.

The exact nature of an attribute will often be intuitively clear (e.g., *height*), and in such cases the definition does not need to be described in the specific aims, perhaps only in later sections of the study proposal. But if there are several competing and rather different definitions of the same attribute (e.g., *social class*), clarifying the attribute in the specific aims may be useful. Attributes can also have a composite nature that needs clarification. For example, attributes often used in experimental research are 'treatment failure' or 'treatment success,' classifications that are entirely dependent on the measurement of other attributes and often-subjective definitions of what constitutes success or failure. Such composite attributes may need to be explained briefly in the specific aims section. Finally, it may be necessary to specify whether the attributes in the occurrence relation are intrinsically continuous (e.g., percent body fat) or have some other scale property (e.g., body mass index between 30.0 and 34.9 kg/m^2), although this issue will be often clear enough without specific mentioning.

5.3.4 Clarifying the Relationships of Interest

Descriptions of specific aims do not just name or graphically depict the phenomena/attributes that constitute the outcome, determinants, effect modifiers, and confounders. For outcomes and determinants, one must specify whether the interest is in the mere existence of a relation between these phenomena/attributes or in any particular shape or strength of a relation (e.g., the interest may be only in the *existence of a difference in adiposity between boys and girls*). With respect to an effect modifier one should be clear about whether it is seen as a factor to control for (perhaps by standardization) or whether there is a specific interest in the strength or shape of the determinant-outcome relation at each or a few levels of the effect modifier.

When describing the occurrence relation one needs to pay attention to the fact that the determinants, confounders, and effect modifiers can be nested. For example, a specific aspect of behavior can be part of larger type of behavior or lifestyle; a specific exposure to a toxic substance may be part of a wider range of undesired exposures in a workplace context; and, a specific bodily dysfunction can be part of a set of related dysfunctions. This potentially nested status of attributes has important consequences when conceiving to adjust a determinant-outcome relation for another attribute (an effect modifier or a confounder). This is illustrated in Textbox 5.1.

Textbox 5.1 Three Scenarios Illustrating the Effect of *Nesting* When Adjusting a Determinant-Outcome Relation for Another Factor

Factor Є Determinant

The adjustment factor is nested (Є) within the determinant. Say, the determinant under investigation is the general level of pollution at an occupational setting. The contemplated covariate to adjust for is exposure to a specific toxic substance. This strategy would make the estimated determinant-outcome relation independent of the specific toxic substance and would thus investigate an association with the entirety of all other remaining exposures.

Determinant Є Factor

The determinant is nested within the adjustment factor. This circumstance should generally be avoided. For example, the determinant may be ownership of a car, and the adjustment factor may be general socio-economic status. This leads to situations where the 'remaining' association after adjustment is difficult to define.

Factor ~ Determinant

The adjustment factor is another determinant. For instance, alcohol consumption and tobacco smoking often go together. Controlling for this factor will make the estimated determinant-outcome relation independent of the adjustment factor. One may have difficulty ascertaining the independent effect of one factor without valid measurement and control for the other.

5.4 The Study Base

The study base is a sample's collective real-life experiences that will need to be empirically measured and related to address the research question. Note that, when the study is of a particularistic nature, as in a survey, the study domain can also be the study base. In experimental and quasi-experimental studies, the experience of the study base is manipulated for study purposes (by an intervention). For reviewers of study proposals it is difficult to acquire a clear idea of the general study design without being informed about the study base, the direct source of empirical evidence. Thus, it is helpful to mention the study base in the specific aims section.

Specification of the study base also requires a stipulated duration and calendar-time of this real-life experience. With respect to the calendar timing of the study base, three basic types are possible:

- *Retrospective study base:* The study base experience has already happened, i.e., before the currently conceived study will be in the data collection phase
- *Prospective study base:* The study base experience will happen after enrollment

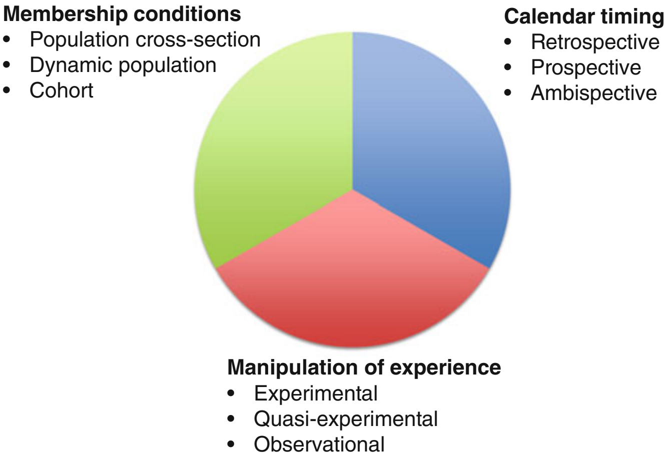


Fig. 5.1 The 3×3 study base wheel. The study base is the sample's collective real-life experiences that will need to be measured. The study base may be defined by three main categories: membership conditions, calendar timing, and manipulation of experience. Within each are three main alternatives. Only one alternative per category may be chosen when defining the study base

- *Ambispective study base:* The study base experience has partly happened already but will partly happen after enrollment into the currently conceived study (Kleinbaum et al. 1982)

This leads to a 'three times three' characterization of the study base, as illustrated in the study base wheel (Fig. 5.1).

5.4.1 Membership Conditions

5.4.1.1 Cohorts

A cohort is a closed population, i.e., a population with fixed membership. Its meaning derives from its use during the ancient Roman Empire, in which a cohort was a subdivision of an ancient Roman legion. Soldiers were enrolled into the cohort as fast as possible and forever. Numbers alive decreased over time. Membership of a cohort is defined by a one-time criterion and membership duration is eternal (though an individual can be lost to follow-up). For example, when someone becomes a member of the 2010 birth cohort in Norway, one was always born in 2010 in Norway, irrespective of time of death or emigration to another country. An illustrious historical example of use of a cohort in epidemiology is the Framingham study, in which a cohort of adults 30- to 62-years-old living in Framingham in 1948, were followed for 20 years to study coronary heart disease (Dawber et al. 1957).

Cohorts are used as a study base in many different study designs, in the whole range from experimental to quasi-experimental to observational studies. In an *experimental cohort study (a trial)*, one investigates the effects of a test intervention in a cohort. In *quasi-experimental cohort study*, one investigates the effects of a researcher-allocated but non-randomized intervention in a cohort. *Observational cohort studies* do not involve any experimental or quasi-experimental allocation of interventions to a cohort, though that does not mean that subjects in observational cohort studies cannot undergo intervention. They can, but not as a manipulated component of the research design. Such observational cohort studies can have diagnostic, etiognostic, or descriptive-prognostic aims.

Another special type is the *test-retest study*, a method-oriented type of study, in which subjects are re-measured after a very short follow-up interval during which no measurable change in the measured attribute is expected. Any observed change in values is therefore due to instrument problems or 'observer error,' though such studies must be careful to control for time-of-day effects (e.g., circadian rhythms) and many other factors. Test-retest studies using quality instruments and appropriate design elements can therefore be done to document observer performance.

Given the fact that cohorts are used as a study base in several very different types of studies, the common use of the term 'cohort study' can be confusing. It would be better to characterize studies by making reference to what really distinguishes between them. Yet, 'cohort study' has now become the standard term to refer to a single particular study type, which is the traditional cohort-based observational etiologic study (*See:* Chap. 6), of which the Framingham study is an example.

5.4.1.2 Dynamic Populations

A dynamic population is an open population, with turnover of membership. The term is borrowed from demography, where a population is not seen as fixed but there are *ins* (births and immigrations) and *outs* (deaths and emigrations). Membership of a dynamic population is defined by a state (Miettinen 1985), for example the state of living in a particular town in a particular year, e.g., Durban in 2010. Membership duration is for the duration of that state. For instance, someone who lived in Durban only in January 2010 was a member for 1 month. As another example, to study coronary heart disease-related mortality over 5 years in a village, rather than using a cohort of all subjects older than 38 years living in the village (as in the Framingham study), one could instead be interested in all subjects older than 38 years that will ever live in the village in a period of 5 years and follow them for the time (within those 5 years) that they are present in the village. In the latter case, the study uses a dynamic population instead of a cohort. Dynamic populations are also used in a range of studies, both descriptive and analytical. For example, they are commonly used:

- As a primary study base in an etiognostic study, as in the example described above (the alternative to the Framingham study design)
- As a secondary study base in an etiognostic study
- In descriptive population surveillance studies

On this basis we propose that the expression 'dynamic population study' should not be used as if it indicated any particular type of general study design (as has been the case for 'cohort study').

5.4.1.3 Population Cross-Sections

A population cross-section is either a cohort at a single follow-up time (usually follow-up time zero; e.g., baseline characteristics of a cohort) or a dynamic population at a fixed point in calendar time (e.g., a survey). It follows that a population cross-section is *not* necessarily a group of people all present at one moment in time. In health research, not everybody can be examined at the same time. At best, a group of people can be selected and examined once within the shortest possible time. For example, a cross-sectional study of presenting symptoms at diagnosis of a rare disease could take 20 years to complete and not all participants may even be alive at the same time. Typical characteristics of a cross-sectional study are that the attributes and experiences of interest are/were assessed once without individual follow-up and that all units of observation were assessed within the shortest possible time.

Similar to cohorts and dynamic populations, population cross-sections are commonly used as the study base in a variety of study designs (*See:* Chap. 6) and therefore 'cross-sectional study' cannot be used to indicate any particular type of study design. Also, as indicated, a population cross-section still concerns either a cohort or a dynamic population (whether or not explicitly defined).

5.4.2 Variation in and Restrictions to the Study Base

There is a general principle that for a study to be informative about a determinant-outcome relationship there should be variation of the determinant in the study base, and, for a study to be efficient, that variation should be wide (Miettinen 1985). For example, if all participants get the same dose, the effect of dose cannot be studied; or if all participants are females, the role of gender as a determinant of the outcome cannot be assessed. Thus, in experimental research there is an interest in highly contrasting two- or three-point designs (i.e., two or three intervention arms differing by dose prescribed), whereas in observational research there is a general interest in choosing a study base with wide variation of the determinant.

When proposing a general design, there is often no objection to being selective about determinant levels. For instance, in an etiognostic study with a cohort as the primary study base and with the only aim to demonstrate the existence of an effect, there may be no objection to limit the cohort to subjects belonging to the non-exposed and highly exposed, leaving out the intermediately exposed. This principle is well recognized in occupational and environmental epidemiology, where it is an aim to have strong representation of the extreme exposure zones in etiognostic research (Corn and Esmen 1979). This is a strategy that can also help to reduce the total sample size required. Thus, when an appropriate study base is identified, this does not necessarily mean that all persons whose experience constitutes the study base must be potential study participants. It may be more efficient to take a representative sample.

5.5 Study Variables

There are three main types of study variables that represent the basic elements of occurrence relations in statistical analyses:

- Outcome variables
- Determinant variables
- Covariates

The term ‘covariate’ is used to denote any variable that would need to be ‘controlled for’ in the analysis when studying the relation between determinant variable and outcome variable. This is used in a broad sense without any connotation regarding whether such a covariate is seen as a potential confounder (in analytical research) or a factor from which the determinant-outcome relation needs to be independent (in descriptive research).

5.5.1 General Requirements for Study Variables

A general requirement for study variables is that the measurement values must be highly correlated with the underlying attribute and come close to measuring the true dimension on average (i.e., high intrinsic validity). For reviewers of study proposals, this tends to be very important information and it is advantageous if the specific aims section gives already a good indication of the intrinsic validity of key variables. For example, it is good practice to avoid using proxy variables to the extent possible. A proxy variable is a variable that does not directly reflect the attribute of interest but is assumed to correlate well enough with it to represent it in an analysis. However, the highest possible intrinsic validity is not always required or affordable. The particular study aims determine the required intrinsic validity of measures, so this issue must be considered on a case-by-case basis. For example, consider an occupational health study and a pharmacological study, both looking at the effects of exposure to a particular chemical substance on a particular health outcome. In the occupational health study it may suffice to measure environmental exposure levels as a proxy for true individual exposure levels, whereas in the pharmacological study it may be required to assess blood/tissue levels of the chemical in each individual.

Comment

A general ethical consideration for all study variables is that they must be based as much as possible on **non-invasive** measurements if human subjects are the units of observation. Invasive procedures are those involving direct entry into living tissues or the exertion of potentially painful and damaging mechanical or physical forces on living tissues. An ‘invasive question’ is a sensitive question. The ‘sensitivity’ may be related to stigma associated with the condition under study or to perceived inappropriateness of the interview questions, e.g., not being culturally acceptable.

Panel 5.3 Types of Variables According to Measurement Level

- A **nominal variable** is defined as a variable measured on a nominal scale, i.e., on a measurement scale consisting of a number of mutually exclusive categories that have no meaningful order. Examples are sex and ethnic group
- An **ordinal variable** is measured on an ordinal scale, i.e., on a measurement scale consisting of a fixed number of mutually exclusive categories in which there is a meaningful order but the differences between categories do not reflect meaningful differences in the 'amount' of attribute. An example is letter grades on a test
- A **discrete numerical variable** is measured on a discrete numerical scale, i.e., on a measurement scale for non-continuous underlying characteristics, consisting of a finite and ordered number of numerical values, with the differences between values having a meaning. Examples are parity and gravidity
- A **continuous variable** is a numerical variable measured on a continuous numerical measurement scale, i.e., on a scale for measuring continuous underlying attributes, expressing measurement values as multiples (with any number of decimals) of a measurement unit. This comprises the interval and ratio measurement scales. Only the ratio scale has a true zero point as the lowest possible value corresponding to the lowest possible amount of attribute. In practice there is not much advantage of a ratio scale over an interval scale except that ratios of measurement values have a more straightforward constant interpretation when a ratio scale is used

Further, one should aim for the highest possible measurement level whenever it is feasible from a budgetary and ethical perspective. Measurement levels are ranked from lower to higher as follows: *nominal* < *ordinal* < *numerical discrete* < *numerical continuous*. Their distinguishing characteristics are described briefly in Panel 5.3. A common advantage of using higher measurement levels is higher statistical efficiency and a wider range of possible statistical analyses. But higher-level measurements tend to be more expensive and sometimes also more invasive. When the underlying intrinsic scale property is nominal (e.g., sex), the variable and measurement scale can only be nominal, too. When the intrinsic scale property is higher, such as continuous (e.g., age), there may be a choice for the variable and its measurement scale between, say, ordinal (young or old) and continuous (age measured as calendar time elapsed since birth). In such instances, the preference generally goes to higher measurement levels.

5.5.2 Variables Expressing Latent Constructs

Sometimes a researcher cannot measure the attribute accurately with a single question or other type of measurement. Instead, (s)he can only think of a series of questions (or other measurements) that each measure some component of the attribute and, if somehow the answers to all of these questions could be taken together, a reasonably accurate measurement could be obtained. Common examples include quality of life (QOL; *See:* Chap. 10), socioeconomic status, and diagnostic questionnaires for psychiatric conditions. In such situations it might be preferable to develop a new measurement tool (Howitt and Cramer 2008; Streiner and Norman 2008), or adapt an existing tool for local circumstances. The term ‘scaling’ refers to such creation of a new tool, often based on a *series of questions*, for the measurement of the latent attribute.

As pointed out in a previous section, every effort should be made to specify the nature of the attributes we wish to measure. When reflecting on this issue in the context of latent attributes, it may appear that there are several aspects to the latent attribute that may need to be measured on a *subscale*. The need for subscales can also be identified by a statistical technique called factor analysis (*See*: Chap. 10). For example, nutritional health-friendliness of schools may be viewed as multidimensional attribute composed of:

- Nutritional care at school
- Provisions for physical activity at school
- Nutritional health education at school
- Other aspects

Different series of questions may then be needed to measure *sub-scores* on the corresponding subscale. In other instances it may seem reasonable to measure the latent attribute on a single scale, using a single series of questions (unidimensional scale). In this case all items should correlate about equally well with the total score. This can be verified using a statistical exercise called item analysis.

For more guidance on developing a new measurement scale, *See:* Chap. 10, and Streiner and Norman (2008).

5.5.3 Outcome Variables

The one study variable that is always necessary is the outcome variable. Outcome variables express a (change in) health-related state or event for each observed individual or other observation unit. When group attributes rather than individuals’ attributes or experiences are the outcomes of interest in a study, that study will often be labeled an ‘ecological study’. The outcome variables of such studies are ‘ecological variables,’ of which there are three types according to Morgenstern (1998), as shown in Panel 5.4. We propose that, similar to ‘dynamic population studies’ and ‘cross-sectional studies,’ the term ‘ecological study’ should not be used as if it

Panel 5.4 Types of Ecological Variables[#]

- **Summary environmental measures:** summarizing for the whole group an exposure that actually *varies considerably at the individual level*, e.g., global level of air pollution in a workplace
- **General environmental features:** exposure that is *identical for each individual* in the group, e.g., existence of a specific law or policy in the area
- **Statistical estimates:** summary statistics of variables that are based on single measurements, repeated measurements, or combinations of several variables; e.g., prevalence estimates of a disease. This type of ecological variable is often *based on individual-level measurements*. Note that ecological studies using this type of ecological outcome variables could also be called 'meta-analytical.'

[#] Panel adapted from Morgenstern (1998)

Panel 5.5 Types of Variables According to Number and Timing of Underlying Measurement Acts: Some Examples

- **Single measurements**
  - Single systolic or diastolic blood pressure reading
- **Combinations of measurements for single assessment**
  - Systolic blood pressure based on average of three replicates
  - Presence of hypertension based on diastolic and systolic blood pressure
- **Repeated assessments**
  - (Baseline-adjusted) change in systolic blood pressure
  - New occurrence of hypertension

represents any particular type of general study design. 'Ecological study' should simply refer to the fact that the outcome variable of the study, whatever its design, is an ecological variable.

One of several ways to broadly classify outcome variables is according to number and timing of underlying measurement acts, as illustrated in Panel 5.5.

5.5.4 Determinant Variables and Covariates

Outcome variables frequently represent health-related constitutional or functional attributes or individual subjective experiences around them. Determinant variables and covariates tend to represent behavioral, environmental, or constitutional factors.

The reasons for this have been explained in Chap. 4: A complex and ever-changing interaction of these three types of factors is what creates each individual's personal life experience. A researcher often uses 'summaries of episodes' of that interaction as determinants. Examples are cumulative doses of exposure over time, and broadly described exposure situations or types of exposure histories, e.g., 'was a manual labor worker (yes/no)'.

Determinant variables and covariates cannot represent experiences or states that temporally follow the outcome. Temporality issues with covariates are important. Time-dependent and time-modified confounding are issues that have only recently started receiving attention (Platt et al. 2009). These phenomena are especially relevant to situations where the outcome variable is derived from time-series data.

5.6 Outcome Parameters

In epidemiology an outcome parameter is a statistic that summarizes the evidence in the data about the occurrence relation under study. Design of the outcome parameter is part and parcel of the general study design (Miettinen 2004). Typical examples of outcome parameters in epidemiology are prevalence and the odds ratio, either crude (unadjusted) or adjusted. The adjustments may be for undesired effects on the outcome parameter estimate such as by confounding, bias, and imprecision of measurement. Outcome parameters traditionally fall into two categories: *estimators* and *test statistics*. Estimators will be discussed in Chap. 22 (Statistical Estimation) and test statistics in Chap. 24 (Statistical Testing). The outcome parameter of a particular study could be a difference in prevalence, which is an estimator. But in the same study, an outcome parameter could also be a chi-square test statistic with P-value addressing the same occurrence relation. In many study reports, estimators and test statistics are reported alongside each other. Estimates have the advantage that they allow for more easy assessment of *magnitudes* of effects in addition to assessing the *existence* of effects.

Hint

The three terms *outcome*, *outcome variable*, and *outcome parameter* sound quite similar but have very different meanings. An 'outcome' is a health-related state or event that is under study (e.g., stroke). An 'outcome variable' is a statistical variate representing the observed values of the outcome in a statistical model or showing them in a database column (e.g., coding 'no stroke' as 0 and 'stroke' as 1). An 'outcome parameter' is a type of statistic that expresses the study 'result' (e.g., an odds ratio).

Estimators can capture a frequency of occurrence (e.g., a single prevalence, or a single incidence rate), in which case they are called 'measures of frequency.' They can also express a contrast of occurrences between two categories/groups (e.g., a difference between two prevalence estimates, or an incidence rate ratio), in which case we call them 'measures of association' or 'measures of causal effect' depending on whether the aim is descriptive or analytical. There are two main approaches to formally contrasting outcome rates among levels of a determinant: the approach using a risk/rate *ratio* and the now less-frequently used approach using a risk/rate *difference*. Miettinen (2004) has pointed out that logistic regression analysis can provide for valid outcome parameters of most types of occurrence relations in epidemiology.

5.7 Presenting the Specific Aims in a Study Proposal

Thus far we have discussed the elements that are typically required or useful to include in a specific aims section. The content and format of the specific aims section may depend on the expectations and guidelines imposed by the particular sponsor or ethics committee for which the document is intended. It is therefore not possible to provide a standardized example of how a specific aims section must be structured. However, systematic consideration of the points raised in this chapter leads to a logical template that is, at the very least, a helpful tool for formulating a specific aims section. Creating such a template for a particular study ensures that the most important information is included.

We propose that the specific aims section first indicates the general aims/objectives of the present study so that the link with the specific aims will become clear. If the study domain is common to all of the ensuing specific aims, it can be included as part of the purpose summary statement and/or as a separate line. In other instances the study domain may be different for some specific aims (perhaps a sub-domain of the study domain), in which case it is advisable to list the domain under the relevant specific aim. We then recommend listing one specific aim after the other. Some investigators prefer indicating a ranking of specific aims, with a primary aim, secondary aims, tertiary aims, etc. The primary aim is considered the main reason why the study is set up. Attempts to achieve an ‘optimal’ sample size are usually geared towards it. As an example, Textbox 5.2 shows a specific aims section of the study proposal in the domain of dentistry. It is an example of a hypothesis-generating descriptive diagnostic research project.

Textbox 5.2 Example of a Specific Aims Section of a Dentistry Study Proposal

Periodontal disease in childhood is associated with substantial morbidity and increases the likelihood of needing costly medical procedures. However, the prevalence of periodontal disease and its risk factors in primary schoolchildren in Cork, Ireland are unknown, making it difficult to plan for related healthcare costs and to intervene if necessary. We therefore propose the following specific aims:

(continued)

Textbox 5.2 (continued)

Study domain: Primary schoolchildren in Cork, Ireland in 2010
Specific aim 1:
To estimate the prevalence of periodontal disease (ICD10-defined) overall and in 1-year age categories in a representative population cross-section (N = 400)
Specific aim 2:

(a) To quantify the differences in prevalence rate of periodontal disease (ICD10-defined) according to degree of body adiposity, as represented by World Health Organization-defined body mass index (BMI)-for-age categories, by taking the category of BMI >18–25 Kg/m^2 as the reference category for the calculation of prevalence odds ratios for the other categories

(b) To examine, by stratified analysis, if the prevalence odds ratio for periodontal disease (ICD10-defined), for 'overweight or obese' relative to 'normal BMI' (as defined above), is modified by usual frequency of brushing teeth (times per week <7 or ≥7)

In this chapter we discussed specific aims and their elements. Patterns exist in the combinations of these elements, some patterns being more common than others because they serve the purposes of general study objectives. These patterns/combinations can be called 'general study designs,' the topic of Chap. 6.

References

Corn M, Esmen NA (1979) Workplace exposure zones for classification of employee exposures to physical and chemical agents. Am Ind Hyg Assoc J 40:47–57

Dawber TR et al (1957) Coronary heart disease in the Framingham study. Am J Public Health 47:4–24

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

Howitt D, Cramer D (2008) Introduction to research methods in psychology, 2nd edn. Prentice Hall, Harlow, pp 1–439. ISBN 9780132051637

Kleinbaum DG, Kupper LL, Morgenstern H (1982) Epidemiologic research. Van Nostrand Reinhold, New York, pp 1–529. ISBN 0534979505

Miettinen OS (1985) Theoretical epidemiology. Delmar, New York, pp 1–359. ISBN 0827343132

Miettinen OS (2004) Knowledge base of scientific gnosis: II. Gnostic occurrence relations: elements and temporal structures. J Eval Clin Pract 10:357–359

Morgenstern H (1998) Ecologic studies. In: Rothman KJ, Greenland S (eds) Modern epidemiology, 2nd edn. Lippincott Williams & Wilkins, Philadelphia, pp 459–480. ISBN 0316757802

Pearl J (2010) An introduction to causal inference. Int J Biostat 6(7):1–59

Platt RW et al (2009) Time-modified confounding. Am J Epidemiol 170:687–694

Streiner DL, Norman GR (2008) Health measurement scales. A practical guide to their development and use, 4th edn. Oxford University Press, Oxford, pp 1–431. ISBN 9780199231881

6 General Study Designs

Jan Van den Broeck, Jonathan R. Brestoff, and Meera Chhagan

Science is imagination in a straight jacket.

Based on J. Moffat

Abstract

In the previous chapter we explained that the necessary elements of general study design are the study domain, the occurrence relation, the study base, the study variables, and the outcome parameters. Different combinations of these elements tend to take different recognizable forms (‘jackets,’ mostly simply referred to as ‘designs’) depending mainly on type of research question. These designs are known under specific names, e.g., survey, forecasting study, randomized controlled trial, etc. For each we discuss here how the ‘jacket’ is tailored with design elements from Chap. 5 (we therefore advise careful study of the two preceding chapters before embarking on this one). In, this chapter we (1) aim to help researchers find the best general study design for a particular research question, and (2) provide a broad classification of general design types that parallels the typology of research questions proposed in Chap. 4.

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

M. Chhagan, Ph.D., FCPaed
Department of Paediatrics, University of KwaZulu-Natal,
Durban, South Africa
e-mail: Chhagan@ukzn.ac.za

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_6,

6.1 Classification of General Study Designs

A classification of mainstream study designs was introduced in Chap. 1 (*See:* Table 1.4). That typology of general study designs described experimental, quasi-experimental, and observational studies. The classification proposed in this chapter – summarized in Table 6.1 – differs from the mainstream classification in that (1) it makes a clear link with a typology of research questions (*See:* Chap. 4); (2) it includes some recent advances in study design, and (3) it includes some designs that are not consistently mentioned within the scope of epidemiology. Table 6.1 contains a series of design labels that fall within each design class. Each design label has been given a code, and these codes will be referenced throughout the chapter to orient readers.

These designs are explained and justified in the next sections by specifying their design elements. Possible design elements have been discussed in Chap. 5. To facilitate the present chapter, Panel 6.1 explains some of the important terms and concepts as they are used here, though additional terms and concepts listed in Panels 2.2, 4.1, and 5.1 are also directly relevant. These four panels may serve as useful resources when reading Chap. 6.

Table 6.1 Classification of general study designs

Designs used for:	Code	Design label
Diagnostic research	1.a	Case reports
	1.b	Case series studies
	1.c	Diagnostic probability studies
	1.d	Traditional diagnostic performance studies
	1.e	Surveys
	1.f	Epidemic pattern studies
	1.g	Cost studies of illness and intervention
	1.h	Meta-analytical diagnostic projects
Etiognostic research	2.a	Traditional etiologic studies (cohort, case-control)
	2.b	The single etiologic study
	2.c	Before – after etiognostic studies
	2.d	Meta-analytical etiognostic projects
Intervention-prognostic research	3.a	Randomized controlled trials
	3.b	Quasi-experimental trials
	3.c	Cross-over trials
	3.d	N-of-1 trials
	3.e	Meta-analytical intervention-prognostic projects
Descriptive-prognostic research	4.a	Clinical prediction studies
	4.b	Forecasting studies
Methods-oriented research	5.a	Procedural validity studies
	5.b	Procedural reproducibility studies
	5.c	Procedural cost studies
	5.d	Procedural acceptability studies

Panel 6.1 Selected Terms and Concepts Relevant to General Study Design

Blinding The deliberate act of not revealing the particular intervention regimen to which a participant has been assigned in a trial[#]

Cases Individuals who have the outcome of interest

Controls Individuals who are members of a reference or comparison group

Cross-sectional study Study in which the participants are/were only assessed once for relevant characteristics (as opposed to a follow-up study where they are/were assessed more than once)

Diagnostic study Study concerned with generating evidence relevant to diagnosis in individuals or relevant to the description of patterns of morbidity/mortality in populations

Directionality Study characteristic of looking at the future occurrence of outcomes according to exposure level (forward directionality, as in classical observational follow-up studies) or of looking at past exposures according to outcome status (backward directionality, as in case–control studies)

Etiognostic study (or **etiologic study**) Study concerned with generating evidence about the possible causal role of one or more determinants of illness or illness outcome, or a study investigating the causal origins of morbidity, mortality, or health care inadequacies

Etiognostic time (or **etiologic time**) Time period during which the outcome can possibly develop due to a causal effect of the exposure

Follow-up study (or **longitudinal study**) Study in which the participants are/were assessed more than once for a same characteristic

Matching Method of selecting participants with the goal of equalizing the distribution of one or more (potential or real) confounders among levels of the exposure of interest

Prognostic study Study concerned with questions about how elements of a prognostic profile predict a particular outcome, future course of illness, or change in morbidity/mortality in a population

Prospective study Study with a prospective study base, i.e., the relevant experiences of the participants are still to happen after the start of data collection

Randomization Method of allocating intervention levels to trial participants, whereby each participant has a known and independent chance of being assigned to a particular intervention level

Retrospective study Study with a retrospective study base, i.e., the relevant experiences of the participants have already occurred by the start of data collection

Treatment arm A group of participants for whom the intervention consists of a specific medical treatment regimen

Trial Experimental or quasi-experimental study looking at the efficacy and safety of a test intervention

[#] Definition contributed by D.Willie

6.2 General Design of Clinical Diagnostic Studies

Clinical diagnostic studies, as described in Chap. 4, aim to generate information that the clinical health worker can use for diagnosing patients. As a brief reminder, the types of research information considered useful for this purpose are:

- Descriptive information from patients with an illness (or a sub-domain thereof) about the natural history of an illness and its antecedents
- Developmental information from persons without illness, attempting to describe the variability of what is 'normal'
- Information about the probability of illnesses as a function of diagnostic profile indicators
- Information about the diagnostic productivity of screening regimens
- Information about the diagnostic performance of tests used to identify a particular illness-related state

6.2.1 Case Reports *(Code 1.a)*

The *study domain* of a case report is a known, usually rare illness, a totally 'new' illness, or a particularly unusual disease presentation. Examples include a disease caused by a 'new' pathogen, or an enigmatic patient with a particular diagnostic profile for which none of the known illnesses seem to have a high probability.

The *occurrence* under study – The unusual aspects of antecedents and course of illness are the main outcomes of interest (Susser and Stein 2009). There is often a hypothesis-generating 'before-after' outlook on sequences of health-relevant phenomena. For example, 'It was noted that the onset of symptoms followed shortly after intake of substance *x*.' This type of particularistic study is therefore often considered relevant for etiognosis too (Vandenbroucke 2008) as well as for prognosis, in addition to being informative about how an illness presentation or course can pose a challenge to a diagnostician.

The *study base* is a cohort of size one (one individual), followed retrospectively from any time in relevant history until the present. Treatment may or may not have been given. Of note, occasionally treatment responses constitute diagnostically relevant information.

The *outcome parameters* can be various types of measures summarizing the unusual character of antecedents and/or the course of illness. Comparisons are made using evidence in the literature, if any is available, or using accepted 'normal ranges' for a parameter. For example, if plasma sodium levels were 160 mEq/L (normal range: 135–145 mEq/L), that individual can be said to have elevated plasma sodium.

6.2.2 Case Series Studies *(Code 1.b)*

The *study domain* of a case series study can consist of persons with an identical illness profile (a case series in the traditional sense) but can, in fact, be any type of target population, ill or healthy. For example, the World Health Organization's

Multicenter Growth Reference Study (WHO-MGRS) aimed at the construction of international child growth standards. The study domain was *healthy* children under 5-years-old who were fed according to international feeding recommendations (WHO 2006). Another example of the utility of this study design is the series of eight cases of Kaposi's sarcoma in homosexual men that initiated the awareness of AIDS as a diagnostic entity (Hymes et al. 1981).

The *occurrence* under study may be aspects of the antecedents and course of illness or aspects of normal physical or mental development, often viewed in relation to basic determinants such as sex, age, and others. For example, in the WHO-MGRS study outcomes included attained triceps skinfold thickness, seen in relation to the determinants age and sex.

The *study base* can be a cohort, a dynamic population, or a population cross-section. In the WHO-MGRS study, use was made of a cohort of newborns followed till age 24 months and of a population cross-section of children 18–71 months.

The *outcome parameters* are often descriptors of distributions of outcome variables by sex and age (e.g., the construction of growth and development standards or 'reference distributions' of any type of attribute with any scale property). As to estimation methods, in longitudinal case-series studies, the description of the age-dependent distributions may require growth reference curve construction methods. For orientation about the choice of such methods, *See:* Borghi et al. (2006). In addition, the LMS modeling method of Cole and Green (1992) is flexible, often appropriate, and easily applicable (e.g., via*:* Growth Analyzer 2009). In the WHO-MGRS study the outcome parameters estimated were selected centile values; Z-score values; and L, M, and S values describing the distributions of attained weight, length, triceps skinfold thickness, and other outcome variables for each age and sex.

6.2.3 Diagnostic Probability Studies *(Code 1.c)*

This type of study, also called the diagnostic prevalence study, has been discussed in depth by Miettinen et al. (2008) and Miettinen (2011b).

The *study domain* of a diagnostic probability study consists of patients presenting with a diagnostic profile containing some defined common key elements, for example 'adult patients presenting with cough and fever.' It is a type of patient that poses a diagnostic challenge.

The *occurrence relation*: The outcomes are the presence (yes/no) of one or more defined illnesses, say pneumonia and flu. The determinants are a range of features that are part of the diagnostic profile of the patient. Those features include elements of the *manifestation profile* (signs, symptoms, test results) and of the *risk profile* (environmental and behavioral risk factors known to be associated with the outcome). The interest is in how these features jointly determine ('predict') the probability of having one or more illnesses (e.g., pneumonia and flu). As was mentioned in Chap. 4, this type of knowledge can then be used to assist the clinician with the diagnostic process and with the evaluation of the informativeness of new diagnostic tests.

The *study base* is a cohort whose current status and past experience is to be documented as from etiognostic time zero (t = 0), which is practically-speaking the time of the first manifestation of the diagnostic profile.

The *outcome parameter* is a diagnostic probability function that allows calculating the probability of the defined illness as a function of the diagnostic profile indicators. The construction of diagnostic probability functions can be based on experts' opinions on the probabilities of illnesses associated with a variety of hypothetical diagnostic profile scenarios (Miettinen 2011b). Alternatively, it can be based on prevalence estimates of the defined illness based on a gold-standard diagnostic procedure applied to real patients. When diagnostic probability functions are available, the informativeness of a diagnostic test can be assessed by producing both a pre-test diagnostic probability function and a post-test diagnostic probability function and comparing their relative ability to arrive at a high enough probability for diagnosis, treatment, or referral. A general way of expressing the contribution of the test would be to quantify its informativeness as 1 – R, where R is the correlation coefficient of the pre- and post-test probabilities (Miettinen 2011a). Another approach would be to model an indicator variable for whether the post-test probability would fall in a 'decisive' range, as a function of the diagnostic profile indicators (Miettinen 2011a). For further information about modeling of diagnostic probability functions, *See:* Chap. 24.

6.2.4 Traditional Diagnostic Performance Studies *(Code 1.d)*

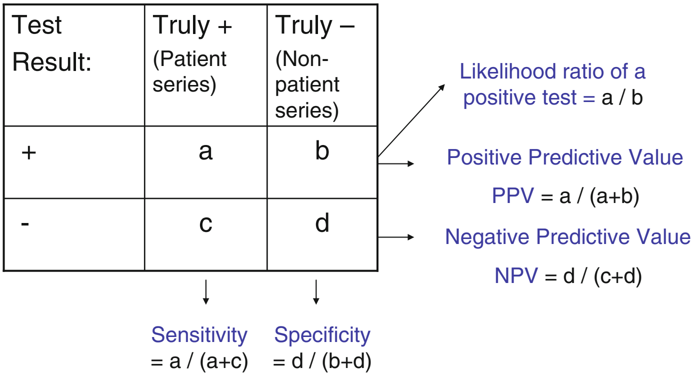
This is a family of designs widely used in clinical epidemiology to evaluate the performance of diagnostic tests and strategies.

The *study domain* consists of all patients with a type of illness for whom the diagnostic value of a 'test' or diagnostic algorithm is of interest. Note that a 'test' can be a sign, symptom, or technical assessment.

In the *occurrence relation* the outcome is illness status and the determinants are the characteristics measured by the tests. Unlike the diagnostic probability study described above, illness status is related to *preceding* test results. The purpose is to determine how well the tests, and sometimes their sequences, '*predicted*' (past tense) illness status. The somewhat odd reverse orientation of this approach (from illness or non-illness back to the test) tends to hamper the generalizability of the findings.

The *study base* is usually a patient series and a non-patient series. The difficulty is that the exact study domain and study base are difficult to define, as the patient series may have peculiar characteristics related to referral patterns or other selection processes, and the same applies to the non-cases. This issue casts doubt about the scientific generalizability of the findings of traditional diagnostic performance studies.

The *outcome parameters* most often used are sensitivity, specificity, positive predictive value, negative predictive value, and the likelihood ratio associated with specific test result levels. Figure 6.1 illustrates these concepts, which are important in practice but one needs to be aware of their scientific limitations as parameters of



Test Result:	Truly + (Patient series)	Truly – (Non-patient series)
+	a	b
-	c	d

Likelihood ratio of a positive test = a / b
Positive Predictive Value
PPV = a / (a+b)
Negative Predictive Value
NPV = d / (c+d)
Sensitivity = a / (a+c)
Specificity = d / (b+d)

Fig. 6.1 The traditional 2×2 table for assessing test performance. The sensitivity of a test is the proportion of patients with a disease who are correctly identified by the test. The specificity is the proportion of non-patients who are correctly identified as not having the illness. The positive predictive value is the proportion of individuals with a positive test result who truly have the disease. The negative predictive value is the proportion of individuals with a negative test result who truly do not have the disease. The likelihood ratio of a positive test is the proportion of patients with a positive test result divided by the proportion of non-patients with a positive test result (i.e., the odds of having an illness if there is a positive test result)

test performance. The estimates can be difficult to interpret due to the abovementioned selection/referral processes and the fact that some parameters strongly depend on the relative sizes of the patient series and the non-patient series.

The *sensitivity* of a test may depend on the mix of illness severity levels among those selected into the patient series. The test may tend to miss mild cases more often than severe cases. Thus, sensitivity as a parameter of test performance only makes sense in relation to a clearly described severity distribution. A similar problem exists with *specificity*, as false positive test results may occur more often in association with certain subject characteristics distributed among the non-patients. Such an association is therefore crucial to know about. *Positive predictive value*, *negative predictive value*, and *likelihood ratio of a positive test* depend on the relative proportion of patients and non-patients in the particular setting in which the test is going to be used. Consequently, these parameters may not be immediately relevant to practical diagnostic challenges. An additional problem with traditional test performance measures is that each of them only summarizes one aspect of the test's utility. Hence, indices have been proposed that integrate several aspects of diagnostic performance. The Clinical Utility Index (Mitchell 2011), for example, is calculated as the product of sensitivity and positive predictive value.

The analysis of a traditional diagnostic performance study will often involve 2×2 tables and may include some adjustments for estimated sources of error. Receiver Operating Characteristic (ROC) curves are also used for assessing optimal diagnostic cut-offs for continuous test results and for comparing diagnostic performance among alternative tests. For further orientation, *See:* Sackett et al. (1991).

6.3 General Design of Community-Diagnostic Studies

Community-diagnostic research is performed in community epidemiology with the aim of 'diagnosing' the burden of illness in (segments of) a population and/or to generate information about a societal 'response capacity' to morbidity (*See:* Chap. 4). Such studies are useful for policy makers to inform decisions about health care resource allocation, or for clinical diagnosticians by creating awareness about an epidemic that might facilitate the diagnostic process. There are some traditional types of general designs that are used in diagnostic-type research in community epidemiology:

6.3.1 Surveys *(Code 1.e)*

The *study domain* of a survey is a particular segment of a demographic population, or, within a geographical area, a collection of institutions or other functional units in health care. The population surveyed can be very large, as in national surveys, or relatively small, as when a single school or village is surveyed.

The *occurrence relation* usually involves many different outcomes of interest. For example, the outcomes can be a range of health-related phenomena. For any particular outcome multiple determinants can be of interest, including demographic sub-populations and various types of risk factors. The interest may be whether a factor has an association with the outcome and/or whether determinant-outcome relations differ according to third variables (i.e., effect modification).

The *study base* is one or several population cross-sections.

The *outcome parameters* depend on which of the following are of interest:

- Outcome frequencies and/or comparisons of outcome frequencies among determinant categories
- Independent and interactive effects of several determinants on the outcome
- Patterns of co-occurrence of several outcomes

6.3.1.1 Outcome Frequency Estimation

If the interest is in the occurrence frequency of the outcome (*not* in relation to a determinant), the commonly used outcome parameters depend on the level of measurement. If the outcome variable is continuous or discrete numerical, multiple descriptors of its distribution may be used, such as their mean and standard deviation or their median and 10th and 90th centiles, etc. Alternatively, the entire frequency distribution can be reported or summarized by a histogram or boxplot. If the outcome variable is categorical, the entire frequency distribution is often described. If the outcome variable is nominal or 2-category ordinal, then the *prevalence rate* is often used. Note that the survey sampling method may necessitate some form of weighted estimation, a topic that is discussed in Chap. 22. *Incidence* is sometimes calculated in surveys based on recorded histories of events over defined calendar periods preceding the survey and with proper adjustment for missing information from persons who have left the dynamic population. However, incidence is not a common outcome parameter in surveys because concerns about recall bias and proper adjustments are difficult to address adequately.

6.3.1.2 Types of Frequency Comparisons

Comparing outcome frequencies among determinant categories can be done using informal, semi-formal, and formal approaches. Informal comparisons are made by mentally comparing rates, distributions, confidence intervals, etc. available from a stratified analysis. In the semi-formal comparison, one transforms estimates to make them more comparable, a process that is accomplished using a method known as direct standardization (*See:* Chap. 22). Formal comparisons are made using statistical estimations of differences between means or proportions, estimations of *prevalence rate ratios*, and statistical testing. Common approaches to making formal comparisons are described in Chaps. 22, 23 and 24.

Regression models can be used to describe how the outcome frequency is related to several determinants. The beta-coefficients obtained from these models are estimates of the independent (descriptive) effects of those determinants. Independent relations of several determinants with a nominal or 2-category ordinal outcome variable are often assessed in *multiple logistic regression* analyses (*See:* Chap. 24). Independent relations of several determinants with a continuous, Normally distributed outcome variable are often assessed using *multiple linear regression* analyses. By inclusion of product terms in a regression model, one can also assess interactions between several determinants.

6.3.2 Epidemic Pattern Studies *(Code 1.f)*

This is a family of study types used to describe changing illness frequencies over calendar time in communities. The occurrence patterns over time and geographical space are considered essential to monitor.

The *study domain* of an epidemic pattern study consists of a population in which a health-related event of interest does or could occur.

In the *occurrence relation* the outcome is frequently the attribute of having acquired the illness of interest. In the case of infectious diseases the two categories ‘ill or not ill’ can be further split into four or more attribute levels, e.g., latently ill, patently ill, susceptible, non-susceptible/immune. The relative occurrence frequencies of these attribute levels over calendar time is then important. Many populations are under surveillance for notifiable diseases and for vital events such as births and deaths. Determinants of interest - in addition to calendar time – are age, sex, geographical area, exposure histories, etc.

The *study base* is a dynamic population under surveillance, a cohort sharing some common exposure history, or repeated population cross-sections.

The *outcome parameters* can be various statistics potentially useful for public health decision-making. In *outbreak studies* important descriptive outcome parameters (which we will not discuss further) include: attack rate, infectious contact rate, the basic reproductive rate, effective transmission factor, the herd immunity threshold, and measures of recurrence. In *surveillance* studies one is interested in time trends in numbers of cases detected or in incidence or prevalence rates. Modeling of the calendar time trends can be done using a moving averages method, cubic spline smoothing, fitting of polynomial functions, or other methods (*See:* Chap. 24).

When the size of a dynamic population is very large and approximately stable, then *pure counts* of cases in successive periods are a valid way of describing the pattern. Any rate calculation, i.e., dividing the counts by population size (prevalence rate) or population time would lead to a useless gain in validity. Additional parameters may be calculated that help with the interpretation as to whether an increase in counts in the population or in a particular segment of it (e.g., in a small area) is truly unexpectedly high. For example, such increases (known as *epidemics* when real) can be caused by new diagnostic methods that allow for the new identification of an already-existing condition, increased public awareness of the phenomenon in question, or altered notification behaviors for the event in question. These issues need to be ruled out to determine whether an apparent increase in counts is reflective of a true epidemic.

Another type of outcome parameter commonly used in epidemic pattern studies is *pseudo-rates* (Textbox 6.1)

Textbox 6.1 Pseudo-rates

Pseudo-rates are ratios that intend to approximate real prevalence or incidence rates. This is typically the case for world health statistics, for which the use of pseudo-rate outcome parameters is entirely justified for efficiency reasons. Examples include the crude birth rates and crude death rates reported by international institutions (e.g., World Bank 2010). *See also:* Chap. 22.

Crude birth rate, a pseudo-rate of population's fertility, is calculated as the ratio of (1) 1,000 times the number of live births to residents in the area in a calendar year, to (2) the estimated mid-year population in the same area in the same year. Both the numerator and the denominator of the ratio could be derived separately from different registries.

Crude death rate, a pseudo-rate of mortality burden in a population, is calculated as the ratio of (1) 1,000 times the number of deaths in the area in a calendar year, to (2) the estimated mid-year population in the same area in the same year

6.3.3 Cost Studies of Illness and Intervention *(Code 1.g)*

The economic burden of disease can be estimated with cost-of-illness studies, while the costs of interventions to prevent or manage disease are estimated in cost-of-intervention studies. The aim of this sub-section is to provide introductory guidance about how to design prospective cost-of-illness or cost-of-intervention studies. With sufficient methodological attention, cost data collected alongside epidemiological studies can later be utilized in full economic evaluations. By combining cost-of-illness and cost-of-intervention data with information about the incremental health effects of alternative actions, the analyst can provide important information that can be used to determine the best way to prevent or manage the disease in situations of resource scarcity.

6.3.3.1 The Cost-of-Illness Study *(Code 1.g. i)*

Cost-of-illness studies measure the economic burden of disease. In other words, they estimate the maximum monetary amount that could be gained if the disease was hypothetically eradicated. The aim could be, for example, to highlight the magnitude of the burden of asthma in school-going children in an area. In addition, cost-of-illness information is commonly used as input data for economic evaluations (e.g., cost-effectiveness studies, *See:* Chap. 24).

The *study domain* is determined by what is considered to be the appropriate cost perspective (i.e., the perspective from which the cost is relevant). Illnesses typically incur costs to patients or their families; to employers; and to governments and/or third party payers, such as insurance companies. Each of these entities represents its own cost perspective. The most comprehensive studies include all the above perspectives, in which case the perspective is called societal.

The *occurrence relation*: Total cost can be decomposed in terms of direct and indirect costs, capital and recurrent costs, or variable and fixed costs (*See:* Chap. 10). Determinants of those can be, for example, sub-sections of society, duration or severity of illness episodes, types of health care seeking behavior, or modifiers of efficacy of treatments.

The *study base* is multiple for analyses with a societal perspective, because the calculation of the outcome parameters is based on cost information from a variety of sources and experiences. The familial costs are often obtained from case-series or intervention study data, whereas the costs of health care utilization may require financial information obtained from health facilities, health authorities, or health insurance companies. There are two broad approaches to cost estimation. The first is to model the cost retrospectively, while the second approach is to undertake prospective costing alongside clinical or epidemiological studies. Calculation of illness costs can be incidence-based or prevalence-based. Incidence-based studies estimate lifetime costs of disease, and will therefore typically provide information about the value of averting a case. Prevalence-based costs are less data-demanding and therefore more commonly estimated; they represent a “snapshot” of the costs for a specified unit of time, and do not attempt to include longitudinal dimensions of disease progression and incidence.

The *outcome parameter* can be the costs (total, direct, indirect, familial, societal, etc.) associated with the illness episode, perhaps as a function of determinants. These estimates are expressed in monetary units, often US$, Euros, or a local currency. When the analyst wishes to adjust the cost outcomes according to purchasing power parities (e.g., to improve comparability across countries), International Dollars may be used.

6.3.3.2 The Cost-of-Intervention Study *(Code 1.g. ii)*

Related to cost-of-illness studies, cost-of-intervention studies attempt to estimate the various costs associated with an intervention.

The *study domain*: The study domain of a cost-of-intervention study is always particularistic because costs depend on pre-existing resources and functionalities within each particular community, and because local epidemiological and socioeconomic factors usually influence the findings. Like cost-of-illness studies, the chosen cost perspective will determine the study domain. A more comprehensive approach, including costs of patients and caretakers, is warranted in the societal perspective.

The *occurrence relation*: The outcome variable is the total projected cost of implementation of the preventive or therapeutic intervention. The costs are provided for a defined population, level of service provision, and time period with respect to the cost perspective of the study. Frequently, the total cost of one intervention is compared to the total cost of 'status quo' care and/or to the cost of one or more alternative intervention strategies. Adjustments are then often made to account for the effectiveness of the alternative interventions.

The *study base* may be a prospective experimental cohort if the cost-of-intervention study is nested within a trial. Several study bases are often used since the cost data and intervention effectiveness data may be from multiple sources, both primary and secondary (*See also:* Evans et al. 2005; Manheim 1998).

The *outcome parameters* may be a difference in costs between alternative interventions, typically called *incremental costs*. Sensitivity analyses or simulations are frequently performed to explore how sensitive the outcome parameters are to uncertainty in single parameters (such as coverage) and to consider the impact of the combined uncertainty in all the parameters.

6.3.4 Meta-analytical Diagnostic Projects *(Code 1.h)*

Although meta-analysis is usually done with trials and etiologic studies, it can also be done with diagnostic studies. Such meta-analytical diagnostic projects are usually ecological studies, though sometimes the individual data of the various studies can be pooled. A more detailed discussion of meta-analytical studies is found in Chap. 25.

The *study domain* and *study base* tend to be the same as for the individual studies.

The *occurrence relations* of interest are of variable nature. Determinants are frequently calendar time and geographical area. The *outcome parameters* are meta-regression parameters and pooled estimates e.g., using prevalence modeling.

6.4 General Design of Etiognostic Studies

Traditional etiologic studies *(Code 2.a)* comprise the traditional cohort study and case–control studies (also called case-referent studies). They are the most important and most widely used designs in observational causal-oriented health research today, although their validity has now been challenged (Miettinen 2010; *See:* Design *2.b*).

6.4.1 Cohort Study *(Code 2.a. i)*

In scientific cohort studies the *study domain* consists of an abstract type of persons who are at risk for the outcome. Cohort studies may have a particularistic study domain, however, if the aim is to study which factors were causal in a particular group of people (rather than an abstract group).

The *occurrence relation* under study is the relation between the health-related outcome and the determinant(s), conditional on potential and known confounders. The occurrence relation might also address mediators of disease and/or effect modifiers. The outcomes can be changes in continuous variables but are usually first-time occurrences of events of interest. Presumed causal exposures must precede the outcome; therefore, the cohort is usually selected based on the absence of the outcome(s) of interest. The exposure experience can be a time-delimited event in the past (e.g., exposure to high-dose radioactive fallout), a summary characteristic of a particular period in the past, or a stable characteristic (e.g., gene variants) present during the period that participants are at risk (i.e., before they develop the outcome of interest). The exposure must have had the time to act upon the development of the illness outcome. This is critical in diseases with a long induction and/or latency period, e.g., cancers. In other words, the exposure period must be etiologically relevant. For example, it makes no sense to study the effect of exposure yesterday on the occurrence of cancer next week. In cohort studies the interest may be in multiple exposures and multiple outcomes. When there are multiple exposures of interest, their independent causal effects, interactions (effect modification), and roles in mediating outcomes may be studied. As to the potential confounding factors, the evident concern is always with prognostic factors at the start of the etiological period and with changes in prognostic factors during the follow-up experience that are not caused by the determinant levels themselves. Special design decisions can help avoiding biases resulting from confounding or from other sources of bias (*See:* below).

The s*tudy base* is a cohort. The cohort's experience can be retrospective, prospective, or ambispective in reference to the start of study implementation.

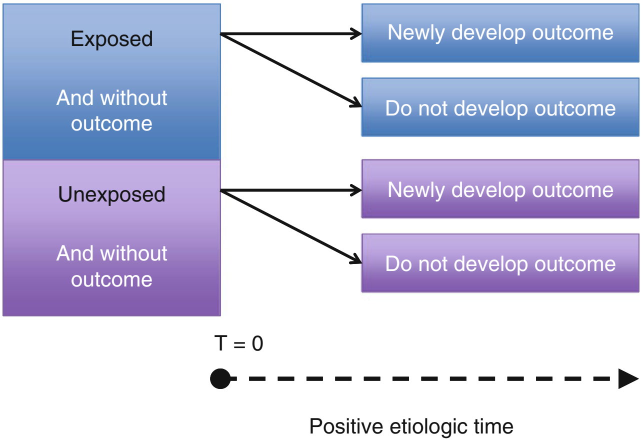


Fig. 6.2 The basic strategy of a typical *cohort study*. A cohort of at-risk subjects is selected first. By definition, at-risk members do not yet have the outcome. Members are followed during a risk period that usually begins at the start of individual follow-up (t = 0). *Positive etiognostic time* represents the interval between t = 0 and end of follow-up, which can occur due to developing the outcome, death, or loss-to-follow-up. Outcome frequency is compared among exposed and unexposed cohort members

The study base is primary, meaning that the cohort of at-risk subjects is sampled first and cases of outcome are identified subsequently in a defined risk period. The start of individual follow-up is usually considered to be the start of that risk period. Thus, etiologic time is considered positive, with time zero (t = 0) being the start of individual follow-up. Secondary study bases are used in case–control studies (next sub-section). The basic strategy of a cohort study is shown in Fig. 6.2.

The *outcome parameters* commonly made use of are adjusted relative risks, incidence rate ratios, hazard ratios, and beta coefficients of a Cox regression or log-linear regression (*See:* Chaps. 22 and 24). Possible approaches to adjust for confounding during analysis are mentioned in Chap. 22. As secondary outcome parameters (i.e., derived from the estimators mentioned above) we mention the attributable fraction (*See:* Chap. 22) and etiognostic probability functions (*See:* Chap. 24).

6.4.2 Case–Control Studies *(Code 2.a. ii)*

In scientific case–control studies the s*tudy domain* consists of an abstract category of persons who potentially have the outcome as a result of the exposure(s). Sometimes, however, the study domain is a particular confined group of people in whom an outcome of unknown etiology occurred. For example, case–control studies are often done to study the causes of a particular infectious disease outbreak (Giesecke 1994).

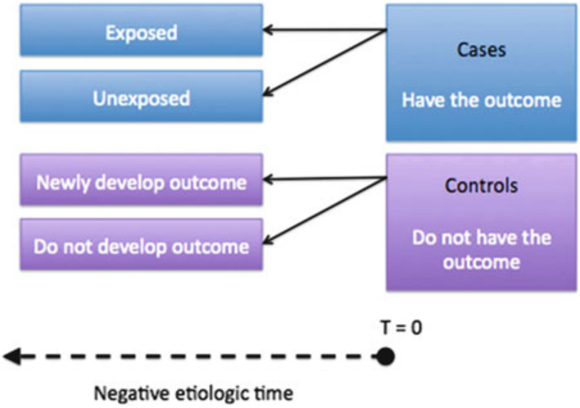


Fig. 6.3 Basic strategy of a typical *case–control study*. Cases are identified first and their exposure histories are characterized. Secondarily, a group of controls is identified, and the exposure histories of the cases and controls are compared

The *occurrence relation* is the potentially causal relationship between the health-related outcome and the exposures(s) under study, conditional on potential and known confounders. There can also be an interest in mediators and effect modifiers.

Similar to cohort studies, presumed causal exposures must precede the outcome, and here too the exposure experience can be a time-delimited event in the past, a feature of a past exposure period, or a current feature. In any case, exposures must have had sufficient time to influence the development of the outcome. Confounding factors are any potential or known determinants of the outcome that could be unbalanced between levels of the exposure under study. Special design decisions can minimize the effects of confounding and sources of bias (*See:* next sub-section).

The *study base* is secondary (in contrast to the primary study base of the cohort study). This means that cases of the outcome are identified first. Thereafter, one secondarily identifies a group of persons (called the 'controls') who collectively represent the source population from which the cases arose. The controls must represent the source population in terms of the 'usual' distribution of exposures, the purpose being to compare this distribution with the 'suspected unusual' exposure distribution among the cases. Proper selection of cases and controls and types of selection bias are discussed in Chap. 9. Definition of the source population and of the secondary study base is briefly discussed below. The basic strategy of the case–control study (Fig. 6.3) is thus to determine whether a past exposure distribution is different between cases and controls. In the cohort study, the anchor point of etiognostic time (t = 0) is the start of follow-up; however, in case control studies, the anchor point of etiognostic time is the time of manifestation (cases) or non-manifestation (controls) of the illness of interest. Etiognostic time in a case–control study is thus considered to be negative: one identifies cases and controls and then counts backward in time to compare past exposure histories. For example, one would ask newly diagnosed lung cancer patients (cases) and appropriate controls about their smoking during the period 5–10 years before the cancer occurred (or did not occur).

The *outcome parameters* tend to be different from those used in cohort studies. Often used in case–control studies are adjusted odds ratios, i.e., beta coefficients of a multiple logistic regression. In Chap. 22 we show that under certain conditions adjusted incidence rate ratios can be calculated, and we provide an overview of methods used to adjust for confounding during analysis. Methods to control for confounding at the study design stage are discussed in Sect. 6.4.3.

6.4.2.1 Defining the Source Population and Study Base in Case–Control Studies

The source population of the cases is the dynamic population from which the cases arose. It is comprised of all people who would have been eligible to be a case had that individual also developed the illness of interest. For example, if cases are identified in a hospital, then the source population is all individuals who would

- Be referred for treatment to that particular hospital if they developed the illness of interest
- Come to the attention of the case–control researchers
- Fit the eligibility criteria for participation

As mentioned, the controls must collectively represent the source population. They are a group of participants with an exposure pattern that is typical of the source population. Control sampling will have to be independent of exposure, i.e., any level of exposure must not be 'over-represented' or the inverse. Controls must not be a special group who actively avoided or engaged in the exposure. This is a frequent problem when using controls identified in hospitals and less of a concern with controls identified from the source population-at-large. When identifying such a group one needs to take into account the implications of the definition of the source population. For example, patients of a doctor who refers cases of the illness to another hospital cannot become controls. If controls can be sampled from a completely enumerated source population (e.g., from a well-defined occupational cohort or a complete list of residences), the case–control study can be labeled as population-based.

In *nested case–control studies*, the source population of the cases is simply the cohort in which the cases were identified. In such studies the controls are usually selected from those cohort members who did not become cases (Gordis 2004; Porta et al. 2008). Nested case–control studies are often carried out to study the effect of exposures that are very expensive or cumbersome to assess, such as those requiring certain biochemical analyses. The nested case–control approach leads to the possibility of doing these assessments only in cases and in a sample of the other cohort members, thereby reducing study costs without compromising study validity.

In *cross-sectional status-based case–control studies* the source population is usually assumed to be adequately represented by the non-cases in the population cross-section. With this design, there is no documentation of past exposure history preceding the manifestation of the illness of interest. The most critical assumption is that the illness of interest cannot cause the exposure (reverse causality). This is easiest to argue in the case of biological sex, socio-economic status, and other stable distal determinants of illness (e.g., genotype). A further condition is that the cross-sectional occurrence of the outcome must be nearly only influenced by frequency of development of the condition, not by the frequency of its disappearance (by death, preferential emigration out of study area, or cure). Study domain restrictions can be very useful to increase the likelihood that the study will approach this ideal. For example, in a cross-sectional status-based case–control study on possible causes of hypertension in adolescents, one may restrict the study domain to adolescents who never used anti-hypertensive medication or followed a salt-restriction diet, thereby reducing the likelihood of including cured cases in the study.

6.4.3 Designs Measures to Avoid Bias in Etiologic Studies

1. *Restriction of the study* to single narrow levels, preferably null levels, of suspected confounders can avoid confounding. For example, in a study of the potential causal effect of *x* on *y*, the confounders of interest may be alcohol consumption, socioeconomic status, and obesity. This confounding would be avoided by doing this study in normal-weight individuals of high socio-economic status who have never consumed alcohol.
2. *Exposure group matching* can be an option in cohort studies. One can try to select the ‘comparison groups’ such that prognostic factors are similarly distributed. For example, in a cohort study of the effect of occupational exposure *x* on outcome *y*, with a concern for confounding by age, one could opt for a primary study base composed of workers from two occupational settings: one in which exposure *x* is frequent and the other in which it is non-existent. If several candidate settings with non-existing exposure to x are eligible, one would choose the one in which the age distribution is very similar to the age distribution in the index setting.
3. *Individual matching* may also be considered in cohort studies. This approach consists of matching ‘exposed’ with ‘non-exposed’ subjects on the basis of potential confounders. It can be a helpful strategy if the study base is primary. When the study base is secondary (case–control studies), the matching of cases and non-cases (‘controls’) is generally ineffective as a means of control for confounding. To see this, simply recall that in order to eliminate confounding, the distribution of the confounder needs to be equal(ized) across levels of the exposure, not across levels of case status. Individual matching in case–control studies must therefore be discouraged (Miettinen 1999).
4. *A prospective design* may avoid some bias by allowing better standardization of aspects of health care; follow-up procedures; the timetable of contacting participants; and the types, accuracy, and precision of measurements. However, even with this added level of control, loss to follow-up may be more strongly related to prognosis in one exposure group than in the other, even if occurring at a similar rate in exposed and non-exposed participants. This may thus spuriously change the contrast in the remaining subjects. Losses to follow-up

(censoring of information) therefore need to be carefully avoided, and reasons for any losses to follow-up need to be recorded. This may be more feasible in a prospective study.

Another issue, typical for prospective designs, is whether there are any design decisions that can prevent dilution or reversal of exposure contrasts. This issue is common in studies where the exposure is not a past event but a 'continuous' exposure ongoing during follow-up. Let us consider an example in which a researcher is comparing a group of smokers (exposed) with a group of non-smokers (non-exposed). During follow-up some of the exposed may quit smoking (i.e., reversal of exposure), and some of the unexposed may start smoking. It is often unethical to influence reversal of exposures. Indeed, it would be unethical to advise people to continue or start smoking. A prospective study, however, allows for monitoring and documenting behavioral changes during follow-up, thereby allowing for adjustments in the analysis stage of the study.

5. *Blinding of the researchers* may be of help to control for confounding. Researchers may have strong expectations about the existence or direction of an association between risk factor and outcome. Indeed, preconceptions can influence the researcher's performance. This may lead to an unintentional trend to positively identify expected outcomes among exposed or to mistakes in the analysis. This can happen with retrospective as well as in prospective designs. Blinding of measurers/investigators as to the exposure status during data collection and analysis can be a useful design decision.

6.4.4 The Single Etiologic Study *(Code 2.b)*

We have mentioned that, in traditional cohort studies, etiognostic time is seen as positive and the study base is primary. In case–control studies etiognostic time is negative and the study base secondary. The single etiologic study proposed by Miettinen (1999, 2004, 2010, 2011b) differs from both these traditional approaches:

The *study domain* and *occurrence relation* are as usual but etiognostic time is always treated as negative, irrespective of whether the *study base* is chosen to be primary or secondary. This is considered necessary because etiognostic time can logically only be negative: etiognostic issues can only be about whether something occurring in a defined period *prior to* an outcome was causally linked to that outcome. Note that the traditional cohort study fails in this respect, as etiognostic time is positive in a cohort study. If a secondary study base is chosen, that study base needs to be a representative sample of the source population of the cases, *without any consideration around case or non-case status* (in contrast to the typical case–control study). No matter whether the study base is primary or secondary, a case group is compared to a reference series, the latter being a group preferably randomly sampled from the source population. Note also that the traditional case–control study typically fails to define the reference series as a random sample of the source population. This representative sampling is necessary because it is the proper way to arrive at valid direct estimation of the incidence rate ratio as the *outcome parameter* in etiognostic studies (Fig. 6.4).

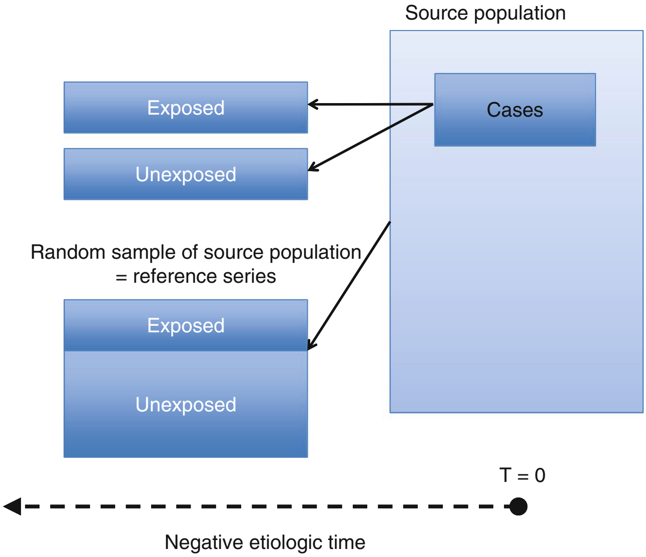


Fig. 6.4 The basic strategy of the *single etiologic study*. Exposure history (in the relevant negative etiognostic time segment) is characterized first in a group of cases arising from a source population, which can be a cohort or a dynamic population. Once the source population is defined, a representative/random sample of it is identified and the exposure history of the reference series is assessed

6.4.5 Before-After Etiognostic Studies *(Code 2.c)*

The before-after etiognostic study, like the previous study types, compares two levels of a determinant in terms of outcome frequency, and control for confounding is attempted albeit in a less formal way. The aim is often to assess the impact of a non-randomized policy intervention in a particular population.

The *study domain* is always particularistic.

As to the *occurrence relation,* the outcome in the simplest before-after etiognostic study is a change in a population's burden of an illness over a specified period, and the determinant is a policy intervention implemented (usually) over the same period (Fig. 6.5). One frequently omits measurement of the outcome under the reference level of the determinant (in an area where the policy was not implemented) based on the assumption that the change in the population's burden of illness would be zero without the intervention, or based on the assumption that some trend or projection (as estimated from an external source) will apply to the null level and can therefore serve as a reference state for comparison (as a 'counterfactual'). Formal control for confounding is also frequently foregone, based on the assumption that, during the observed period, no other major factors caused a change in the population's burden of illness aside from the implemented policy.

The *study base* is the dynamic population under study.

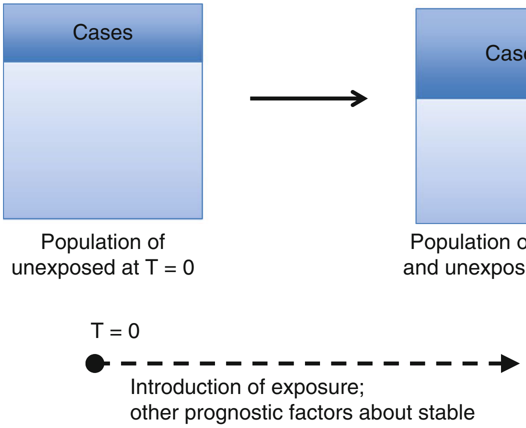

Cases
Cases
Population of unexposed at T = 0
Population of exposed and unexposed at T = 1
T = 0
Introduction of exposure;
other prognostic factors about stable

Fig. 6.5 Basic strategy of a typical *before – after etiognostic study*. The population burden of the outcome is assessed before and after the exposure occurs or is introduced. The change in population burden is attributed to the exposure assuming there that the influence of confounders was negligible or was adjusted for

The *outcome parameter* is usually the change in population burden as such, or, the amount by which the change differs from an expected value.

It should be noted that before-after etiognostic studies can be valid and yield convincing results, provided of course that the mentioned assumptions are reasonable. For example, a dramatic decrease of incidence of a water-born infectious disease may be shown to follow a massive immunization campaign against this disease, while over the same period, no simultaneous policies to improve sanitation and hygiene were implemented and no decrease due to seasonal variation is expected.

6.4.6 Meta-analytical Etiognostic Projects *(Code 2.d)*

The discussion on meta-analyses here is shortened for space constraints; however, a further discussion of meta-analyses is found in Chap. 25.

The *study domain* of a meta-analytical etiognostic consists of the type of individuals or other observation units who potentially have the outcome of interest as a consequence of the exposure(s).

The *occurrence relation* is usually the overall relationship between the determinant(s) and the health-related phenomenon, adjusted for potential and known confounders. However, there is often also interest in effect modifying factors that cause heterogeneity in individual study results. In other words, meta-analyses often aim to determine the overall exposure-outcome relationship and seek to identify sources of heterogeneity that might explain why some studies report one result whereas others report a contradictory result.

The *study base* is the collective original cohorts, dynamic populations, and/or population cross-sections.

The *outcome parameters* are statistics demonstrating heterogeneity as well as overall fixed and/or random effect estimates summarizing the collective evidence from the original individual studies (*See:* Chap. 25).

6.5 General Design of Intervention-Prognostic Studies

This section deals with the general design of some typical types of intervention-prognostic studies: the randomized controlled trial, the quasi-experimental trial, and the cross-over trial, as well as one rare type of intervention study: the N-of-1 trial. We also briefly mention meta-analyses of trials. A *trial* is a follow-up study looking at the effects, intended and unintended, of one or more levels of at least one test intervention (assigned in the context of the research study) on outcomes of interest. One type or level of intervention is chosen to be the reference with which the other intervention levels (called index levels) are compared. A *randomized trial* is a study in which each observation unit has an independent and known chance of being allocated to each of the intervention schemes under study. In a *quasi-experimental trial*, considerations other than independent chance alone determine the allocation of intervention levels (e.g., considerations of personal preference of participants or communities). In a *cross-over trial*, observation units undergo randomized sequences of intervention levels. In an *N-of-1 trial*, which is a particularistic intervention study, there is only one observation unit, but the interventions under study are allocated in successive, presumably independent treatment periods.

6.5.1 Randomized Controlled Trials: RCT *(Code 3.a)*

This class is a family of related general study designs and, before describing the design elements common to all, we give an overview of the most prominent family members, some of which form a family of their own.

1. The *traditional clinical trial* compares the course of illness between index and reference patients over a specified and fixed period of individual follow-up time after the start of an intervention. Individuals are assigned to the index or reference groups in a random fashion by randomization, with the aim of making the groups prognostically comparable at the start of the intervention. The basic strategy of a typical traditional clinical trial is illustrated in Fig. 6.6.
2. The *dose–response trial* addresses several levels of a same type of treatment and has a special interest in the shape of the relationship between treatment level and outcome variable. In a simple dose–response design there is randomization into interventions of different intensity (e.g., differing drug doses). Usually, it has more than two intervention arms. In a *randomized withdrawal design* randomization is into two groups with equal initial interventions, but then one arm's intensity of intervention is gradually decreased during follow-up.
3. In the *cluster-randomized field trial* intervention levels are targeted to (members of) groups (clusters). They can be useful for a number of reasons (Smith and

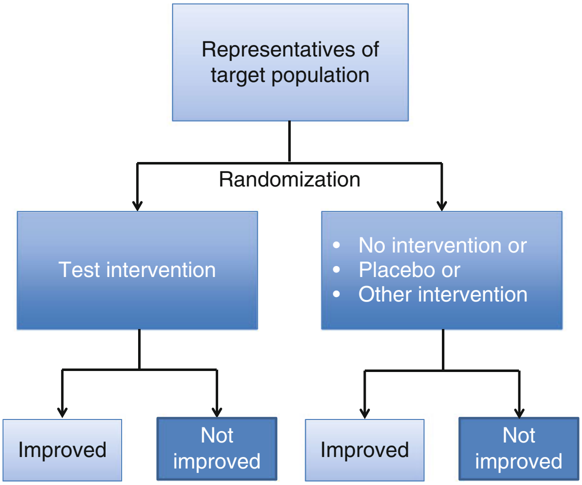


Fig. 6.6 Basic strategy of a typical *traditional clinical trial*. A sample of representatives of the target population is randomly assigned to either receive a test intervention or a comparison intervention. Outcomes of interest (in the example, improvement yes/no) are assessed during a fixed follow-up period

Morrow 1996). First, the tested intervention may naturally concern whole communities, e.g., education or sanitation interventions. Second, there are logistical advantages of individuals living close together. Third, if all individual members of a selected cluster are treated, one avoids the potential embarrassment created by visiting only certain individuals in close communities. In the classic design, clusters would be randomized to study intervention at baseline. The *stepped wedge design* is an alternative design where all clusters would eventually receive the intervention, though commencement of the interventions occurs at multiple time points. Indeed, it is the use of multiple time points, at which clusters cross from no intervention to the study intervention that allows comparison between clusters. An example is the introduction of a new vaccine or new vaccine protocol following demonstrated *efficacy* in Phase 3 trials. For operational logistic reasons the vaccine would be introduced in districts or other geographical areas (clusters) in staggered sequence. This staggered introduction could be done in the context of a trial with a stepped wedge design.

4. *Multiple intervention designs* allow the investigator to study multiple intervention types simultaneously. The most frequently used is the *factorial trial*, in which all possible combinations of the different treatments define the study arms. In the simplest case, the 2×2 factorial design trial, there are two treatments, *a* and *b,* to be combined into four intervention levels: *a* alone, *b* alone, *a* plus *b*, and no *a* plus no *b*. The advantage of the 2×2 factorial design is that the total sample size required to estimate the main effects is only half the total sample size required to do the two intervention experiments separately (this is because one intervention is equally distributed within the other, so data can be used to study the effects of both interventions). An additional advantage of the factorial design is that it allows estimation of the effects of one treatment at each level of the other treatment, which allows for examining interactions between treatments. There exist many other typical designs that combine the study of

many intervention types in a single experiment. Unlike the factorial design, they do not combine all possible levels of all treatments and are therefore referred to as '*incomplete experimental designs*.' These types of studies are often done in industrial and agricultural sciences, yet are seldom in epidemiology. Examples are Latin squares, Graeco-Latin squares, and incomplete block designs. For further reading on multiple invention studies, *See:* Kirk (1994) and Armitage and Berry (1988).

Common design features of randomized controlled trials include the following:

The *study domain* consists of a type of individual or population in which the index intervention level(s) could have a beneficial effect and a favorable safety profile.

In the *occurrence relation* the interventions are the exposures of interest. The outcomes of interest may be the occurrence and/or timing of desired and undesired events, or changes in continuous health-relevant attributes. The outcome variable is sometimes chosen to be a biomarker (Textbox 6.2) or a dichotomous variable expressing the attribute of 'treatment failure or success'. Sometimes a set of possible 'endpoints' (e.g., death, abandonment of treatment, illness worsening, etc.) may be involved in the definition of a composite attribute. There is always an interest in controlling for confounding, and there may also be an interest in effect modification. As to the potential confounding factors, the evident concern is always, as will be described below, with how successful the randomization was in balancing prognostic factors among intervention arms, and with possible changes in prognostic factors during follow-up (that are not caused by the intervention levels themselves). This is equivalent to the concerns about confounding in follow-up based etiognostic studies.

The *study base* is a prospective experimental cohort (or rarely a dynamic population). The duration of follow-up in intervention-prognostic studies can be planned according to interests in short-, medium-, or long-term effects of the intervention, but in practice the actual duration is guided by ethical principles relating to monitoring for changing degrees of equipoise and shifting balances of safety parameters.

The *outcome parameter* can be an (adjusted) relative risk, incidence rate ratio, hazard ratio, difference in incidence risk/rate/median time till events, or an (adjusted) difference in change in a continuous outcome variable. When the study aim is qualitative (about the existence of an effect) rather than quantitative (about the magnitude of an effect), the P-value from a statistical test is the typical outcome parameter. *Intention to treat analyses* produce estimates and P-values based on comparisons of intervention groups *as initially randomized* (White et al. 2011), irrespective of whether a participant is known to be non-compliant with the randomly assigned intervention. In RCT's a secondary outcome parameter can be, among others, the 'number needed to treat' (Cook and Sackett 1995), and the preventive fraction. Miettinen (2010, 2011b) has proposed that evidence from trials can be used to construct prognostic probability functions (*See:* Chap. 24), and presented as a smooth-in-time risk prediction function (Hanley and Miettinen 2009) that uses intervention level as well as individual prognostic factors as predictor variables.

Textbox 6.2 Biomarkers

Biomarkers include cellular, biochemical, or molecular indicators of biological, subclinical, or clinical effects (Porta et al. 2008). Biomarker levels are most commonly used as primary endpoints in clinical trials in situations where the biomarker is strongly correlated with the clinical outcome-of-interest and yet measurable earlier in the course of disease/follow-up. An example would be the use of biochemical lipid profiles as an outcome rather than coronary events, even though reduction in risk of coronary events is the long-term goal. Beyond the use of biomarkers as primary endpoints, biomarkers can also be used in interim analyses of clinical trials as a *basis for stopping rules*. If a 'validity trial' indicates that a surrogate biomarker is a valid predictor of an adverse outcome-of-interest, that biomarker can be analyzed during the trial to monitor for potential harm to participants. Such interim analyses are often conducted by independent Data Safety Monitoring Boards.

6.5.1.1 Methods of Randomization

Random allocation of intervention types or levels aims to ensure that treatment groups are comparable as to the baseline prognostic factors that could act as confounders (known and unknown). Several methods of randomization exist, though not all are optimal. Randomization must be fully concealed, i.e., it must be impossible for the researcher to know what treatment the next enrolled subject will get. Poorly concealed methods – such as tossing a coin, the sealed envelope method, and lists of random numbers for sequential allocation – are amenable to undue manipulation and are therefore sub-optimal. Better methods include third-party randomization, in which an independent person allocates a random assignment (over the phone, online, or by some other method) and keeps the randomization list secret until after the study analysis is complete. A similar approach might involve an independent pharmacist, who prepares randomly numbered medication or placebo packages and conceals the chemical contents from the patient and physician; packages are then sequentially allocated to patients. In general, randomization methods that involve third-party randomization are considered best.

With non-stratified randomization, chance alone decides the allocation. However, there may still be considerably more patients with a worse prognosis in one of the randomized groups. *Stratified randomization* (in strata of a prognostic factor) can alleviate this issue by optimizing the prognostic comparability of treatment groups. In other words, stratified randomization increases the likelihood that both groups are equal in prognosis. Such a method involves separate randomization lists within prognostic groups, e.g., subjects with a good prognosis are randomized using one randomization list and subjects with a poor prognosis are randomized using another list.

Block randomization aims to ensure a constant balance of numbers enrolled in the different arms throughout the enrollment period. It bases the allocations on randomly ordered intervention arms within small blocks of a fixed size. The rationale is usually to maintain balanced treatment group assignments even when enrollment is prematurely halted.

Minimization is an alternative to blocking and stratification (Treasure and MacRae 1998). With minimization, the chances of a next participant being allocated to a particular treatment arm depend on any random imbalances among treatment arms in terms of important prognostic factors. For example, if there is an accumulation of individuals with a bad prognosis in group *A*, the next individual with a bad prognosis is more likely to be assigned to group *B*, the result being to decrease the imbalance between groups.

6.5.2 Design Decisions to Avoid Bias in Trials

In addition to randomization, there are many other design decisions that can enhance the internal validity in a trial, including the following:

1. *Standardization of procedures* – To maintain comparability between intervention groups during follow-up, the groups need to receive equal attention (through standardization of procedures) with regards to study contacts (frequency, duration); concomitant treatments and support; and types and quality of measurements. Put most simply, differential treatment of groups needs to be avoided.
2. *Blinding* – To avoid prognostic divergence for reasons other than the intervention, it is crucial to include blinding of treatment allocations. In *single blinding*, the patient does not know his or her type of intervention (but the study staff is not blinded). This approach is usually inadequate, as study staff may have a conscious or unconscious bias towards one treatment group or another. In the more preferable *double blinding*, neither the patient nor the investigator/data collectors know what type of intervention the patient is receiving. In *triple blinding*, the data analysts are also blinded from the treatment allocation and are only told whether a participant has been assigned to group *A, B, C,* etc. It is worth noting that for blinding to be successful, the intervention's appearance, taste, texture, and other discerning properties need to be the same in each treatment arm. If the study participant, investigative staff, or analysts can discern one group from the other, then blinding is not possible.
3. *Promotion of retention and adherence* – The degree to which the patient correctly adheres to the allocated treatment is a prognostic factor. This generally means that good compliers tend to fare better than poor compliers, even under placebo treatment. Propensity to compliance can be randomized but, especially without double blinding, compliance rates may diverge over time, e.g., if groups are differentially influenced by participant-staff interactions or by whether the participants believe they have been assigned to a particular group. Drop-out for reasons unrelated to prognosis tends to only affect the power of statistical analysis. But if drop-out is more related to prognosis in one group

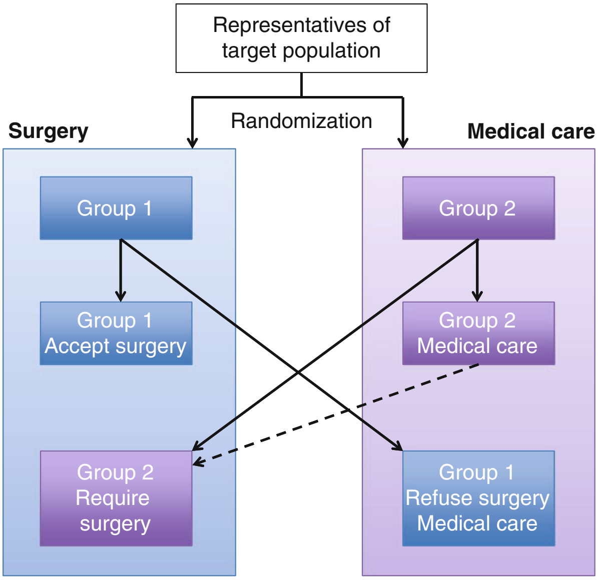


Fig. 6.7 An example of *unplanned cross-over* among treatment arms in a trial. Participants are randomly assigned to receive either medical or surgical treatment. Part of those assigned to surgery refuse surgery and get medical treatment, and part of those assigned to medical treatment require surgery during follow-up. The consequence is an unintended change in contrast between the initially randomized groups

than in another, then drop-out tends to also affect comparability and create bias. Efforts to avoid drop-out should therefore be equal in all comparison groups, and the reasons for dropout should be monitored with a special focus on whether the drop-out is prognosis-related. Additional issues of retention and adherence are discussed further in Chap. 17.

4. *Avoidance of unplanned cross-over* – A special form of lack of adherence is unplanned cross-over, which changes the exposure contrast between intervention groups. Maintaining the exposure contrast during follow-up can be challenging. Figure 6.7 illustrates this for a hypothetical scenario in which patients are be randomized to receive either medical or surgical treatment. In this particular example, cross-over is largely unavoidable. Occasionally one may be able to define the treatments not strictly as 'medical' and 'surgical' but as broader types of treatment strategies (including rules about changes in treatment under certain conditions). This could then be accompanied by a restriction in study domain to patients with levels of severity that allow any of the alternative strategies to be initiated.

6.5.3 Quasi-Experimental Trials *(Code 3.b)*

In a quasi-experimental trial other considerations than chance determine the allocation of intervention levels.

The *study domain* of a quasi-experimental trial, in principal, does not differ from the study domain of a randomized controlled trial.

The *occurrence relation* is also similar to the randomized controlled trial, but in the quasi-experimental trial confounding cannot usually be effectively controlled for by design decisions. Thus, confounders more often need measurement and formal inclusion into the occurrence relation at the analysis stage. In quasi-experimental trials, intervention-level allocation can be preference-based (preference of patient and/or physician) or involve some other systematic approach, e.g., alternate allocation, alternate-day allocation. The treatment allocation method is thus influenced by factors other than chance alone. These other factors may be related to prognosis, e.g., the physician may be convinced that the test treatment is superior and make sure all of the high-risk patients get it, leaving the control arm to the low-risk patients. The result is that the treatment arms are unlikely to be prognosis-equivalent at baseline. Randomization is therefore generally preferred, as it intends to randomly distribute prognostic factor levels over the treatment arms, and thus to make the treatment arms comparable in terms of prognosis. Nevertheless, adjustment of all relevant baseline prognostic factors in the analysis can help to overcome the aforementioned disadvantage of quasi-experimental trials.

The *study base* of a quasi-experimental trial is usually a cohort. It is generally preferable to have *concurrent controls*. However, sometimes person-time from *historical (treated or untreated) controls* is used, in which case the researcher must rely on the quality of the controls' medical records, and special concerns arise about comparability of information with the treated group.

For *outcome parameters*, *see*: randomized controlled trials. When historical controls are used, efficacy of the test intervention is sometimes estimated by (1) constructing a prediction model of the outcome of interest by using the controls' data, and (2) comparing model-predicted outcomes with observed outcomes among those receiving the test intervention.

6.5.4 Cross-Over Trials *(Code 3.c)*

Cross-over trials differ from randomized controlled trials in that the observation units are not randomized into a single intervention level but into a sequence of several intervention levels, often including a run-in null intervention period and sometimes also 'washout' null intervention periods between active intervention periods (Fig. 6.8). The purpose of the cross-over trial is *not* to learn about the relative effects of particular sequences but about the effects of the component treatments, obviously under the assumption that 'carry-over' effects between phases are negligible. Consequently, cross-over trials are not useful if one or several treatments included in the sequence have a substantial effect on the course of illness, e.g., if it cures the disease within follow-up time or if one of the treatments is surgical.

The *study domain* is usually a chronic disease about which the treatments being evaluated might provide relief of symptoms.

The *occurrence relation:* The outcome is illness status or symptom status; the determinants are the intervention types or levels. Confounders are taken into account through randomization. Effects can be modified by period (period effects).

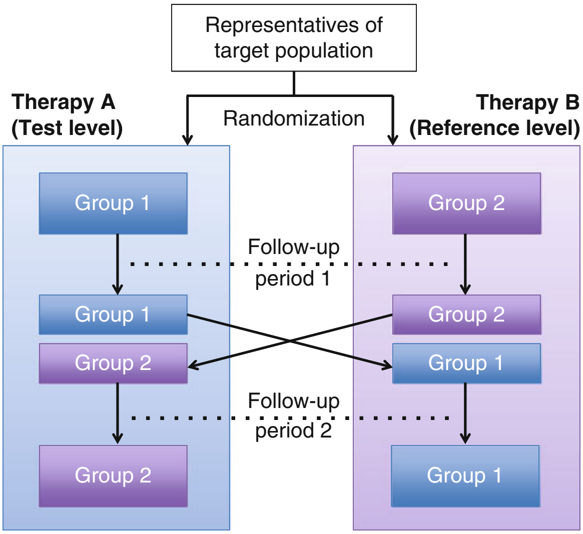


Fig. 6.8 Basic strategy of the typical *cross-over trial.* Enrolled participants are randomly allocated to a sequence of treatment periods, possibly interrupted by short washout periods

The *study base* is a cohort experience over specified successive periods. A run-in period may allow the researcher to do extensive baseline assessments and sometimes to increase comparability of the groups, e.g., by standardizing diet for 2–4 weeks prior to the first intervention period. When a small carry-over effect from one intervention period to the next is considered plausible, then a washout period of an appropriate length may be inserted before start of the next period.

Outcome parameters: The main effects of the treatments are usually studied with t-tests. Even with washout periods, there can be period effects, i.e., the effect of a treatment may depend on whether it comes first or second. Therefore, more advanced analytical approaches may be necessary to investigate treatment-period interactions. For further reading on this topic, *See:* Armitage and Berry (1988).

6.5.5 N-of-1 Trials *(Code 3.d)*

N-of-1 trials can be useful in illnesses where individual variation in treatment responsiveness is known to be large and the optimal dose or type of intervention cannot be assessed in another way.

The *study domain* is particularistic because the aim of an N-of-1 trial is to select the optimal treatment for a single individual patient.

That individual solely constitutes the study base (i.e., a cohort of size 1). Generalizability (external validity) beyond the individual patient is obviously limited. The N-of-1 trial must therefore be seen as a structured attempt at therapy or as a procedure to assess patient responsiveness rather than as a scientific experiment. The methodology of N-of-1 trials developed from investigations of adverse drug reactions, as in 'de-challenge and re-challenge' tests. Similar to such tests and to cross-over trials, the N-of-1 trial is based on an assumption that changes in treatment will have fast effects.

As to the *occurrence relation*, we briefly mention the most widely used method (Sackett et al. 1991). This method involves a single patient who undergoes a series of pairs of treatment periods. Each pair of treatment periods includes one period of the 'experimental therapy' and one period of a placebo or other control treatment. The order of treatment type within each pair is determined by randomization, and to the extent possible, both the clinician and patient are blinded to the treatment being given during any given period. The outcomes are usually relief of symptoms/signs and these are monitored continuously or very often.

6.5.6 Meta-analytical Intervention-Prognostic Projects *(Code 3.e)*

In meta-analyses of trials the s*tudy domain* consists of the type of person who potentially benefit from the intervention.

The *occurrence relation* under study is (1) the overall relationship between the health-related outcome and the intervention(s), conditional on the potential and known confounders, and (2) modifying factors related to heterogeneity in study results.

The *study base* is the evidence from the included cohorts and/or dynamic populations.

The *outcome parameters* are statistics demonstrating heterogeneity as well as overall fixed and/or random effect estimates summarizing the aggregate effect observed in the collection of individual studies included in the analysis (*See:* Chap. 25).

6.6 General Design of Descriptive-Prognostic Studies

Addressing descriptive-prognostic research questions typically involves the construction of risk functions and/or survival functions for outcome events of interest. Descriptive-prognostic studies can be set up specifically and uniquely with this goal in mind. More often, however, descriptive-prognostic research questions are part of a set of research questions addressed in one study that also addresses etiognostic or intervention-prognostic questions.

6.6.1 Clinical Prediction Studies *(Code 4.a)*

The *study domain* of a clinical prediction study can be either a type of persons for whom the risk of development of a particular illness or a particular sickness pattern is of interest, or, a type of ill persons for whom the risk for a particular course of illness is of interest.

The *occurrence relation* of a descriptive-prognostic study only concerns the outcome and determinants, and there are *no concerns about confounding or mediation*. The outcome is an event potentially occurring sometime after the prognosis is made. The predictors tend to be individual prognostic indicators, individual treatments, behaviors, and adherence issues.

The *study base* can be an experimental cohort, though observational cohorts are most often used.

The *outcome parameters* are the coefficients and terms of the risk prediction function, together with the results of validation studies of this prediction model (*See:* below and Chap. 24). The construction of the risk function tends to be a straightforward extension of the analysis of a trial or follow-up etiognostic study. For example, a logistic regression function, which is a common output from etiognostic studies, can readily be transformed into a risk function. Similarly, a trial can be analyzed using treatment level as one of the independent variables in a logistic regression analysis, and the logistic regression results can be transformed into a risk function (Miettinen 2010). Nevertheless, issues of over-fitting and useful model reductions may require special attention. An extensive overview of issues and practical methods of developing validated clinical prediction models can be found in Steyerberg (2009).

6.6.2 Forecasting Studies *(Code 4.b)*

The community medicine equivalent of the clinical prediction study is called the forecasting study. The general design is equivalent.

The *study domain* of a forecasting study consists of a type of community for which there is an interest in the future risk for a particular pattern of morbidity or in the risk of a particular change in an existing pattern.

In the *occurrence relation* the outcome is represented by an ecological variable, e.g., the prevalence of malaria or increase of malaria incidence above a chosen alarming threshold value. The predictors also tend to be only ecological variables although individuals' data may be involved when a multi-level approach to analysis is taken. In the example of alarming malaria incidence, the predictors in a dynamic forecasting function may include, among others, mosquito density, species composition and behavior, hydrological and meteorological variables, parameters of resistance to anti-malarial drugs, and other variables characterizing the epidemic pattern of malaria in the same or surrounding areas.

The *study base* is one or several experimental or observational cohorts that have provided measurement values of outcomes and predictors.

The *outcome parameters* are the coefficients and terms of a forecasting function, usually together with the results of validation studies. If a surveillance system is in place, then predictor information can be used to give early warnings about a community health issue. It is important to note that the predictors themselves may require separate modeling exercises.

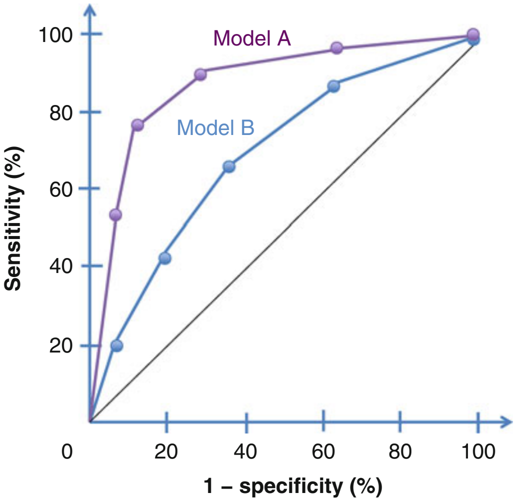


Fig. 6.9 Example of the use of Receiver Operating Characteristic (ROC) curves in the evaluation of prognostic models. Two alternative prognostic models for death within a specified period are compared. Model A appears superior since the ROC curve is closer to the *upper left corner* (i.e., the point of perfect sensitivity and specificity) and has a larger area under the curve (AUC)

6.6.3 Descriptive-Prognostic Model Validation

The validation of a clinical prediction or forecasting model can be internal or external.

Internal validation uses part or all of the data that were used for model construction. Classical methods of internal validation include the split-sample method and bootstrap validation. *External validation* is based on the application of the risk model to a group of subjects whose data were not used for model construction, e.g., patients from another site. In instances where the risk prediction score can be dichotomized and then validated against a gold standard dichotomous outcome, the area under the ROC curve (a plot of sensitivity against 1-specificity) can be used to compare the performance of alternative prediction models (Fig. 6.9).

The accuracy of the prediction model can be assessed by the goodness-of-fit between the predicted and observed risks, which can be done separately in categories of predicted risk. Making a risk function applicable to another population may require re-calibration of the function and models, and any validated function may require updating after some time (*See:* Chap. 24).

6.7 General Design of Methods-Oriented Studies

The purpose of any methods-oriented study is to create evidence about some aspect of the performance or utility of a research procedure (*See:* Chap. 4).

The domain of a methods-oriented study is the type of methodological issue or situation about which evidence is created. That domain can be either general-scientific or it may be restricted to a particular situation (i.e., particularistic).

For example, information on the agreement between measurement values obtained by two highly standardized but different measurement methods (for the same attribute) may be highly generalizable. On the other hand, the acceptability and the validity of a tool may pertain to a particular cultural setting alone and not be generalizable to other settings.

The *study base* of a methods-oriented study can be a cohort followed for a very short period (e.g., in test-retest studies), but it can also be a population cross-section.

In the remainder of this section, we give a brief overview of possible *occurrence relations* and *outcome parameters* in a number of broad sub-types of methods-oriented studies.

6.7.1 Procedural Validity Studies *(Code 5.a)*

In such studies the interest may be in the internal consistency of a single complex procedure, such as a multi-item measurement scale. Cronbach's α is a frequently used outcome parameter. There may also be an interest in rates of missing values, particular rates of misclassification, or of erroneous outliers, perhaps as a function of different circumstances. The interest of a procedural validity study may also be in determining causes of bias. In such instances, regression methods can be used to model bias as a function of a set of determinants. The aim is then descriptive or analytical.

It is sometimes possible to compare a procedure's results with those of a 'gold standard' procedure. The results of these comparisons, based on paired measurements, are frequently expressed as rates of misclassification, average bias, limits of agreement (Bland and Altman 1986), Kappa statistics (Siegel and Castellan 1988), Sign test statistics, and correlation coefficients. For more details, *See:* Chap. 11. As an extension of this logic, sometimes the relative validity of two procedures can be assessed by comparing both with a gold standard method. The procedure leading to the lowest misclassification rates, highest agreement, etc. would then be considered the more valid of the two. If there is no gold standard, then the relative validity of two alternative procedures can sometimes be compared using their convergent validity. In other words, if the attribute being measured by the two alternative procedures is known to be very strongly related to another attribute, one can examine which of the two procedures leads to the strongest relationship with that other attribute. The procedure with the strongest relationship would then be considered the more valid of the two.

6.7.2 Procedural Reproducibility Studies *(Code 5.b)*

In this study type, the interest is in the reproducibility of a single procedure. Based on independent replicate measurements, one often calculates one or more of the following reproducibility statistics: coefficient of variation (CV); technical error of measurement (TEM); reliability coefficient (RC); and intra-class correlation coefficients (ICC). These can be calculated in test-retest studies. In a test-retest study, the replicates can be made by the same observer, a different observer, or a mix of both. In this way inter- and intra-observer reproducibility statistics can be obtained. For more detail, *See:* Chap. 11.

The interest may also be in the reproducibility of a procedure under special circumstances. Comparison is then needed with the degree of reproducibility obtained in usual circumstances. F-tests may be used to make these comparisons in the case of continuous outcome variables. *Determinants of reproducibility* can be studied by modeling error rates or variances.

6.7.3 Procedural Cost Studies *(Code 5.c)*

The overall purpose of procedural cost studies is to determine the cost of specific research procedures or to determine the most cost-effective procedure for a given research aim. This may be accomplished through one or more of the following activities:

- To collect information on and evaluate costs associated with various stages of a research study. This activity is usually performed in 'pilot' or feasibility studies. Examples include assessing the cost of different sampling and recruitment strategies, such as conducting surveys by going door-to-door versus sending surveys in the post. The financial feasibility will differ by geographical location, characteristics of the study population, nature of the study questionnaire, and availability of study personnel, among other factors
- To compare the efficiency of two or more alternate procedures, usually weighing financial costs with validity

The cost of a research design or procedure consists of capital costs (e.g., procurement of necessary machinery and tools), operational costs (e.g., wages and infrastructure resources), cost to research participants (e.g., reimbursements for travel), and time allocation on the part of researchers and participants. There are also cost implications of different sample size scenarios, sampling schema, recruitment strategies, retention strategies, and follow-up procedures. Pilot studies may be conducted specifically to define these attributes and hence the costs of a research study.

6.7.4 Procedural Acceptability Studies *(Code 5.d)*

Even if one procedure is determined to be more valid than another, sometimes it may be less acceptable and therefore be less desirable. If study participants find the procedure to be unacceptable, they will not consent at the time of enrollment, and they may refuse a certain procedure after enrollment. These are undesirable circumstances; therefore, when planning a study it is important to select the procedure that best balances validity and acceptability. Documenting acceptability may be challenging, but is usually assessed indirectly using refusal rates and directly by asking participants questions about acceptability. Qualitative research methods (e.g., focus group discussions) are commonly seen as a useful alternative.

When the specific aims are clear and the appropriate general study design is chosen and described, the essence of the research plan is in place and some practical and efficiency considerations are next on the agenda of the proposal developer. First, there are considerations around study size, which will be a major determinant of a study's efficiency and the amount of information that will be produced. This brings us to Chap. 7.

References

Armitage P, Berry G (1988) Statistical methods in medical research. Blackwell Scientific, Oxford, pp 1–559. ISBN 0632015012

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet i:307–310

Borghi E et al (2006) Construction of the World Health Organization child growth standards: selection methods for attained growth curves. Stat Med 25:247–65

Cole TJ, Green PJ (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. Stat Med 11:1305–19

Cook RJ, Sackett DL (1995) The number needed to treat: a clinically useful measure of treatment effect. BMJ 310:452–454

Evans DB et al (2005) Methods to assess the costs and health effects of interventions for improving health in developing countries. BMJ 331:1137–1140

Giesecke J (1994) Modern infectious disease epidemiology. Edward Arnold, London, pp 1–256. ISBN 0340592370

Gordis L (2004) Epidemiology. Elsevier Saunders, Philadelphia, pp 1–335. ISBN 1416025308

Growth Analyser, version 3.0 (Application) (2009) Dutch Growth Foundation, Rotterdam. www.growthanalyser.org. Accessed Sept 2012

Hanley J, Miettinen OS (2009) Fitting smooth-in-time prognostic risk functions via logistic regression. Int J Biostat 5:1–23

Hymes KB et al (1981) Kaposi's sarcoma in homosexual men – a report of eight cases. Lancet 2:598–600, Sep 19

Kirk R (1994) Experimental design. Procedures for the behavioral sciences. Wadsworth, Belmont, pp 1–921. ISBN 9780534250928

Manheim L (1998) Health services research clinical trials: issues in the evaluation of economic costs and benefits. Contr Clin Trial 19:149–158

Miettinen OS (1999) Etiologic research: needed revisions of concepts and principles. Scand J Work Environ Health 25:484–490

Miettinen OS (2004) Epidemiology: Quo vadis? Eur J Epidemiol 19:713–718

Miettinen OS (2010) Etiologic study vis-à-vis intervention study. Eur J Epidemiol 25:671–675

Miettinen OS (2011a) Epidemiological research: terms and concepts. Springer, Dordrecht, pp 1–175. ISBN 9789400711709

Miettinen OS (2011b) Up from clinical epidemiology & EBM. Springer, Dordrecht, pp 1–175. ISBN 9789048195008

Miettinen OS et al (2008) Clinical diagnosis of pneumonia, typical of experts. J Eval Clin Pract 14:343–350

Mitchell AJ (2011) Sensitivity × PPV is a recognized test called the clinical utility index (CUI+). Eur J Epidemiol 26:251–252

Porta M, Greenland S, Last JM (2008) A dictionary of epidemiology. A handbook sponsored by the I.E.A, 5th edn. Oxford University Press, New York, pp 1–289. ISBN 9780195314496

Sackett DL et al (1991) Clinical epidemiology. A basic science for clinical medicine, 2nd edn. Little Brown, Boston, pp 1–441. ISBN 0316765996

Siegel S, Castellan NJ Jr (1988) Nonparametric statistics for the behavioral sciences, 2nd edn. McGraw-Hill International Editions, New York, pp 1–399. ISBN 0070573573

Smith PG, Morrow R (1996) Field trials of health interventions in developing countries: a toolbox. Macmillan, London

Steyerberg E (2009) Clinical prediction models. A practical approach to development, validation, and updating. Springer, New York, pp 1–497. ISBN 9780387772431

Susser M, Stein Z (2009) Eras in epidemiology. The evolution of ideas. Oxford University Press, Oxford, pp 1–352. ISBN 9780195300666

Treasure T, MacRae KD (1998) Minimisation: the platinum standard for trials. BMJ 317:362–363

Vandenbroucke JP (2008) Observational research, randomized trials, and two views of medical science. PLoS Med 5(3):e67

White IR et al (2011) Strategy for intention to treat analysis in randomized trials with missing outcome data. BMJ 342:d40

World Bank (2010) World development reports. http://econ.worldbank.org. Accessed Sept 2012

World Health Organization (2006) WHO child growth standards. Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age. WHO, Geneva, pp 1–312. ISBN 924154693X

Study Size Planning

7

Jonathan R. Brestoff and Jan Van den Broeck

> *Validity considerations alone are often sufficient to imply that zero is the optimal size.*
>
> Olli S. Miettinen

Abstract

In planning and proposing a study, a paramount concern is the likelihood that the study will provide useful or meaningful information. An important factor in demonstrating that a study will be informative is *sample size*. If a study has a sub-optimal number of subjects, it may be under-powered to detect statistical significance even in the presence of a true effect, or estimates produced by the study may lack useful precision. On the other hand, if a study has too many subjects, one may encounter resource limitations and ethical issues associated with exposing an unnecessarily large number of subjects to risk. An optimal study size therefore balances the need for adequate statistical power or precision, the limited nature of resources, and the ethical obligation to limit exposure to risk. As such, study proposals and scientific papers often include sections on the planning of study size. This chapter begins with an exploration of various factors that contribute to optimal study size. We then briefly review some useful sample size calculations in the contexts of surveys, cohort studies, case–control studies, and randomized trials.

J.R. Brestoff, MPH (✉)
Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck, M.D., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistr0079,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_7,

7.1 The Concept of Optimal Study Size

Assuming that a study is perfectly valid, we are presented with an issue of how informative that study will be, a main determinant of which is *sample size* or *study size*. These synonymous terms are used to describe the number of subjects in a study. A concern of most reviewers of study proposals is whether the amount of information produced by a study will be large enough to make a 'substantial contribution.' Therefore, study size becomes a critical issue not only for planning studies but also for obtaining funding.

This situation raises many difficult questions: What is a 'substantial contribution?' Is a 'substantial contribution' the same for all stakeholders and for all study objectives? Is there – or under what conditions is there – a way to determine optimal study size? These questions demand complex answers that have been the subjects of many textbooks. Space constraints preclude a fully comprehensive review of this topic here; however, the present chapter deals with these questions in an introductory manner and provides useful tools and equations to make arguments about study size. Selected terms and concepts relevant to this discussion are listed in Panel 7.1.

Panel 7.1 Selected Terms and Concepts Relevant to Study Size Planning

Clinical relevance Potential to have a meaningful effect on clinical practice

Community health relevance Potential to have a meaningful effect on community health

Dropout rate Proportion (or percentage) of *enrolled* participants who have an unplanned early cessation of individual follow-up

Optimal study size A desirable yet realistic number of study participants based on ethical, scientific, and efficiency considerations. Only participants who contribute data to the analysis (i.e., those who do not drop out of the study) are included in optimal study size figures

Population size (Abbr., *N*) The total size of a target population

Power The probability of detecting a statistically significant association of a particular magnitude or greater when a true association exists

Precision (of an estimate) Degree of lack of random error. In practice, precision is taken to be the narrowness of the confidence interval

Refusal rate Rate of non-participation among *eligible* observation units invited to participate (those who refuse to participate cannot, by definition, contribute to a dropout rate)

Sample size (Abbr., *n;* Syn., study size) (1) Number of sampled individuals to be approached for possible inclusion as participants. (2) Number of observation units that contribute data available for analysis

(continued)

Panel 7.1 (continued)

Sample size calculation An aspect of study size planning consisting of a statistical determination of the number of participants needed to satisfy concerns about precision of estimates or power to detect an anticipated effect

Sampling fraction Sample size *n* divided by target population size *N*

Sample size planning Determination of optimal study size

Significance level (of a null hypothesis test) A pre-determined P-value (α) below which an obtained P-value is labeled 'statistically significant' or at/above which an obtained P-value is labeled 'statistically non-significant'

Stakeholders (of a research study) Persons, institutions, or communities with an interest in a research study or that can be affected by a study or its results

7.1.1 Factors Influencing Optimal Study Size

The planning of study size is often presented as a pure matter of statistical calculations of sample size and/or power. We urge readers who have not already done so to embrace a broader view: that there are generally no statistical means of determining what the optimal study size is (Miettinen 1985) because non-statistical factors greatly influence what is considered to be the optimal study size. A few reflections may clarify and illustrate this point.

First, problems with internal validity reduce the amount of information gained by all studies to varying degrees, even when that information is generated from very large studies. If we imagine that an idealized perfect study provides 100 'units of information,' any real study will provide only a fraction of that amount. In a real study, the maximum amount of information that a study can provide depends on the design of the study, and as that study progresses, information is lost from various errors or issues. Thus, even if it were possible to design a perfect study with the potential to provide 100 'units of information,' measurement error and the most minor ethical mishaps (both of which are impossible to avoid completely) will cause information to be 'lost.' Since we cannot measure the amount of information that a study can potentially provide or how much has been lost, we cannot account for 'information loss' using statistical means of determining optimal study size. By an extension of logic, therefore, using *only* statistics to determine optimal study size is impossible. Statistical calculations of study size do indeed help us to maximize the likelihood that a given study will be provide an *acceptable* amount of information; however, study size calculations must always be contextualized and modified with non-statistical factors.

Second, apart from validity concerns and 'information loss,' a variety of additional factors co-determine what study size will be perceived as useful or optimal, and these factors tend to differ according to the study design. For example, the optimal size of a Phase-1 trial is very low (typically n = 6–12) because of ethical considerations. This type of study carries high risk because it represents the first time humans are being exposed to a new pharmaceutical formulation. The risks are unknown and therefore considered to be very high; therefore, an optimal sample size might be n = 10 in spite of statistical arguments suggesting that n = 50 is better. In this case, ethical considerations are valued more highly than the additional information that 40 more participants would provide. On the other hand, a Phase-3 trial might be very large (on the order n = 10,000). At this stage, the short-to-medium-term safety and tolerability of a pharmaceutical formulation is better understood, and only reasonably safe interventions are considered suitable for a Phase-3 trial. In this study design, the primary objective is to assess effect sizes and medium-to-long-term safety, both of which can be quite small and relate to rare events. Consequently, the optimal size of a Phase-3 trial using the same formulation as a Phase-I study might be three orders of magnitude larger.

Third, financial and other resource limitations can ultimately weigh heavily on the perceived usefulness of study size. Since resource availability is sometimes dynamic during a study, the perceived optimal size of a study can change during the data collection phase of a prospective study. Such changes in perceived optimal study size can relate to an entire study, but sometimes optimal study size changes for one specific aim but not another. To illustrate this point, let us consider a 3-year-long prospective study in which Specific Aim 1 is to investigate whether zinc deficiency increases the risk of acquiring acute cholera and the severity of the disease, and Specific Aim 2 is designed to test whether various factors are effect modifiers for the effect of zinc deficiency. In year 2, the funding agency experiences financial difficulties that force redistribution of research funding, requiring the research team to scale back the study. Since effect modification tends to increase the optimal study size considerably, the research team and the funding agency meet and agree to re-craft Specific Aim 2 to address only the two most important effect modifiers. Such a decision reduces the overall study size by 30 % and reduces the optimal study size for Specific Aim 2, while potentially having no study size consequences for Specific Aim 1.

From the above we deduce that scientific, ethical, and practical concerns drive study size planning. Though we highlighted only one factor for each of the three dimensions, many factors may need to be considered when optimal size is to be determined. Table 7.1 summarizes several of these factors, most of which are derived directly from the general principles of epidemiology (*See:* Chap. 1). The discussion above and Table 7.1 indicate that study size optimization is complex process involving the simultaneous consideration of numerous counter-acting phenomena.

7.1.2 Useful Precision or Power

One of the major considerations listed in Table 7.1 concerns a desired limit of precision for an estimate or a minimum power and significance for the detection of an anticipated effect, beyond which evidence is considered increasingly useless. The epidemiological

Table 7.1 Considerations for determining optimal study size for a single specific aim

Dimension of concern	Factor influencing optimal study size	Usual direction of influence on study size
Ethical	Need to maximize societal relevance through stratified analysis for sub-layers of society (Particularistic studies)	↑
	Need to minimize cumulative burden of study participation in non-minimal-risk studies	↓
	Need to minimize potential harm	↓ to 0
Scientific/methodological	Scientific interest in effect modification	↑
	Internal validity problems that cannot be adequately adjusted for in the analysis	↓ to 0
	Need for adjustments of outcome parameter estimates	↑
	Need for efficiency and parsimony of design	↓
	Minor design decisions, e.g., choice for a continuous outcome variable rather than a categorical one	↓
	Interest in effect size or shape of relationship rather than the mere existence of an effect	↑
	Existence of a desired limit of precision for an estimate; a minimum statistical power and significance level for the detection of an anticipated effect, beyond which evidence is considered increasingly useless (e.g., to a main stakeholder)	To within size leading to useful precision
	Meta-analysis	↑ to maximum
Practical	Existence of an upper threshold of study budget or a requirement to minimize costs	↓
	Existence of a restriction in access to a vital implementation resource, e.g., a particular type of study personnel	↓
	Natural limits to the amount of accessible observation units or information	↓ to limit
	Expected refusal and dropout rates	↑

and statistical literature on sample size has mainly focused on this aspect, and later in this chapter we will further expand on such statistical aspects of study size planning. As mentioned previously, the use of statistics to determine the optimal study size must be contextualized with the factors discussed above and in Table 7.1.

7.1.2.1 The Range of Useful Precision

Outcome parameter estimates consist of a point estimate, surrounded by an interval estimate. For example, a point estimate of a prevalence rate is surrounded by a 95 % confidence interval. The interval estimate is an expression of the uncertainty surrounding the point estimate and derives mainly from sampling variation as well as measurement variation/error. In general, the degree of uncertainty is inversely related to the size of the study. On one hand, if a study is too small, the uncertainty may increase to a level considered to be undesirable or useless (an exception, however, is that well-designed small studies may contribute meaningfully to later meta-analyses; *see:* Chap. 25). On the other, as the study size increases, the degree of uncertainty decreases, and the interval estimate becomes narrower.

In the latter case, increasing study size to achieve higher precision may be considered undesirable and useless beyond some threshold. For example, in case–control studies it is generally accepted that having more than 4 'controls' for each case is inefficient. In practice, the starting question is thus often: *What range of study sizes* – at analysis – *will give us an interval estimate narrow enough to be considered useful but not so exceedingly narrow as to be inefficient?* Note that this refers to final precision, after any necessary adjustments in the analysis, e.g., after corrections for misclassification of outcome or determinant.

Can this range in fact be determined, considering that 'usefulness' and 'optimal' are both subjective perceptions? We argue that subjectivity does not imply total arbitrariness. As pointed out by Snedecor and Cochran (1980), what is needed is careful thinking about the use to be made of the estimate and the consequences of a particular margin of error. In this respect, the researcher planning study size may consider that:

- Perceived usefulness of a particular precision is often influenced by the fact that narrower confidence intervals enhance the precision of any subsequent projections of cost or efficiency of envisaged larger-scale policies. Thus, when the research study falls within a comprehensive evaluation of a possible new policy, high precision tends to become a necessity.
- It may be necessary to get the opinion of some stakeholders on the matter (especially those that are providing funding). Sometimes there is an explicit wish of a sponsor to obtain evidence with a specific margin of uncertainty, e.g., a desire to know the prevalence 'within ± 1 %.' This is frequently the case in diagnostic particularistic studies, such as surveys. The stated reasons for this are not always clear. Perhaps a similar margin was used in a previous study about the same occurrence and, if a similar level of precision is reached at the end of the new study, a 2 % or higher increase in prevalence could then roughly be seen as evidence for the existence of a real change in prevalence, although this is not the ideal way make such a determination (Altman et al. 2008). When there is such a desirable margin of uncertainty, the required study size to achieve this is usually easy to calculate (*See:* Sect. 7.4). When using this approach one should not forget to take into account possible necessary adjustments, perhaps for an expected refusal rate, the sampling scheme, finite population correction, measurement error, covariate adjustment, or other reasons, as will be discussed below.
- The perceived clinical or community health relevance of particular effect sizes is important. Stakeholders sometimes set a prior threshold for an effect size as a basis for decisions, e.g., about pursuing further research, about further development of a drug or clinical strategy, or about further exploration of a public health policy. For instance, it may be stated that 'only if the effect can, with reasonable certainty, be larger than *x* can it be considered clinically relevant.' This type of expectation is generally easier to take into account using a power-based outlook

(*See:* below) rather than a precision-based outlook, although the latter has also been proposed (e.g., Greenland 1988; Bristol 1989; Goodman and Berlin 1994).

- When there is a desire for very high precision, it is unrealistic to aim for a precision that is so high that it approximates the expected variation due to measurement error.
- There are possible (dis)advantages of wide or narrow confidence intervals, beyond issues of cost and feasibility. Narrow confidence intervals can give a false impression of validity and a false impression of generalizability (Hilden 1998). Wide confidence intervals often give a false impression of lack of validity.

Sometimes, given the multiplicity of factors influencing optimal study size (Table 7.1), there is little room for choosing a sample size in studies that plan for estimation. For example, there may be an upper limit to study size that is lower than statistical calculations suggest. The question may then become: given the maximum sample size imposed, will the precision be useful and worth the effort, resources, and potential risks? The issue of sample size calculation then becomes an issue of precision calculation.

7.1.2.2 Power to Detect an Anticipated Effect with a Chosen Confidence

Many studies plan for statistical testing. In such studies the outcome parameters are test statistics with P-values, e.g., a t-test statistic with an associated P-value. Statistical power is interpreted as the probability of detecting a statistically significant association of a particular magnitude or greater (Daly 2008). An important question in study size planning is then often: What range of study sizes – at analysis – will give enough statistical power (e.g., one often uses a power of 80 % at a 95 % level of confidence) to detect true differences of magnitudes considered meaningful? If the true effect is smaller than this anticipated meaningful effect, then we can accept a non-significant test result (Daly 2008). This refers to *final* power after any necessary adjustments in the analysis, e.g., after corrections for misclassification of the outcome or determinant (Edwards et al. 2005; Burton et al. 2009). Based on this, a statistical sample size calculation can often be done. In the later sections of this chapter examples will be given.

In current epidemiological practice, the abovementioned type of sample size or power calculation is frequently performed, not only in studies that plan for testing but also in studies that plan for the estimation of effects. This may partly be because the methods for precision-based sample size calculation are not yet fully part of epidemiological tradition and are less well known, less developed, and sometimes more difficult to use. This is one of the factors that perpetuate the use of statistical testing in studies that do not need it.

Sometimes there is little room for choosing a study size in studies where statistical testing is planned. The question that needs to be addressed in that case may be whether the statistical power of the study is expected to be useful or whether the power would only provide for detecting effects that are so extreme that one might as well abandon the study plans. The issue of sample size calculation then becomes an issue of power calculation.

7.2 The Process of Study Size Planning

As we have noted above, the process of study size planning is not only a matter of statistical calculations but also a matter of many other considerations. Below we describe the usual process of study size planning in studies where there is indeed room for choice:

In the balancing of considerations listed in Table 7.1, the focus should first go to those factors that would require reductions of the study size to zero. In other words, any major issues about design validity and ethics should be addressed and solved first. This is obviously something that should already have been done at the stage of designing general objectives, specific aims, and general study design. However, it happens regularly that proposal reviewers, statisticians consulted for sample size and power calculations, and even article reviewers come across problems of this nature. This suggests that it is valuable to reconsider this aspect at this stage of the study planning process.

Conditional on satisfying concerns about design validity, design efficiency, and ethics, remaining major issues are the useful precision of statistical estimates (or, when statistical testing is planned, the statistical power to detect useful effects with some degree of certainty) and the costs of various hypothetical study sizes. Both may need to be calculated and balanced. At this stage statistical methods may be useful. Once an opinion is formed regarding the optimal study size, the next step is to project what sizes at preceding study stages (recruitment, sampling, eligibility screening, and enrollment) are expected to lead up to this optimal size at analysis. This determination will require considerations of expected rates of non-contact, refusal, and attrition as well as anticipated adjustments for measurement error and confounders, etc.

The process described is repeated for each specific aim separately. It may then turn out that optimal sizes, at analysis or before, for different specific aims are incompatible. This may even lead to the abandonment of one or more of the initial specific aims or to their 'downgrading' to a secondary or tertiary level aim because of expected lack of useful precision or power. The balancing exercise may also entail other study design changes, e.g., a choice for a more efficient design, a choice for another measurement level for the outcome variable, a reduction in the size of a reference series, etc. The balancing effort may even lead to the conclusion that financial resources, time, and availability of subjects do not allow for continuation of the study plans (Miettinen 1985).

7.3 The 'Sample Size and Power' Section of the Study Proposal

Written justifications of the chosen study size are usually located in the 'sample size and power' section of the study proposal or the methods section of a paper. These justifications may need to include elements listed in Panel 7.2.

Panel 7.2 Elements for Inclusion in the 'Sample Size and Power' Section of the Study Proposal

- Specify for which specific aim the calculations were done and why, or present the calculations separately for each specific aim
- Indicate whether a precision-based or power-based approach (or both) was used
- Indicate whether or not a more-or-less fixed or maximum study size imposed itself; if so, explain the rationale and how a precision- or power-calculation was done
- If a precision-based approach was used, mention what precision was desired and why
- If a power-based approach was used, indicate what anticipated or desired effect size was used for the calculation and what level of significance was used
- Specify formulas used, assumptions made, and sources of inputs into the calculation method
- Mention if, why, and how adjustments of estimated study size were done to allow for expected refusals, dropouts, measurement error, subgroup analyses, and control for confounders
- Mention the results of the calculations

Let us now look at study size planning and sample size calculation in some particularly common situations. In the next sections we will discuss the optimal size of surveys, cohort studies, case–control studies, and trials. For each of those we will discuss how optimal sample size is influenced by ethical, scientific, and practical concerns. As far as sample size calculation is concerned, the next sections give examples both of the precision- and power-based approach. However, it should be noted that both approaches are not given for a given scenario; if the alternative approach is desired, we recommend consulting Kirkwood and Sterne (2003) or another book of medical statistics or sample size planning.

7.4 Size Planning for Surveys

A typical survey addresses multiple specific aims, each of which often contains two or more sub-aims. These sub-aims frequently entail subgroup analyses, such as comparisons of estimates for different catchment areas or across subgroups (e.g., age categories). The planning of study size therefore often requires an extensive exploratory phase to determine the size requirements for different subgroup analyses and to use this information to derive an optimal size for the entire study. A common approach in determining optimal study size is to prioritize the specific aims and sub-aims and to consider each in the context of resource limitations. This process may then lead to revisions to or refinements of one or more specific aims and/or sub-aims. Such revisions sometimes involve abandoning certain sub-aims (especially if the associated sub-aim is very resource intensive). A common alternative approach is to retain the sub-aim in question while acknowledging that findings may be statistically imprecise or underpowered.

As discussed in Chap. 9, target populations are rarely studied in their entirety, unless that population is very small and is contained within a small area. Attempting to survey an entire target population becomes increasingly inefficient as the size of the target population or its catchment area increases, introducing an ethical issue concerning the appropriate use of limited resources. Therefore, large surveys are more likely to require statistical sampling and, as will be discussed below, the sampling proportion then becomes an important consideration in the sample size calculation. When exploring the study size implications of the various research questions addressed in the survey, the following sample size calculations may be helpful.

A note on notation:

N	Capital N	Refers to the size of the **target** population
n	Lower-case n	Refers to the size of the **sample**

7.4.1 Sample Size Calculation for Estimating a Prevalence

When the purpose of a survey is to estimate the prevalence of a health phenomenon, a main concern is the degree of confidence in the prevalence estimate. Therefore, it is said that sample size calculations for estimating prevalence are *precision-based*. The following formula is often used (Kirkwood and Sterne 2003):

$$n = \frac{p(1-p)}{e^2} \tag{7.1}$$

Where:

n = **sample size for estimating a prevalence**

p = expected proportion (e.g., 0.12 for a prevalence of 12 %)

e = desired size of the standard error (e.g., 0.01 for ±1 %)

As an example, consider a proposed Study A, in which one purpose is to estimate the prevalence of depression in people over the age of 18 years. Based on similar studies from another region in the same country, the researchers predict that the prevalence of depression will be 12 % ($p = 0.12$) in their study. They want to achieve a standard error of 1 % around their estimate (i.e., 12 % ±1 %). In order to achieve this degree of precision, the researchers will likely need to have at least $n = 1056$ participants (based on Eq. 7.1), a value that can be usefully rounded up to $n = 1100$.

It is important to note that the above equation is valid only if the calculated sample size (n) is less than 5 % of the target population (N). The value of n/N is known as the *sampling fraction*. If n/N is greater than 0.05 or 5 %, then a *finite population correction* is necessary. The use of a finite population correction is usually not necessary, as most studies do not usually sample 5 % or more of the target population. Since such a scenario is rare, we do not discuss the topic further in this chapter.

7.4.2 Sample Size Calculation for Estimating a Mean

A similar approach is used when the purpose of the survey is to estimate the mean value of a continuous health-related parameter. Again, a main concern is the degree of confidence in the estimated mean; therefore, in this case too a precision-based approach is useful. The following formula is often used:

$$n = \frac{\sigma^2}{e^2} \tag{7.2}$$

Where:
n = **sample size for estimating a mean**
σ = expected standard deviation, and
e = desired size of the standard error

As an example, consider a sub-aim of Study A, in which the goal is to compare the mean body mass index (BMI) of participants with depression and those without depression. Previous studies of the target population indicate that the standard deviation of BMI is expected to be 4.0 ($\sigma = 4$), and the desired standard error is 0.5 kg/m^2 ($e = 0.5$). Based on Eq. 7.2, Study A will likely need at least $n = 64$ participants in each group to achieve the desired degree of precision. This is a very realistic proposition because the researchers anticipate a 12 % prevalence of depression with a sample size of $n = 1{,}100$; therefore, the smallest group in which BMI will be measured will likely be $0.12 * 1{,}100 = 132$ participants. This sub-aim of Study A will therefore likely achieve greater-than-desired precision.

7.4.3 Sample Size Calculations When Comparing Proportions

When the outcome parameter is the difference between two proportions, such as prevalence estimates, a power-based approach is usually taken to calculate sample size. The following series of formulas can be useful:

$$n = \frac{c_{pp}\left[p_1\left(1-p_1\right)+p_2\left(1-p_2\right)\right]}{\left(p_1-p_2\right)^2} \tag{7.3}$$

Where:
n = **sample size for estimating the difference between two proportions**
c_{pp} = constant defined by the selected P-value and desired power
p_1 = expected prevalence in group 1
p_2 = expected prevalence in group 2

The constant c_{pp} is determined by taking the square of the sum of the Z scores for the selected P-value and desired power:

$$c_{pp} = \left(Z_\alpha + Z_\beta\right)^2 \tag{7.4}$$

Where:
c_{pp} = **constant defined by the selected P-value and desired power**
Z_α = Z score defined by the P-value (*See:* Table 7.2)
Z_β = Z score defined by the statistical power (*See:* Table 7.2)

For example, the Z score for a P-value of 0.05 is equal to 1.96, and the Z score for 80 % power is 0.840. The sum of these values is 1.96 + 0.84 = 2.8, and this quantity squared is 7.8. This calculation has been performed for various common P-value and power combinations; the results of these calculations are shown in Table 7.2.

It is critical to note that Eq. 7.3 assumes that groups 1–2 are of equal size. Such a scenario is fairly uncommon, however. Therefore, one may need to adjust the value *n* to account for unequal group sizes. After using Eq. 7.3, one can employ Eq. 7.5 to execute the adjustment:

Table 7.2 Values for c_{pp} based on common P-value and statistical power Z scores

		Power (1-β)			
		80 %	90 %	95 %	99 %
P-value (α)	Z scores	$Z_{\beta,80}=0.840$	$Z_{\beta,90}=1.282$	$Z_{\beta,95}=1.645$	$Z_{\beta,99}=2.326$
0.05	$Z_{\alpha,0.05}=1.960$	7.84	10.51	13.00	18.37
0.01	$Z_{\alpha,0.01}=2.576$	11.67	14.88	17.82	24.03

$$n' = \frac{n(1+k)^2}{4k} \tag{7.5}$$

Where:

n′ = **calculated sample size with adjustment for unequal group sizes**

n = calculated sample size assuming equal sample size (i.e., unadjusted)

k = the ratio of planned sample sizes of the two groups, where the larger group's size is divided by the smaller group's size.

Equation 7.5 can be used to adjust for unequal group sizes in other sample size calculations, not just for the comparison of two proportions.

As an example, imagine a study in which one is comparing the prevalence estimates of ovarian cancer in women aged 45–50 years versus 70–75 years. Significance was set at $P<0.05$, and power of 90 % is considered acceptable for this study. For this P-value and this power, the correct value for c_{pp} is 10.51 (Eq. 7.4 and Table 7.2). Based on previous studies, it is hypothesized that the prevalence of ovarian cancer will be 1 % in the younger age group and 4 % in the older age group. Using this information and Eq. 7.3, N is calculated to be 564 people per group.

This value assumes that one desires groups of equal size. However, if one anticipates or desires different group sizes, the value 564 must be adjusted to account for unequal group sizes. If your study will involve 3-times as many women in the younger age category than in the older age category (as the prevalence of ovarian cancer in women aged 45–50 years is much lower), then the ratio k will be $3/1=3$. Using this value and $n=564$ (the value to be adjusted), the total sample size for the entire study, n', will be 752. This value can be usefully rounded to 800 participants in total. Let us assume that only women in the 45–50 and 70–75-year-old age groups will be enrolled in the study. Since one plans to enroll 3-times as many women in the younger age group than the older age group, the 800 total participants will be composed of $800 \div (3+1)=200$ women aged 70–75 years and $800–200=600$ women aged 70–75 years ($600 \div 200 = k = 3$).

7.4.4 Sample Size Calculation When Comparing Means

When the outcome parameter is the difference between two means, the following equation can be used to calculate the sample size:

$$n = \frac{c_{pp}\left(\sigma_1^2 + \sigma_2^2\right)}{D^2} \tag{7.6}$$

Where:
n = **Sample size for estimating the difference between two means**
σ = expected standard deviation of the mean difference
c_{pp} = constant defined by the selected P-value and desired power (Table 7.2)
D = expected minimum difference between the means

For example, consider a study in which one wishes to compare the magnitude of weight loss in ovarian cancer patients being treated with regimen A or regimen B. Both groups are expected to lose weight on average, however, one hypothesizes that regimen A will be associated with greater weight loss than B. A pilot study allowed the investigator to predict the standard deviations of the weight loss for A to be 7 kg (σ_1) and for B to be 9 kg (σ_2). The investigator considers a minimum difference in weight loss of 3.0 kg (D = 3.0) to be clinically important. A power of 95 % and P-value of 0.05 were considered adequate for this study; using these parameters and Eq. 7.4, c_{pp} was determined to be 13.00 (Table 7.2). These pieces of information can be plugged into Eq. 7.6 to calculate a total sample size of n = 188. As discussed in the previous sub-section, Eq. 7.5 can be used to adjust this result to account for unequal group sizes.

7.5 The Size of an Observational Etiognostic Study

The sample size calculations discussed thus far relate to fairly straightforward, common scenarios in epidemiology, the estimation or comparison of proportions or means. Yet many investigators wish to address questions about etiology, or the causal factors that contribute to a health phenomenon. In this section we discuss sample size calculations in two typical etiognostic research scenarios, the traditional cohort study and the traditional case–control study. Such studies require additional considerations based on specific details of the study design. For example, choosing to contrast more levels of a determinant, to study a larger number of causal co-factors, or to study more effect modifiers will tend to increase the required sample size beyond what sample size calculations suggest. On the other hand, making a strategic decision to use a more sensitive measure that does not compromise specificity will tend to reduce the required sample size (Miettinen 1985). Study design factors will need to be considered on a case-by-case basis and their implications for sample size may need to be addressed.

7.5.1 Sample Size Calculation for a Traditional Case–Control Study

In case–control approaches the following formula is often found helpful for developing an argumentation about study size:

$$n = \frac{2\left[Z_{\beta}\sqrt{p_1(1-p_1)+p_2(1-p_2)}+Z_{\alpha}\sqrt{2p_{ave}(1-p_{ave})}\right]^2}{(p_2-p_1)^2} \tag{7.7}$$

Where:

n = **sample size of a case–control study**

Z_α = Z score for the desired level of significance

Z_β = Z score for the desired power

p_1 = expected proportion of exposure among controls, as derived from p_2 and an anticipated odds ratio (*See:* Eq. 7.8)

p_2 = expected proportion of exposure among cases

p_{ave} = average of p_1 and p_2

To use this equation, one must establish an anticipated odds ratio (OR). This quantity can sometimes be based on knowledge of the strength of association for other risk factors, but ultimately, the anticipated OR should be driven primarily by the hypothesis being tested. A second piece of information that must be obtained is an expected proportion of exposure among cases (p_2). The value p_2 can often be anticipated using external survey data, employing pilot studies, or locating relevant literature. These two pieces of information, the anticipated OR and p_2, can be plugged into the following equation to compute p_1:

$$p_2 = \frac{p_1(OR)}{1+p_1(OR-1)} \tag{7.8}$$

As an example, consider a case–control study aimed at investigating whether chronic chewing of smokeless tobacco is associated with increased odds of developing any form of mouth cancer. Patients with mouth cancer (cases) and without mouth cancer (controls) are recruited to the study. A pilot study allowed the investigator to estimate that 24 % of cases will have a history of chronic chewing of smokeless tobacco ($p_2 = 0.24$). The investigator anticipates that the OR of having a history of chronic chewing of smokeless tobacco is 6. The values $p_2 = 0.24$ and OR = 6 can be plugged into Eq. 7.8 to determine that $p_1 = 0.05$. In other words, this investigator predicts that 24.0 % of cases and 5 % of controls will have a history of chronic chewing of smokeless tobacco. The average of p1 and p2 is equal to 0.145 (p_{ave}). The investigator has set the level of significance at 0.05 and desires to achieve a power of 90 %; the corresponding Z scores are $Z_\alpha = 1.96$ and $Z_\beta = 1.282$ (Table 7.2). With all of these pieces of information at hand, it is possible to calculate that this study will likely need to include N = 141 participants in total, a value that can be usefully rounded up to 150 total participants. Assuming that this case–control study uses a fairly common ratio of 4 controls for each case, this study should include approximately 30 cases and 120 controls (based on Eq. 7.5).

7.5.2 Sample Size Calculation for a Traditional Cohort Study

To develop an argument around study size for a traditional (independent) cohort study, a slightly more complicated calculation can be useful:

$$n = \frac{\left[Z_\alpha \sqrt{\left(1 + \frac{1}{m}\right)\bar{p}\left(1 - \bar{p}\right)} + Z_\beta \sqrt{\left[\frac{p_1\left(1 - p_1\right)}{m}\right] + p_2\left(1 - p_2\right)} \right]^2}{\left(p_1 - p_2\right)^2} \tag{7.9}$$

Where:

n = total **sample size of a cohort study**

Z_α = Z score for the desired level of significance

Z_β = Z score for the desired power

m = the number of unexposed participants per exposed participants

p_1 = the probability of event in unexposed participants

p_2 = the probability of event in exposed participants

$\bar{p} = (mp_1 + p_2) / (m + 1)$

To illustrate this equation, let us consider a study in which an investigator aims to determine whether obese adults are more likely to develop colon cancer than are non-obese adults. The investigator has set the level of significance at 0.05 and desires to achieve a power of 80 %. The investigator hypothesizes that obese participants will have a twofold increase in the risk of developing colon cancer compared to non-obese participants (i.e., a relative risk or RR = 2). Since $RR = 2 = p_2/p_1$, the probability of developing esophageal cancer in one group (either non-obese or obese) is sufficient to predict the probability of developing the disease in the other group. Assuming that the study will last for 5 years and that the probability of developing colon cancer in non-obese participants during that time is estimated (based on previous work by other researchers) to be 5 % ($p_1 = 0.05$), the expected probability of developing colon cancer in the obese group is 10 % ($p_2 = 0.10$). The investigator plans to enroll three non-obese participants per obese participant (m = 3). Knowing p1, p2, and m, it is possible to calculate the value $\overline{p}$ = 0.0625. With all of this information on hand, the investigator executes Eq. 7.9 and determines she will likely need to enroll at least 271 participants, an estimate that can be usefully rounded up to $n = 300$. Since the investigator plans to enroll 3-times as many non-obese people as obese people, she used Eq. 7.5 to determine that her study should include approximately 75 obese participants and 225 non-obese participants.

There is an important caveat, however, that must be considered in cohort studies on the existence of relatively minor effects. In such cases, sample size calculations tend to produce under-estimates. One way to address this issue is to increase the number of exposed participants, though this approach requires more resources. A more efficient approach is to over-represent extreme degrees of exposure in the exposed group. For example, of the 75 obese participants in the example above, it may be wise to include more severely obese participants than might be enrolled by chance alone. Two critical assumptions of this approach are that there is a dose-dependent relationship between the exposure and the outcome, and that all potential confounding exposures in obese and severely obese participants are similar. If such an approach is taken, it should be clearly reported in the methods section, and these assumptions should be tested with their results disclosed in detail in the results section.

7.6 The Size of an Intervention Study

Study size gains elevated importance in intervention studies, such as randomized controlled trials (RCTs), because the risks associated with intervention studies are generally greater than diagnostic or etiognostic studies. Any intervention poses some degree of risk to the participants; therefore, over-enrollment could expose an unnecessarily large number of participants to a potentially harmful intervention when fewer participants would have been sufficient. In other words, in intervention studies, sample size becomes a major ethical issue, where the main concern is to balance the needs for attaining useful results and for limiting potential harm.

The degree of importance of sample size in an intervention study is directly proportional to the riskiness of that study and is informed by the aims of the study. For example, the optimal size of a Phase-1 clinical trial, the first time a new drug is given to humans for safety and tolerability testing, is very heavily influenced by ethical considerations. Consequently, Phase-1 trials tend to be very small (e.g., n = 6–12). In a Phase-2 study, when a drug's dosing regimen is being evaluated, safety and tolerability are better established but still unclear; therefore, Phase-2 studies tend to be larger than Phase-1 studies but still relatively small (e.g., n = 15–60). The increase in study size in Phase-2 studies often allows preliminary hypothesis testing of efficacy and negative outcomes, though only large effect sizes tend to be detected in such small studies. In a Phase-3 study, on the other hand, the planned study size must be larger than in a Phase-3 study because effect size must be determined. In order to get to Phase-3, safety was previously established in Phases 1–2; therefore, the risk of harm is lower in a Phase-3 study. Consequently, depending on the goals of the study and the predicted effect size, Phase-3 studies can be as small as $n = 200$ and as large as $n = 22{,}000$ (e.g., Physician's Health Study-I).

In any intervention study, it is generally useful to consider ways to increase the efficiency of the study without compromising statistical power. One commonly employed approach is to tweak elements of the study design in order to increase the total number of participants who are likely to experience an outcome of interest. Such tweaks often include: selecting subjects from high-risk populations, lengthening the follow-up period, maximizing compliance, and minimizing drop-out/attrition rates (*See:* Chap. 17). In making such tweaks, one must be very careful to avoid introducing unacceptable bias or ethical errors, and it is critical to report these intentional tweaks in the methods section so others can critically appraise the results.

When calculating sample size for an intervention study, by far the most common approach is to use formulae for the calculation of study size for the comparison of means or proportions (*see:* Eqs. 7.3 and 7.6). Therefore, in this chapter we will not further discuss study size planning formulae for intervention studies. However, there are many comprehensive resources covering a wide range of intervention scenarios. Readers interested in this topic may find useful additional information in statistical textbooks and books on clinical trials, e.g., Meinert (1986).

7.7 Accounting for Attrition

In every study there will be some proportion of participants who withdraw from the study or are otherwise lost to follow-up. The best way to account for these phenomena is to increase study size calculations by a known factor based on previous studies or a pilot study. However, such information is not always available, so a common approach is to round up to a useful study size (e.g., n = 271 can be usefully rounded up to n = 300). Although this approach provides some leeway, some researchers have advocated for a simple further adjustment: to add an additional 10 % (e.g., n = 271 is rounded to n = 300 and 10 % is added to make n = 330). Though such accounting is helpful for planning and budgeting studies, it should be noted that there is no standardized approach to dealing with this issue. Indeed, there is a great deal of controversy regarding approaches to account for attrition. In order to limit potential criticisms, an important task for writing successful grants or other funding requests, we recommend making adjustments to sample size based on pilot studies or literature, and resorting to the 10 % add-on approach if no such pilots or literature are available.

An ideal scenario is to plan the size and procedures of a study from a scientific point-of-view only and, consequently, to aim for very high power/precision and to employ only the most accurate, sophisticated procedures (which are often the most expensive). However, in practice, such an ideal scenario is quite rare in part because stakeholders typically put some level of restriction on such ambitions. It is therefore quite evident that interactions with stakeholders, a topic that is discussed in the next chapter, are crucial if one wants to develop a realistic and ethical plan for a study.

References

Altman DG et al (2008) Statistics with confidence, 2nd edn. BMJ Books, London, pp 1–240. ISBN 9780727913753

Bristol DR (1989) Sample sizes for constructing confidence intervals and testing hypotheses. Stat Med 8:803–811

Burton BR et al (2009) Size matters: just how big is BIG? Int J Epidemiol 38:263–273

Daly LE (2008) Confidence intervals and sample sizes. In: Statistics with confidence, 2nd edn. BMJ Books, London, pp 139–152. ISBN 9780727913753

Edwards BJ et al (2005) Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. BMC Genet 6:18

Goodman SN, Berlin JA (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 121:200–206

Greenland S (1988) On sample size and power calculations for studies using confidence intervals. Am J Epidemiol 128:231–237

Hilden J (1998) Book review of Lang TA and Secic M: ‘How to report statistics in medicine. Annotated guidelines for authors, editors and reviewers’. Med Decis Making 18:351–352

Kirkwood BR, Sterne JAC (2003) Essential medical statistics, 2nd edn. Blackwell Science, Malden, pp 1–501. ISBN 9780865428713

Meinert CL (1986) Clinical trials. Design, conduct and analysis. Oxford University Press, Oxford, pp 1–469. ISBN 0195035682

Miettinen OS (1985) Theoretical epidemiology. Delmar, New York, pp 1–359. ISBN 0827343132

Snedecor GW, Cochran WG (1980) Statistical methods, 7th edn. The Iowa State University Press, Ames, pp 1–507. ISBN 0813815606

Funding and Stakeholder Involvement

8

Jan Van den Broeck and Jonathan R. Brestoff

Donors don't give to institutions. They invest in ideas and people in whom they believe.

G.T. Smith

Abstract

This chapter clarifies the concept of research sponsorship and provides general advice for seeking and applying for research grants. Funding bodies and other research sponsors represent an important group of stakeholders, and their involvement and roles in research projects are discussed here as well as in Chap. 30 (Dissemination to Stakeholders). Sponsors and research institutions carry ethical obligations that are directly relevant not only to society but also to the researchers applying for funding. Some of these obligations are discussed here. After identifying funding bodies and sponsors, submitting a grant application typically initiates a variable and competitive review process; knowledge of a funding body's review process is useful in constructing successful grant applications. An ongoing process after the receipt of funding – grant management – is critical to achieving the grant's specific aims and to completing studies within resource constraints, and practical advice on grant management is accordingly provided.

J. Van den Broeck, M.D., Ph.D. (✉)
Faculty of Medicine and Dentistry, Centre for International Health,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_8,

8.1 Sponsors and Other Stakeholders of Research

Sponsors and other stakeholders of research play a pivotal role at all stages of the epidemiological research process. In this section we introduce the concepts of 'stakeholdership' and sponsorship and we zoom in on sponsor roles in general and on funding of research. Selected terms and concepts relevant to these topics are found in Panel 8.1.

8.1.1 Stakeholders of Research

Stakeholders are all individuals, institutions, or communities who are interested or can be affected by a research project. This implies that epidemiological research has many stakeholders at many levels, from individuals to society-at-large. Pivotal stakeholders are the research participants, the investigators themselves, and the funding bodies supporting the research project. Table 8.1 lists some categories of stakeholders and the types of research studies they often support.

8.1.2 Sponsors of Research

A research sponsor is an individual, company, institution, or organization taking responsibility for the initiation, management, and/or financing of a research study.

Panel 8.1 Selected Terms and Concepts Relevant to Funding and Stakeholder Involvement

Funding (of a research project) (Provision of) availability of financial resources in support of a research project

Grant management Administrative management of the use of a research grant

Grant proposal Research proposal document submitted to a funder of research in view of the obtainment of a research grant

Peer review (of a grant proposal) A check, by scientists knowledgeable of the type of content at issue, of the scientific soundness, feasibility and acceptability of the grant proposal

Research grant An amount of money allocated to a specific research study by a funder of research (variably accompanied by other forms of support)

Sponsor An individual, company, institution, or organization that takes responsibility for the initiation, management, and/or financing of a study

Stakeholders (of a research study) Persons, institutions or communities who have an interest in a research study or can be affected by its activities or by the study results

Table 8.1 Some categories of stakeholders and types of research they often sponsor

Stakeholder category	Examples of types of studies frequently supported
Research institutions, academia, individual investigators	All types of studies
Public health authorities, government	Surveys
	Surveillance studies; outbreak investigations
	Forecasting studies
	Community intervention studies
	Cost-effectiveness studies
Industry	Intervention studies
	Methods-oriented studies focusing on performance of new devices
Patient advocacy groups	Studies on the particular illness of interest to the group
Public interest groups, community organizations	Studies on community diagnosis, etiognosis, prognosis
Charitable individuals (Volunteers, Maecenas), **foundations, trusts**	Studies on particular illnesses or public health problems of interest to the charitable entity

This definition implies that sponsors can facilitate research in several ways, including sponsorship without financial support. It also means that for any given research study there can be several sponsors, each taking on one or more roles. Their sponsorship may include:

- Funding
- Protocol development
- Organization of scientific and ethical oversight and support
- Facilitation of recruitment, sampling, and enrollment
- Training, quality assurance, and quality control support
- Technical assistance with devices, instruments, drugs, etc.
- Dissemination of results

The scope of responsibilities of a sponsor can be very wide and nowadays encompasses setting research priorities (Textbox 8.1). For investigators and grant applicants it is very important to have a good insight into what specific responsibilities a sponsor is willing and able to take. For example, many sponsors will not directly engage in infrastructure building (unless the grant is specifically for this purpose, such as a Construction Grant), and some sponsors provide guidance on study design and protocol development. Clarity about mutual responsibilities helps facilitating contacts between investigators and sponsors. Researchers should ask about the sponsor's expectations about these and other issues:

- Time commitment of the investigators
- Frequency and content of progress and budget reports
- Frequency and content of any data and safety monitoring reports; and collaboration with study monitoring activities
- Which Good Clinical Practice guidelines should be followed

Textbox 8.1 Sponsor Responsibilities in Setting Research Priorities

Sponsors of research, especially funders, are increasingly involved in setting research priorities

Research priority setting needs to be based on existing evidence about illness burdens and knowledge gaps. It needs to involve intensive communication between expert scientists and policy experts. Sponsors and funders of research all over the world should give consideration to international ethical issues, such as equality in access to research on a global scale (e.g., requirements to publish papers in an open-access manner), fair distribution of research burdens, possibilities for research into specific health issues of developing countries. Similarly, national sponsors should pay attention to equality/fairness and special research issues within a country, and the same applies to other levels. Viergever et al. (2010) provide a checklist with advice for good practice in these matters. *See also:* Tomlinson et al. (2011).

- Issues of conflict of interest and intellectual property (*See:* Chap. 31)
- Issues of contractual agreements, e.g., about publications

Researchers often appeal to multiple stakeholders to take on some sponsorship role, especially that of funding. Some sponsors will fund studies when approached in the right way; thus, fundraising skills are important. Strategies for fundraising include:

- Seeking project grants
- Seeking core grants and program grants
- Partnerships with the private sector
- Investment in research (often for-profit companies)
- Capital fundraising (often for-profit companies)

Only the seeking of project grants is further discussed in this chapter because of its particularly high relevance to many individual epidemiologists.

8.2 Project Grant Seeking

Project grant seeking is the main form of fundraising by most investigators and their institutions.

8.2.1 General Principles of Grant Seeking

Some sponsors/funders are open to novel research ideas and designs and will consider grant proposals within the scope of their mission. Others have strictly defined, specific research areas or even specific research questions that they are interested in only. It is essential that investigators-applicants are well aware of such specifications. For example, (1) the United States National Institutes of Health (NIH) regularly call for proposals in specific health domains, and (2) pharmaceutical companies often fund research on topics with clear commercial prospects. Internally, sponsors tend to allocate their funding budgets over periods of 1–5 years to keep pace with changing priorities. In support of budgeting processes, some sponsors involve members of the scientific community or other stakeholders in outlining their specific areas of interest.

General rules of scientific writing (*See:* Chap. 28) apply to grant proposals in principle, but in practice applicants are expected to abide by the rules, guidelines, and preferences of the specific funder (even with respect to preferred concepts, terminology, and style). Funders only invest in ideas and people whom they believe will further their own goals and who have followed their preferred procedures from the earliest contact. In principle, key to an applicant's pecuniary success in grant seeking is the ability to induce grant proposal reviewers to have a perception of scientific rigor, feasibility, innovation, and potential for having a large impact. Grants are obtained when the sponsor believes in the principal investigator, research team, and institution. If that belief is reinforced by successful completion of a project, it may become easier to obtain future funding.

8.2.2 Project Grant Seeking as a Process

A basic introduction to grant seeking is found in Devine (2009). The author describes the process of grant seeking from the perspective of the investigator-applicant. A synopsis (slightly adapted) of the key steps follows:

- First, one develops a pre-proposal or a typical proposal (*See:* below) and discusses it with potential collaborators, institutions, and non-funding stakeholders
- Having become familiar with the current interests and granting methods of relevant sponsors, one shortlists sponsors whose interests and methods fit best with the specific research question addressed in the (pre-)proposal
- Usually one first creates a revised version of the (pre-)proposal that maximizes the fit with each shortlisted sponsor; it is important to discuss this revised (pre-) proposal with scientific collaborators and institutions involved
- Next, one contacts one or more potential sponsors-funders at an appropriate time with respect to the sponsors' 'funding cycles.' Many sponsors use triage methods such that a full proposal is not required initially but only at the final stage
- To develop a full grant proposal (if required), one should carefully follow the sponsor-specific guidelines
- The research institution(s) of the investigator should provide administrative review of the grant application. By authorizing to send a proposal to a potential sponsor, the institution is confirming that, in it's best estimation:
  - The research project can be performed at the proposed funding level
  - Any unique policies of the institution have been considered
  - The proposal meets the requirements of the potential sponsor
  - The institution will comply with all legal requirements
  - The investigator will adhere to the institutional policies

8.2.3 Developing a Project Grant Proposal

As highlighted by Schroter, Groves, and Højgaard (2010), currently there are no uniform requirements for the format of grant proposals, although a large proportion of funders would be sympathetic to that idea. The particular instructions of each sponsor must be followed. Yet, there are elements that tend to be common to all, allowing for the construction of a typical detailed grant proposal structure (Panel 8.2). This structure is useful in the early stages of proposal development and can be easily adapted to follow a sponsor's particular instructions (usually available on the sponsor's website). In addition, one can use a 'model proposal,' i.e., a proposal that was previously successful. Someone in the team has to accept the main responsibility of proposal development, and consequently, must drive that process.

Writing a grant proposal is very similar to writing a scientific paper albeit with other emphases and without results and discussion sections. Much of the advice that

Panel 8.2 Typical Structure of a Detailed Project Grant Proposal

1. Title and summary
2. Investigators, collaborating institutions, and research capacity
3. Table of contents
4. General objectives and specific aims
5. Background and significance
6. Preliminary studies
7. Research design and methods
   (a) Study design: study area, target population, general design, interventions
   (b) Outcome measures, case-definitions
   (c) Study eligibility
   (d) Study procedures
   (e) Laboratory methods
   (f) Sample size and power
   (g) Data management
   (h) Data analysis
   (i) Time frame
   (j) Possible pitfalls and alternative strategies
8. Ethical issues
9. Literature cited
10. Budget and budget justification
11. Possible addenda
   (a) Curriculum vitae of investigators
   (b) Letters of institutional support
   (c) Letters of support from other stakeholders
   (d) Questionnaire drafts
   (e) Standard operating procedures

Textbox 8.2 Key Sections of a Grant Proposal

The soundness of the **specific aims section** is seen by most reviewers of study proposals and grant proposals as an important sign of scientific quality. Adherence to the principles and guidelines of general study design (*See:* Chaps. 5 and 6) must be apparent in the description of the specific aims as well as throughout the **section on research design and methods**. If there are several specific aims, then the following should normally be briefly *described separately for each*: study attributes of interest and their proposed relations (outcomes, determinants, effect modifiers, and confounders), sample size/ power descriptions, study variables (measurements, case definitions), outcome parameters, and analysis plan.

will be given in Chap. 28 (Scientific Writing) applies. For example, to overcome writer's block, one can simply start by typing out the structure of the document with necessary sections and sub-sections. Since the process of proposal development can be very complex, we recommend that those new to the process or looking for new approaches begin writing their proposals by adding bullet points under each subsection. This approach is efficient and useful because the proposal development process is non-linear; ideas relevant to different sections are often inspired non-sequentially. The bullet points then can be progressively developed into text and new ones added as ideas come up.

The title and the summary are crucial elements of the proposal because they play a key role in most sponsors' review processes (Bordage and Dawson 2003). The summary is written last and focuses mostly on reflecting specific aims, research design, and methods. Literature evidence on the topic is concentrated in the section on background and significance, although literature on methodological issues is useful to cite (if available) in the sections on research design and methods. A systematic approach to reviewing the existing evidence is typically viewed favorably (*See:* Chap. 25, Sect. 25.1 on Systematic Literature Reviews). All scientific statements must be correct and referenced; inaccuracies and incorrect interpretations strongly work against the proposal and may themselves lead to poor reception by reviewers. It may be useful to include a list of abbreviations and a glossary of technical terms. The section on specific aims (object design) is one of the most important sections of a grant proposal (Textbox 8.2).

8.2.3.1 Feasibility Arguments

Proposals do not typically have a section dedicated to feasibility, but one can embed feasibility arguments in various sections of the proposal. For example, Sect. 2 (Investigators, collaborating institutions, and research capacity) and 6 (Preliminary studies) of the grant proposal allow one to put appropriate emphasis on some aspects of feasibility. One can show that the investigator and team are qualified, experienced, and motivated by including the bio-sketches of investigators and collaborators, an overview of previous publications in the field of research, and evidence of serious time commitment. If laboratory measurements are to be used (e.g., enzyme-linked immunosorbence assays or ELISA's), then preliminary data showing that the specific research team can in fact measure the parameter using the stated technique is very helpful. Other aspects of feasibility are documented by letters of support from institutions and other stakeholders and by providing evidence of access to the necessary infrastructure.

In order to show competence it is also important to deliver a proposal with correct syntax and spelling, a consistent format, and a polished scientific writing style expressing coherence and sound logic. Further signals indicating investigator competence may be due attention to data management and ethical issues. In addition, timelines should be realistic and take account of the time necessary for study preparations, training, recruitment, follow-up, analysis, and writing. Timelines should also take into account the expected turn-around time of ethical and grant review. An unrealistic or overly ambitious timeline carries many risks, including potentially being seen as naïve to the tasks required to completing the study. Finally, the budget should be realistic and well justified, and it should be made in consultation with experienced epidemiologists and the home institution's grant officials to increase the likelihood that adequate funding is being requested.

8.2.3.2 Budget Plan and Justification

Failing to request enough financial resources necessary to complete a project can ruin an otherwise strong proposal. A well thought-out, realistic financial plan is important to ensure that the project will be sufficiently funded. Concerns over cost reduction should not compromise the validity of a proposed study; in fact, it is good practice for budget item justifications to repeatedly refer to their contributions to study validity. Even if the potential funder only requires a global budget, the provided budget must be based on a careful costing exercise for the entire study. If cost estimates are very high and are felt to compromise the likelihood of being funded, it may be worth approaching multiple funding bodies (an approach that should be made clear in the application and perhaps in advance of formal submission to notify all funding bodies) and/or considering whether some elements of the study are, in fact, unnecessary and can therefore be cut.

Budget items may fall into the broad categories listed in Panel 8.3. This list should not be considered all-inclusive, as the range of possibilities is wide.

One should check which budget items are allowable by the sponsor and be exhaustive in listing justifiable budget items. The financial resources needed are obviously very study-specific and the nature of budget items may be quite different according to the type of study. In intervention studies, for example, the budgeting must be inspired by cost estimations around the use of drugs, devices, and other treatments. Consideration should be given to their purchase/donation, shipment, storage, stocks, administration to participants, adherence assessment, side-effects management, and adverse events monitoring and reporting. It is worth noting that some items are easily forgotten in budget plans, examples of which include:

- Costs of data management (including data cleaning) and quality control
- Costs associated with ethical review and scientific oversight

Panel 8.3 Budget Item Categories Often Used in Budget Specifications (Exact Categories and Terms to Be Used in a Grant Application are Sponsor-Specific)

- Overhead costs (Indirect costs; this is dependent on the home institution)
- Direct costs
  - Personnel-related costs
  - Procurement of Research and Development services
  - Equipment purchase
    Transport-related
    Communication-related
    Measurement procedures-related e.g., for lab analyses
    Intervention-related
  - Operating costs
    Travel and dissemination
    Training
    Space rental
    Miscellaneous costs

- Costs associated with systematic reviews and meta-analyses, e.g., staff salary, internet and library access, article purchase, copying and printing costs, and translation of foreign language articles
- Costs of grant management

8.3 Reviewing a Project Grant Proposal

8.3.1 Internal Review of Grant Proposals

Extensive internal review is required before sending a grant proposal to sponsors. When reviewing, in addition to verifying completeness of the content officially required by the sponsor, use can be made of the checklist of areas of concern listed in Panel 8.4.

8.3.2 External Peer Review of Grant Proposals

A newly submitted full project grant proposal will first be assessed internally by the funder to check for completeness and for compatibility with the sponsor's interests. Any missing elements (even minor ones) in the submission may lead to immediate rejection. Many sponsors then use external reviewers to assess the project grant proposals that survived the initial screening. The reviews are currently based on funder-specific guidelines. Usually three or more grant reviewers are involved.

Panel 8.4 Checklist for Reviewing the Quality of a Grant Proposal

- Informative title
- Sufficient and convincing abstract
- Clearly stated specific aims
- Scholarly, pertinent background and rationale
- Appropriate referencing and use of citations
- Relevant prior work, pilots, expertise
- Sufficient space, human resources, time, and commitment
- Appropriate target population and sampling strategy
- Efficient recruitment methods; realistic projected enrollment and attrition rates
- Accurate and precise measurements
- Detailed quality control plan
- Detailed data management plan
- Adequate sample size and/or power
- Scientifically sound analysis plan for each specific aim
- Ethical issues addressed: oversight, consent, confidentiality, privacy, safety, fairness
- Tight realistic budget without compromising quality
- Realistic timetable
- Clear, concise, well-organized document
- Helpful table of contents and subheadings
- Good schematic diagrams and tables
- Neat and free of errors

Based on the reviewers' reports proposals are ranked and a shortlist is made of proposals recommended for funding. The recommendation is made to a board that will make the final choices. Some sponsors organize review committee meetings to rank the proposals before recommendation to the board.

8.3.2.1 Why Project Grant Proposals Fail

There are tree main groups of reasons why grant proposals fail to be successful. The first group of reasons has to do with the interests and budget of the sponsor; the second with quality characteristics of the grant proposal itself; and the third with quality of the grant proposal review system:

1. Failure due to the interests and budget of the sponsor

   At any point in time sponsors have a set of areas of research that interest them most. A high quality grant proposal may fail if the sponsor sees the proposal as unrelated or only tangentially relevant to the main interests of the moment. In addition, sponsors may receive many more good quality grant applications than they are able to fit within their budget of the period.

2. Failure due to proposal-related factors

   Reasons given by grant reviewers in a study of the NIH (Cuca and McLoughlin 1987) were:

   - Questionable or unsuitable methodology
   - Inadequately defined hypothesis (lacking, faulty, diffuse, or unwarranted)
   - Confusing data collection procedures or inappropriate instruments, timing, or conditions
   - Inappropriate composition of the study group or control group
   - Vague or unsophisticated data management and analysis plans that are unlikely to give accurate and clear-cut results
   - Determination that the proposal is unimportant, unimaginative, or unlikely to provide new information or insight
   - Assessment that the principal investigator has inadequate expertise, familiarity with the relevant literature, poor past performance, or insufficient time
   - Inadequate study setting, support staff, lab facilities, access to the appropriate patient population, or collaboration

3. Failure due to reasons related to the peer review process

   When peer review is used to assess grant proposals, as is usually the case, the poor reproducibility of the peer review process (Hartmann and Neidhardt 1990) combined with a small number of reviewers per grant proposal, may lead to misclassification of a good quality grant proposal as mediocre or poor. Occasionally, not all reviewers may be equally well-qualified, or not all of the well-qualified reviewers may have approached the task at hand with the same seriousness. Also, rare examples are known of good quality grant proposals falling victim to sex bias, theft of ideas, cronyism, and bias against less reputable applicants or institutions (Wessely 1998; Groenveld et al. 1975). Another ethical problem around grant reviewing is that, along with the increased competitiveness in grant seeking, such vague criteria as ‘elegance,’ ‘sophistication,’ and ‘innovation’ are now weighing more and more heavily in comparison with the criteria of ‘importance of the targeted knowledge’ and ‘validity of methods.’ The danger is that these vague and ethically less important criteria overshadow the truly important ones. To avoid disappointment, young investigators-applicants do well keeping in mind that important studies, targeting new knowledge that is very needed in public health, nowadays have a reduced chance of being funded if the methods required to validly achieve this knowledge are standard, common, or easy. However, this should not detour one from pursuing genuinely good ideas on important topics if the required methods are ‘simple.’ Rejection of a grant proposal is more common than acceptance, and there may be opportunities in the future to obtain funding for that study.

8.4 Project Grant Management

Institutions and investigators share responsibility for the financial management of research grants. Institutions are often the official receivers of the grant. And grant management requires devoted time and expertise. The project grant manager may be an administrative person employed at the institution who may have more than one project to manage. Within a project, grant managers often take on other administrative management roles than only financial ones. The roles of a grant manager may include any of the following and more:

- Accountancy
- Human resources management
- Stocks and flows management; purchases
- Liaison with sponsor
- Liaison with regulatory authorities
- Liaison with collaborating institutions and other stakeholders
- Harmonize study activities with institutional policies

There is a strong ethical dimension to grant management. Consider, for example, that misuse of research funds (deviation of funds to items not initially budgeted by the sponsor-approved protocol) is unethical. Grant managers help to harmonize study activities with institutional policies on a variety of issues, including:

- Suspected scientific misconduct
- Mentorship of young researchers
- Support for and from students
- Sexual harassment; avoidance of dual relationships within research teams; discrimination
- Intellectual property rights

As pointed out in this chapter, grant proposals explain the main purpose of a study and highlight the specific aims. They also contain information on a range of study-specific issues, such as recruitment, sampling, enrollment, measurements, quality assurance, data handling, and data analysis plans. In the following chapters we discuss the planning of these study-specific aspects. The next chapter deals with methods of securing the most precious of all resources: the research participants themselves!

References

Bordage G, Dawson B (2003) Experimental study design and grant writing in eight steps and 28 questions. Med Educ 37:376–385

Cuca JM, McLoughlin WJ (1987) Why clinical research grant applications fare poorly in review and how to recover. Cancer Invest 5:55–58

Devine EB (2009) The art of obtaining grants. Am J Health Syst Pharm 66:580–587

Groenveld L, Koller N, Mullins N (1975) The advisors of the United States National Science Foundation. Soc Stud Sci 5:343–354

Hartmann I, Neidhardt F (1990) Peer review at the Deutsche Forschungsgemeinschaft. Scientometrics 19:419–425

Schroter S, Groves T, Højgaard L (2010) Surveys of current status in biomedical science grant review: funding organizations and grant reviewers' perspectives. BMC Med 8:62

Tomlinson M et al (2011) A review of selected research priority setting processes at national level in low and middle income countries: towards fair and legitimate priority setting. Health Res Policy Syst 9: 19 http://www.health-policy-systems.com/content/9/1/19. Accessed Sept 2012

Viergever RF et al (2010) A checklist for health research priority setting: nine common themes of good practice. Health Res Policy Syst 8:36

Wessely S (1998) Peer review of grant application: what do we know? Lancet 352:301–305

The Recruitment, Sampling, and Enrollment Plan

9

Jan Van den Broeck, Ingvild Fossgard Sandøy, and Jonathan R. Brestoff

Research has shown that people who volunteer often live longer.

A. Klein

Abstract

In the previous chapters we discussed specifications of research questions, general study designs, and study size. The next step in developing a study proposal is to create a practical plan for how to find and enroll participants or other observation units. It is important to be clear about what particular characteristics are needed (inclusion and exclusion criteria), how to identify an appropriate number of participants (recruitment, sampling, and eligibility screening), and how to get the necessary permissions to access secondary data or to obtain new information after informed consent and enrollment. The principles and guidelines for each of these tasks are described in this chapter (terminology in Panel 9.1), except that we devote a separate chapter (Chap. 16) to the management of the informed consent process.

9.1 Defining the Study Population

Earlier in the development of a study plan, one defines the study domain (target population) and one chooses a type of study base (cohort, dynamic population or population cross-section; prospective, retrospective, etc.…) and a study size

J. Van den Broeck, M.D., Ph.D. (✉) • I.F. Sandøy, M.D., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no; Ingvild.Sandoy@cih.uib.no

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_9,

Panel 9.1 Selected Terms and Concepts Related to Recruitment, Sampling and Enrolment

Cases Individuals who have the outcome of interest

Controls Individuals who are members of a reference or comparison group

Eligibility screening Checking potential participants' characteristics against the inclusion and exclusion criteria

Enrolment (1) (– procedure) Interactive process composed of sampling, eligibility screening and informed voluntary consent, intended to lead to actual study participation (2) (– act) Actual inclusion as participant

Informed consent process Process of fully informing potential study subjects about the study and of obtaining their voluntary agreement to participate or to continue participation

Recruitment (1) Study activity of informing potential participants and their communities about general features of a study to enhance enrollment (2) Sometimes used as synonym for enrollment

Refusal rate Rate of non-participation among eligible observation units invited to participate

Sampling Process of identifying and establishing access to potentially eligible observation units or to existing information about them

Sampling fraction Sample size divided by population size

Selection bias Bias in the statistical study result caused by problems of selection or retention of study participants

(*See:* Chaps. 5, 6 and 7). Another step in the planning consists of selecting and describing the actual study population/sample; that is, the individuals whose collective experiences will serve as the study base. The definition of the study domain has a great influence in defining the study population; however, in every study, one must further specify characteristics of the study population. Such specifications usually involve a clear description of the study area and setting as well as of practical inclusion and exclusion criteria that will be used to identify individuals eligible for enrollment.

9.1.1 Describing Study Area and Setting

A clear description of the study area and study setting (where study activities take place) helps one to properly target recruitment, sampling, and enrollment efforts. By highlighting important characteristics of the study area and setting, one gives study proposal reviewers insight into the contextual factors that may be important for planning tasks and for interpreting study findings. General study area characteristics of usual interest are:

- Socio-economic status profile
- Ethnicity profile

- Urban – rural distribution
- Burden of diseases
- Any information on the population distribution matrix of modifiers and confounders of interest

Specifications of the study setting may concern, among others:

- Clinical and/or community-based study setting, e.g., home visits
- Type, number, and distribution of clinical settings, e.g., all hospitals and health centers in the study area
- Location of the study coordination center
- Justifications for the choices made

9.1.2 The Difference Between Inclusion and Exclusion Criteria

Inclusion criteria and exclusion criteria are sometimes difficult to conceptualize. Imagine a 10-year-long prospective study in which one aims to address whether using oral contraceptive pills is a causal factor in the development of breast cancer. Though men may develop various forms of breast cancer, cases in men are rare. In addition, the likelihood of developing breast cancer is low for, say, a 30-year-old woman. For these reasons, the investigators decide to include in their study only women older than 55-years-old. Which are the inclusion criteria? Which are the exclusion criteria? Indeed, by including only women, one by definition excludes men. And by only enrolling women 55-years-old or older, one by definition excludes women under the age of 55-years-old. It is no surprise that there is considerable confusion about these terms.

What distinguishes inclusion criteria from exclusion criteria? To answer this question, we suggest that there are two fundamental phases in defining the study population. The first phase is an attempt to define observation units that broadly represent the target population. Such criteria are *inclusion criteria*. In other words, inclusion criteria allow us to define a preliminary study population that approximates the target population. In the second phase, one defines characteristics that whittle the preliminary study population to the actual study population. To use sculpture as an analogy, inclusion criteria are like the additive process of molding clay into an approximate shape resembling the form of a sculpture (where the sculpture is the study population), and exclusion criteria are like the subtractive process used to give final form to the sculpture.

Since inclusion criteria broadly reflect the definition of the target population, they often relate to the study area/setting as well as age category, sex, and other features that are definitional to the target population. In the above example, inclusion criteria might be:

- Women living within 75 km of each of four study-affiliated clinical sites
- 55-years-old or older
- No prior history of any form of breast cancer
- No current diagnosis of any form of breast cancer

Exclusion criteria may be very extensive and are usually intended to increase internal validity, by eliminating the influence of a known confounder and/or reducing attrition, increase statistical efficiency or to avoid an ethical issue. Common reasons for exclusion are:

- It is impossible for the subject to have the outcome of interest
- It is impossible to measure the health-related state or event of interest or the exposure(s), e.g., plans to emigrate relatively shortly after enrollment
- The subject has a particular contraindication for the test intervention
- It would be unethical to include a subject with this particular characteristic because of vulnerability-related reasons (*See:* next sub-section)
- The subject is not able to collaborate because of disease or mental disability
- Informed consent was not obtained or is not obtainable, e.g., due to refusal (*See:* Chap. 16)
- The subject has a characteristic which is a rare effect modifier, and exclusion makes the study domain more homogeneous
- The subject may have a characteristic which is a relatively uncommon confounder whose influence can be eliminated by restriction of the study domain
- Another individual from the same family (or household) is already enrolled in the study, a scenario that may complicate the analysis by increasing the level of non-independence of observations

It should be noted that if many exclusion criteria are applied, this may limit generalizability of the findings to other relevant populations.

9.1.3 Ethical Issues Around Inclusion and Exclusion Criteria

General epidemiological principles (Panel 1.1) prescribe respect for autonomy, avoidance of harm, and minimization of burdens, among others. At the stage of recruitment and enrollment one is often faced with the reality that some potential participants are especially vulnerable to coercion, harm, or burdens. Not excluding them could expose them to those risks. However, research on such individuals may be justifiable if they have a particular health issue that needs to be studied. For example, pregnant women are a vulnerable group of people, but the disease pre-eclampsia can only be studied in a sample of pregnant women.

According to the general ethical principle of fairness and justice there should be a fair distribution of the burdens and benefits of research among all layers of society and among societies. The selection of participants in research should be fair, with persons being selected only because of the specific subject area being studied (e.g., pre-eclampsia), and not because of their easy availability or their reduced autonomy.

Vulnerable persons are all those who have:

- Diminished ability to protect their own interests
- Reduced capacity to give informed consent
- Incapacity to understand or communicate
- No position to make a voluntary decision
- Increased risk of harm or an increased burden of participation

Examples of vulnerable persons are given in Panel 9.2.

Special justification is required to invite such persons to participate in research, and the CIOMS Guidelines (Council for International Organizations of Medical

Panel 9.2 Examples of Vulnerable Persons Whose Inclusion Requires Special Justification

- Pregnant women
- Prisoners
- Children
- Fetuses
- Mentally disabled or mentally ill patients
- Terminally or seriously ill patients
- Persons in dependent positions
- Educationally or economically disadvantaged persons
- Persons who are under the influence of drugs or alcohol
- Traumatized individuals

Sciences 2009, 2010) require that additional measures be taken to protect their rights and welfare. The principle of fairness and justice further dictates that the distribution of research burdens and benefits should not be inspired by racial, gender-related, sexual, or cultural considerations. In practice this implies that special justification will be needed if the investigator wishes to restrict the study to one gender, race etc.

9.2 Recruitment Before the First Study Contact

When new information is to be collected, obtaining a sufficient volume of quality data strongly depends on enrollment rates, which, in turn, depend on how well potential participants and their communities are reached and informed about the study. Such outreach efforts can be strongly influenced by recruitment activities that occur before the first study contact.

9.2.1 Overview of Recruitment Strategies

Frequently used recruitment strategies, before sampling and first study contact, are listed in Panel 9.3.

We expand briefly on the use of study information sheets and media coverage because these can be important for the success of the recruitment and enrollment process. For more information on obtaining community consent, we refer to Chap. 16.

9.2.2 Study Information Sheets

Information sheets, flyers, or brochures are often used to raise awareness of the existence or arrival of a study among potential participants and other stakeholders.

Panel 9.3 Examples of Recruitment Strategies Before Sampling and First Study Contact

- Information sheets
- Posters at strategic points
- Media coverage
- Meetings with opinion leaders, traditional leaders, and local authorities; attempting to obtain so-called 'community consent'
- Community information meetings; mobile shows; drama
- Community Advisory Board involvement
- Personal contacts in person or by mail, email, or telephone
- Meetings with health facility staff
- Meetings with neighborhood health committees or community-based organizations
- Patient advocacy groups

They can be useful to inform populations about large upcoming population-based studies (e.g., surveys and surveillance systems) and smaller-scale studies. All information sheets need to be approved by the research ethics committee. For possible content of information sheets, *See:* Panel 9.4.

There are some advantages and disadvantages to the use of study information sheets. Their main advantages are that they:

- Can be part of a strategy to boost enrollment rates
- Allow people to think about and discuss with others the pros and cons of participation
- Potentially avoid situations in which people are taken by surprise when approached for eligibility screening
- Can make approaching people more acceptable
- May avoid some unnecessary screening contacts with non-eligible subjects
- Make the informed consent process easier once subjects are found to be eligible
- Can be of use after enrollment as part of an ongoing informed consent process
- Raise awareness and potentially enhance the reputation and status of the investigators and the research institution
- Are often perceived as a sign of transparency in participant selection

There may be some downsides as well. For instance, some people do not like the necessary shortness and lack of detail in a brief information sheet and may perceive that as a deterrent.

9.2.3 Media Coverage During Recruitment

Local media coverage can be useful whenever maximum participation rates are required in population-based studies. An effective recruitment strategy might be

Panel 9.4 Frequent Concerns About Upcoming Research Projects and Possible Responses for Inclusion in Information Sheets and During Media Coverage

- Are the researchers trustworthy and competent?
  - Provide information on research institution and main investigators
  - For media coverage, introduce yourself before communicating
- Is the topic of research relevant to me and my community?
  - Describe the health problem in lay terms
  - Mention the burden of the problem in the community
  - Mention the importance of the new information that may be obtained
- Are they communicating with me in a respectful manner?
  - When communicating with individuals, use a personalized approach in an appropriate style
  - Express appreciation to prospective participants
- What is in this for me? Will I be part of something big and exciting?
  - Emphasize that participants will contribute to something important
- What will they ask from me if I participate? Is it going to be easy?
  - Make it clear whether there will be an intervention and what kind
  - Provide an idea of timing (start, duration) of participation
  - State whether there will be home visits, clinic visits, biological samples
  - Specify whether there will be several rounds of data collection
- Is it safe to participate?
  - Provide an idea of the general level of risks and discomfort imposed by the study
  - Re-assure that safety protections, confidentiality, anonymity, and privacy will be complied with
  - Give opportunities for questions and discussion by providing a telephone number, a website, and/or an email address
- What do other people think about this project?
  - Mention support from community leaders and opinion leaders
- How many people do they want to participate?
  - Mention targeted sample size

to first publish an article about the upcoming study in the local newspaper and then to insert a copy of this article in the invitation letter or add it to the information sheet.

Communication about a study in the early recruitment phase needs to address concerns that most people have about research. It is, in fact, the same kind of information that will need to be provided later during first study contacts and in the informed consent form, although usually not in as much detail. Some frequent concerns about a new study are listed in Panel 9.4.

9.3 Overview of Sampling Methods

In a broad epidemiological sense, the term 'sampling' refers to the process of facilitating access to a suitable selection of observation units or to the existing information about them. Sampling helps to create opportunities for first contact with potentially eligible individuals or their data. There are two general types of sampling methods: statistical and non-statistical sampling. The former involves generating a list of potentially eligible observation units (i.e., defining a sampling frame) and using a statistical scheme to select from the sampling frame a number of units to be approached for enrollment. Non-statistical sampling does not involve a sampling frame or such statistical schema. Though there is a perception that a study population must be statistically representative of a large population, that idea is a misconception (Miettinen 1985). In fact, statistical sampling methods tend to be restricted to large surveys, cluster-randomized trials, and some etiognostic study types. Indeed, in epidemiology non-statistical sampling methods tend to be suitable for most studies.

9.3.1 Non-statistical Sampling Methods (Non-probability Sampling)

There are many types of non-statistical sampling methods. Perhaps the most commonly used are consecutive sampling, convenience sampling, and snowball sampling. These methods are described in Panel 9.5. These sampling methods are frequently used in cross-sectional studies, observational follow-up studies, and in experimental and quasi-experimental studies. A basic assumption of these methods is that the mix of recruited subjects will be roughly typical of the target population. Each has distinct advantages and disadvantages (e.g., snowball sampling can be useful to recruit individuals who are difficult to reach, such as drug addicts).

Non-statistical sampling methods are sometimes used to achieve a quota of units with defined characteristics. Such quotas are intended to ensure that a sufficient number of units in different exposure or outcome levels are achieved, or to balance a known confounding or effect modifying characteristics across groups. For example, in a study of how ethnicity modifies an outcome parameter, one may sample an equal number of participants from different ethnic groups. This approach is often called *quota sampling*.

9.3.2 Statistical Sampling Methods (Probability Sampling)

These methods, unlike non-statistical sampling methods, use sampling frames. Statistical sampling methods are mostly used in surveys, cluster-randomized trials, and sometimes etiognostic studies. The main goal of statistical sampling is to achieve a study population that is statistically representative of the target population. Statistically, the ideal scenario is to sample a complete target population

Panel 9.5 Common Types of Non-statistical Sampling Methods Used in Epidemiology

- **Consecutive sampling**
  With this method, all eligible subjects are found consecutively. These units can be found sequentially or in regular intervals. For example, the investigators approach every *n*th patient presenting to the emergency room (where *n* is 1 if every patient will be approached). Alternatively, the investigators could approach all patients presenting on every *n*th day (e.g., every Wednesday).
- **Convenience sampling**
  In this method, subjects are approached at the time of data collection. This approach is particularly useful if attempting to recruit subjects in a public location, such as a shopping center. This approach can be used in studies with very broad inclusion criteria, e.g., 'adults.'
- **Snowball sampling**
  Participants are successively recruited through referrals from other participants. For example, in a study on cocaine addiction, one might ask a participant to refer others with cocaine addiction to the study. This approach is particularly useful for patient populations that are difficult to reach.

(100 % sample), as this avoids sampling error and is, by definition, the most representative study population possible. However, complete sampling is practically impossible in almost every conceivable scenario in epidemiology. If a sampling frame exists or can be constituted, statistical sampling methods can be affordable and efficient. They allow us to sample a fraction of the target population (sampling fraction); though smaller sampling fractions introduce error, they can also increase internal validity because data collection may be managed by a smaller team of people. Thus, it may be more feasible to find experienced data collectors and to supervise and pay them properly. Such teams tend to collect more accurate data than an army of less-well-trained temporary staff hired on an extremely tight budget.

A statistical sample can be only as good as the sampling frame (Herold 2008). If the sampling frame is biased, so too will be the sample. Therefore, if there is either certainty or serious suspicion about the lack of quality of an existing sampling frame, the only solution may be to constitute a new sampling frame in preparation for a study. Table 9.1 gives examples of survey sampling frames with expected limitations in relation to representativeness of the target population.

9.3.2.1 Random Sampling with or Without Replacement

With random sampling each member of the sampling frame has a known and fully independent chance of being selected. The preferred way to execute random sampling is:

- To assign a random unique number (generated using a random number function in statistical software or a spreadsheet) to each member of the population,

Table 9.1 Examples of sampling frames and their limitations in relation to representativeness of the target population

Sampling frame	Limitations
Census	A proportion of individuals may never be listed because they were never found at home during the census
	Homeless people or itinerants may be missed
	If the census was not conducted very recently, it may be outdated in areas with substantial in- and outmigration
Taxpayer list	People may try to avoid being listed as a taxpayer
	Only approximately representative if it can be shown that only a small proportion avoids tax
List of postal or email addresses	Mail addresses, business addresses and living addresses are sometimes different
	People may have several mail addresses and several living addresses
	In rural areas or informal settlements houses may not be numbered or have a clear postal address
	Variation in number of subjects per postal address
	People may have several email addresses
	Many people do not have an email address, and these people may be different from those who have email addresses
List of (landline) phone numbers	Decreased probability of inclusion of several types of individuals, such as those who have no landline phone (e.g., those who have cellular phones only, or, those who are too poor to afford any type of phone), those who are never or rarely at home, those whose landline does not function for whatever reason, et cetera
	Variation in number of subjects per landline phone
List from hospital or health center information systems	Sick people only
	Rapidly outdated; Patients may frequently change health care provider
	If lists are obtained from public facilities only, the listed patients may differ from those who seek services at private facilities
List of schools, pupils, villages, employees, or administrative areas	Lists of schools are sometimes only available for the public sector
	Rapidly outdated
List of geo-referenced homesteads; satellite maps showing bounded structures	Not all bounded structures are inhabited
	Variation in number of subjects per homestead

- To rank the sampling frame according to the randomly assigned numbers, and then
- To select the first *n* of the ordered random numbers, where *n* is the required sample size

Other methods, such as the lottery methods and the use of tables of random numbers, are more prone to human error but are useful alternatives in situations where no statistical software package is available.

Each individual has exactly the same probability of being selected in ‘simple random sampling with replacement’ (SRS-WR), i.e., when the sampled individual continues to be part of the sampling frame (thus possibly giving a sample with duplicates). Each individual has *approximately* the same probability of being included in ‘simple random sampling without replacement’ (SRS-WOR) if the sampling frame is very large, and this method is often preferable since duplicates can be effectively avoided.

9.3.2.2 Systematic Sampling

With systematic sampling every *nth* person or unit is selected from the sampling frame, where the selection interval *n* is determined by dividing the size of the sampling frame by the study sample size. The first unit is usually sampled randomly. Systematic sampling with a random starting point is not fully random because the chance of a unit being selected is not independent of the prior unit selected. The likelihood of being sampled is, in fact, dependent on the selected starting point, and this non-randomness comes at a cost. *Starting point bias* can arise if there is a pattern in the characteristics of the sampled units that runs in phase with the sampling interval. For example, this may occur if the sampling frame is the list of consecutive houses in a specific street and every *n*th house is mostly a corner house or a shop.

9.3.2.3 One-Stage Cluster Sampling

When the population is large, widespread, and not completely enumerated, cluster sampling may save time, money, and effort. Rather than engaging in a complete census prior to the study and sampling widely scattered participants after the census, it could be advantageous to randomly select some clusters and then try to enroll all eligible subjects in those clusters. The clusters can be villages, electoral districts, schools, households, any natural grouping of people, or even artificial groupings like grid cells placed over a satellite photograph. The practical advantages of cluster sampling are considerable, as participants in each cluster will usually live relatively close to each other, making them more easily accessible. If all individual members of a selected cluster are visited, one avoids the potential embarrassment, discontent, or stigma created by visiting only certain individuals in close communities. The disadvantage is that there is usually some loss of statistical precision compared to what could have been achieved with SRS with the same number of participants. This is because the variation between individuals from the same cluster is often smaller than the variation between individuals from different clusters. A small number of clusters and a small sampling fraction may lead to poor representation of the target population. This could happen, for example, by sampling less than ten clusters that represent less than half of all the clusters. Larger numbers of clusters or pre-sampling information on cluster heterogeneity for variables of interest may be needed.

Table 9.2 illustrates the essential differences between the main forms of statistical sampling.

Table 9.2 Illustration of random sampling, systematic sampling, and cluster sampling

Type of statistical sampling	Sampling unit: example	Sampling frame: list, sampled units in bold
Random sampling	Individual school children in a region	1, **2**, 3, 4, **5, 6,** 7, **8**, 9, 10, 11, **12, 13, 14**, 15, … *(Individuals are randomly chosen from the list)*
Systematic sampling	Individual school children in a region	1, **2**, 3, **4**, 5, **6**, 7, **8**, 9, **10**, 11, **12**, 13, **14**, 15, … *(Every* n*th individual is chosen from list)*
Cluster sampling	Classes of school children in a region	1, 2, **3, 4**, 5, **6,** 7, 8, 9, **10, 11, 12**, 13, 14, **15**, … *(Classes are randomly chosen from the list; all pupils from the selected classes are invited)*

9.3.2.4 Multi-stage Cluster Sampling

Cluster sampling is done in stages for successively-smaller hierarchically-nested groups within the population until the required observation unit level (usually individuals) is reached. It starts with cluster sampling and can end with random sampling of individuals. For example, in a *two-stage sampling* exercise one may first take a random sample of schools and then take a random sample of children from each school. Multi-stage cluster sampling can also involve several successive cluster sampling steps. For example, a *three-stage sampling* exercise could consist of randomly sampling schools first, classes within each school next, and then pupils within each class.

Clusters may differ in size (e.g., large villages, small villages; large households, small households), so if a fixed *number* of individuals is selected from each cluster, individuals living within a large cluster would have a lower probability of being selected. Weights would need to be applied during analysis to adjust for this. Alternatively, one can apply *self-weighted sampling* (Armitage and Berry 1988), where in the first stage the chance of selecting each particular cluster is proportional to the size of the population within it. The second-stage samples can then have a fixed number without creating bias. Another version of self-weighted sampling would be to select clusters with equal probability and then select a number of individuals from each cluster that is proportional to the size of the cluster.

9.3.3 Additional Aspects of Survey Sampling

9.3.3.1 Stratifications in Survey Sampling

Stratified sampling divides the population into non-overlapping subgroups (strata) according to some important characteristic, such as sex, age category, or socioeconomic status, and selects a sample from each subgroup. The number of individuals sampled from each stratum can be made proportionate or disproportionate to the frequency of the characteristic in the population. Disproportionate stratified sampling is sometimes used to ensure that persons belonging to a less common subgroup or a certain category of a potential modifier are represented in large enough numbers to enable the calculation of precise enough estimates for this subgroup. For example, if old age is a potential modifier for a phenomenon under study, one may decide to disproportionately 'over-sample' the oldest age group to enable the calculation of an adequately precise estimate for that age group. When disproportionate stratified sampling is used, it will still be possible to estimate an overall outcome parameter (e.g., for all ages combined) and achieve a robust standard error by using procedures called weighting (*See:* Chap. 22). Stratified sampling can even reduce the overall sampling error if there is a lot of heterogeneity in outcome parameter estimates between strata (Armitage and Berry 1988). Note that disproportionate statistical sampling is a type of quota sampling (*See:* Sect. 9.3.1)

9.3.3.2 The Use of Subsamples in Surveys: 'Multi-phase Sampling'

In large surveys the amount of information that can be collected on each participant is often limited because of logistical and budgetary constraints. If more detailed information is desired (e.g., plasma lipid profiles), it may be cost-efficient to gather that information only in a nested subsample. The process of defining a nested subsample is known as *multi-phase sampling* and, in its simplest form, involves two phases of random sampling, where the sample frame for the subsample is the entire study sample. The precision of the estimates in the subsample will be less than in the study sample. However, surveys are often designed with large sample sizes to produce sufficiently precise estimates of primary outcomes for several sub-regions, ethnic groups, and age-sex categories. Therefore the size of even a 10 % subsample may be large enough to produce adequately precise estimates of secondary outcomes for the entire target population, perhaps even if stratified on a variable of interest (e.g., sex).

9.3.3.3 Complications Created by Non-enrolments in Surveys

Sampling of individuals creates opportunities for initial contact with potentially eligible individuals. Complications can arise if many of the sampled subjects are not enrolled because of missed contacts (after several attempts), lack of eligibility, or refusal. For example, after a systematic sampling exercise involving visits to every nth house, it may appear that only 90 % of the targeted sample size was reached. In order to find the remaining 10 %, should one continue with a second round of systematic sampling, with the same selection interval but from another starting point? This strategy could create bias as the remaining 10 % of participants would be found mostly in the beginning of the round in a relatively small area not representative of the total area. To avoid this problem a new larger selection interval must be used in the second round. Another solution may be to find an immediate replacement for any missed enrolment, perhaps the nearest eligible person. Alternatively, an anticipated 10 % non-enrolment rate can be taken into account in the calculation of the selection interval *n* for the first round, but this may still result in a slight over- or under-enrollment. Similarly, when simple random sampling or cluster sampling is used, a certain percentage can be added to allow for non-enrollments. To enable evaluation of possible selection biases one should try to collect information on the non-enrolled.

9.4 First Study Contact, Eligibility Screening, and Maximizing Response Rates

First study contacts are made personally, via an invitation letter, email, telephone call, or by a house-visit. During first contacts, the same common concerns listed above in the section on recruitment activities should be kept in mind (*See also:* Textbox 9.1). If the first contact is via a letter or email, the message should be clear, brief, personal, and professional. It should also have an attractive layout, use the header of the institution, and be signed. If the first contact is face-to-face, it is important that the researcher behaves respectfully and complies with culturally acceptable dressing, language, and etiquette. In some cultures this implies greeting and informing the head of household before any other household members. Introductory letters and wearing personal IDs with a picture will usually increase the credibility of and trust in the researchers. In a telephone survey, respectful and culturally appropriate language and tone of voice are important.

In (e-)mail or telephone surveys the response rate strongly depends on the number of attempted contacts, on flexibility and variation of the contact strategy for individual cases, and on whether candidates are given enough time for reflection. Whether there should be multiple contact attempts – and, if so, when and how frequent these should be – is very culturally dependent. A common strategy is to make two or three attempted contacts. An approach that has worked well for mail surveys in the U.S.A. is to start with a pre-notice (a phone call or a letter) followed by mailing of the questionnaire and a cover letter (Dillman 2000). If no response was received, up to three reminders were sent that were slightly different in formulation. In that study setting, inclusion in the mail of small incentives in the form of cash, checks, lottery tickets, or pencils was associated with better response rates. After failing to contact a person by mail one may switch to a telephone- or visit-based strategy, possibly making multiple attempts to phone or visit if necessary.

After a proper introduction and briefing about the study, it is usually natural to ask a few simple and straight-forward questions (e.g., about age and residence) to determine whether an individual is eligible. Eligibility screening is usually conducted before an individual is asked to give informed consent to participate, but if particularly sensitive information is needed to determine eligibility, informed consent should be obtained first.

White and colleagues (2008) provide a good overview of what is known about factors associated with participation rates and selection bias in Anglo-Saxon high-income countries. Their overview suggests, among others, that non-participants tend to be poorer, have an unhealthier lifestyle, and are more likely to be male and non-white. Younger age has also been reported among important factors associated with non-participation (Moorman et al. 1999). However, examples of studies showing the contrary also exist (Galea and Tracy 2007). Anyhow, these factors may have limited relevance for research in other cultural settings and in low- and middle income

Textbox 9.1 Selected Ethical Aspects of First Study Contact and Eligibility Screening

In instances when sampling is done from client registries of care facilities, it is appropriate to have the **list of** statistically sampled **candidate subjects reviewed by the caregivers** before any contact is made with the candidates. This allows exclusion of terminally ill persons, persons with severe mental illness and other persons with characteristics that are exclusion criteria. It may also prevent unnecessary efforts to contact persons who are no longer clients or prevent bothering family members of persons who recently died.

Efforts must be made to ensure that **invitation letters or calls** by themselves do not cause any unwarranted health or confidentiality concerns. Letters, information sheets and other recruitment strategies, informed consent forms, personal introductions by enrollers, and questions and exams related to eligibility, all need to be culturally adapted to the local setting and must express respect, empathy, professional seriousness as well as give reassurance about common concerns. If this is not ensured, enrollment rates are bound to be affected.

Endeavors at maximizing participation by **repeatedly attempting to contact persons** who do not respond to invitation letters or are not available when visits or calls are made must be balanced against the risk that people perceive that their privacy is being invaded. Non-response and unavailability may reflect unwillingness to participate, and in such situations repeated reminders may create antipathy, also among other community members, and thereby impinge on their potential study participation.

In communities, the fact that some persons are visited and others not can lead to **embarrassment and stigma**. This problem can occur more frequently with certain sampling schemes. For this reason, sometimes all community members are indeed visited but detailed information collected only from those required to undertake the study.

It is usually inappropriate to offer **monetary or other incentives** beyond compensation for costs of traveling and time. When making first contact with persons who will be 'cases' in a case–control sampling strategy, offering monetary or other incentives is often perceived as inappropriate or even offensive (Coogan and Rosenberg 2004).

Although **eligibility screening** is usually a non-invasive process, in some studies it **may require invasive procedures** such as blood sampling and generation of sensitive personal information such as HIV status. Informed consent is always needed for this kind of eligibility screening and the informed consent process needs to make it clear that the subject may end up being non-eligible.

countries. More methods-oriented research is needed worldwide on the factors that influence enrollment and refusal rates.

Finally, it is crucial to make a plan for monitoring accrual and refusal rates and for gathering information about reasons for non-participation. These issues will be discussed in detail in Chap. 17 (Accrual, Retention and Adherence).

9.5 Sampling and Enrollment in Cohort Studies

We will now discuss some particularities of sampling and enrollment in etiognostic-type studies. We focus on sampling and enrollment procedures ('selection') for cohort studies in the present section and for case control studies in the next one. For each, we will point out the possible sources of selection bias.

9.5.1 Selection Strategies in Cohort Studies

In cohort studies there are some special issues in relation to inclusion and exclusion criteria. The most notable issue is that subjects should, to the extent possible, be excluded if they are not at risk of the outcome. This concerns those who already have the outcome and those who cannot logically ever develop the outcome. Furthermore, in prospective cohort studies there should be a reasonable possibility for follow-up and repeated assessment of study attributes. Generally, it is better to exclude those who have near-term emigration plans or other characteristics that will likely lead to rapid loss to follow-up.

Two modes of selecting members into a cohort can be distinguished:

- *Cohort selection mode-1*: selection of the exposure groups separately. For example, one may select workers of a factory using a dangerous substance and, separately, workers of another nearby factory where the same substance is not used. Mode-1 is often the preferred mode when the exposure is relatively rare, such as exposure to radiation during pregnancy. Group matching and individual matching for confounding variables (*See:* Chap. 6) can be helpful as part of this approach.
- *Cohort selection mode-2*: the commonly preferred method, consisting of selection of one single group, with consideration of exposure levels during measurement and analysis. For example, the Framingham Study population was enrolled irrespective of their smoking status, and later split up according to smoking habit categories. This mode can be more expensive than mode-1 when the exposure is relatively rare.

With either mode, non-statistical sampling methods are often used for the formation of the cohort. Sometimes a statistical sample is used. For example, a subsample of a survey can be selected for inclusion into a cohort study. Participants of large case–control studies may, under certain conditions, also be used for a subsequent cohort study. When the controls of a case–control study are truly a representative sample of the source population, they may form a natural group of candidates for follow-up in an ensuing cohort study. Strategies have also been described for selecting both the cases and the controls of a case–control study into a subsequent cohort study. An example of this is known as the 'reconstructed population method' (*See:* Sommerfelt et al. 2012).

9.5.2 Selection Bias in Cohort Studies

The purpose of a cohort study is to set up a valid contrast of outcome frequency between exposure levels. This means that one should try to ensure that the exposed and unexposed groups have a comparable prognosis at baseline (i.e., a comparable mean risk of developing the outcome) and, further, one should try to avoid prognostic imbalances arising during follow-up (except those mediated by the exposure). If this cannot be achieved, imbalances in prognostic factors at baseline and during follow-up should be measured and adjusted for during analysis. With cohort selection mode-1 (separate selection of exposure groups) one tries to achieve the ideal baseline prognostic equivalence by carefully selecting the groups and making sure they have similar distributions of confounders, sometimes by using individual matching.

It is not uncommon, though, for a researcher to select the groups to the best of her/his abilities but remain uncertain about or be unaware of some prognostic imbalances. Consider the example of a study in which the outcome frequency among workers in an industrial setting (exposed) is to be compared to the outcome frequency in a group selected from the general population (unexposed). A 'healthy worker effect' can occur if healthy persons with relatively good prognosis are more likely to be employed in the industrial setting or if those at risk of the illness are more likely to stop working or switch to different types of jobs. In this case, the exposed and unexposed would have different baseline prognoses, and it would be unclear how this prognostic imbalance could be measured accurately enough for adequate adjustment in the analysis. Consequently, a biased outcome parameter estimate would be expected. On the other hand, a 'sick worker effect' can occur if the bias resulting from a baseline prognostic imbalance is created by a specific job that attracts people with poorer health prognosis on average, e.g., night watchmen (Miettinen 1985). It is a problem in epidemiology that several types of individual prognostic factors, such as an inclination to follow health advice, a tendency to react poorly to stressful situations, and other susceptibilities to important behaviors are difficult to measure accurately.

Selection bias can also occur through erroneous determinations or assessments of eligibility criteria. For example, in a cohort study comparing the rate of appendicitis among smokers and non-smokers, bias can arise if enrollers neglect to verify appendectomy as a study exclusion criterion (and if this is not adjusted for in the analysis). This example can also be used to illustrate the point that sub-optimal selection processes can contribute to confounding.

Finally, remember that in Chap. 2 (Basic Concepts in Epidemiology), biases resulting from various patterns of loss-to-follow up were also treated as a form of selection bias in cohort studies.

9.6 Sampling and Enrollment in Case–Control Studies

The general design of case–control studies has been discussed in Chap. 6. This included a discussion of the concepts of *source population* and *secondary study base*, both of which are important to keep in mind when reading this section. Here, we expand on practical strategies of sampling and enrollment and highlight common sources of selection bias in case–control studies.

9.6.1 Selection of Cases in Case–Control Studies

In the typical case–control study, the selection of cases and controls constitutes two quite different activities. We therefore discuss them separately, starting with case selection.

9.6.1.1 Incident Versus Prevalent Cases

An important decision to make is whether the study will target prevalent cases or incident cases. The distinction between the two is that incident cases (i.e., new cases) cannot include individuals who manifestly have had the illness for longer than a defined time cut-off, whereas prevalent cases can. When incident cases are selected, the study tends to be less prone to certain types of bias. For example, with long-standing prevalent cases there are more frequently recall problems about the exact nature of the diagnosis, timing of diagnosis, and antecedent exposures. This can be especially problematic when diagnostic and exposure-related information is obtained via interview, e.g., if the identification is based on questions such as ‘have you ever been diagnosed with asthma?’ Note that people with mild chronic conditions may remember symptoms more easily than the correct medical term for their condition.

A separate problem arises if the illness has a high fatality rate. Prevalent cases may then represent a special select group of long-term survivors. And if the exposure under study is a true cause of the development of the illness, it is likely to be also a causal determinant of the course of illness and the outcome. Thus, when prevalent cases are used in such instances, the preponderance of survivors among the cases may under-represent the exposed, which is expected to result in an underestimation of the odds ratio. On the other hand, with incident cases it usually takes longer to get adequate numbers of participants (an efficiency concern).

9.6.1.2 Case Ascertainment and Eligibility Assessment in Case–Control Studies

Selection of cases involves case ascertainment, which requires clear and valid case definitions. Up-to-date accepted diagnostic criteria are preferred as a basis for the case definition. A choice of incident cases or of severe cases will require incorporating extra criteria into the definition of case eligibility. Additional criteria might include accepted grading systems to assess severity and a specific maximum time since first manifestation of illness to distinguish incident from prevalent cases. High sensitivity and specificity of case ascertainment is necessary and the use of proxy variables should be avoided if possible.

9.6.1.3 Sources of Cases in Case–Control Studies

Case Recruitment in the Community

Cases can be identified during surveys. With this approach the identified cases are likely to be representative of all cases, and the source population for the subsequent selection of controls can be clearly defined. However, consider that, although the cases are recruited in the community, referral bias (*See:* next subsection on Case Selection Biases) is still possible. The cases may be identified during home visits by asking the question 'have you ever been diagnosed with illness x'. This illness may be one that is typically diagnosed in a hospital after referral, and this referral may be associated with the exposure. The selected cases could thus be a group with increased exposure odds in comparison with all true cases (some of whom remained undiagnosed).

Cases Identified in Disease Registers

National or regional disease registers can be a useful source of cases, but since cases must have come to diagnostic centers, they may represent a selected group of all cases. Referral bias arises when the cases' inclusion in the register was influenced by whether or not they were exposed. All eligible cases can be included or they can be randomly sampled if that is needed for budgetary purposes.

Case Recruitment in Care Settings

Historically, this has been the most frequently used source of cases in case–control studies. Enrollment activities can be conducted in hospitals, clinics, private practices, or combinations thereof. There are some advantages to this approach, not the least being the ready availability of cases in a setting that may easily allow the use of valid up-to-date diagnostic procedures. If the care settings have well-defined catchment areas and, nearly all cases occurring in these catchment areas are expected to end up in the local facilities, then defining the source population becomes easier. If not, selection of controls truly representing the source population of the cases can be difficult to achieve and demonstrate. Health care utilization surveys can be helpful for this purpose. Such surveys could show, for example, that the initially targeted referral center(s) only catch(es) a minor proportion of cases developing in the surveyed area. This would indicate a need to include more referral centers for case identification, or a need to redefine the catchment area/source population.

A requirement for case recruitment in care centers is that the whole process of referral, case diagnosis and enrollment should be independent of exposure (*See:* Sect. 9.6.3). This requirement is more likely to be fulfilled for *severe* cases. Hence, some epidemiologists have suggested that such case–control studies should be done with severe cases only (e.g., Miettinen 1985). When recruiting cases from care settings, one should preferably target cases from several care settings in the region because risk factors (antecedent exposure) may be unique to a single hospital due to referral patterns and other factors. If one would involve only a tertiary care hospital a problem could be that this hospital has a very large catchment area with a complex referral pattern. This may hamper a clear definition of the source population.

Cases Recruited by Snowball Sampling

This approach tends to involve identification of some cases in care settings or surveys, followed by the identification of additional cases via snowball sampling. This type of case recruitment has been used mainly when eligible persons are difficult to reach, such as intravenous drug users. A limitation to this approach, however, is that defining the source population of these cases can be particularly challenging.

Cases Developing During Follow-Up of Well-Defined Cohorts or Dynamic Populations

In traditional nested case–control studies, the cases are usually all new cases developing in the defined cohort or, more rarely, in an enumerated dynamic population. Sometimes only a sample of all newly developed cases is taken. The cohort can be a research cohort, an occupational, or educational cohort, or any cohort for which relevant exposure and follow-up data are or can be made available.

9.6.2 Selection of Controls in Case–Control Studies

A subject is eligible as a control if one can answer "Yes" to this question: "If the subject had been sick with the case-defining illness, would (s)he have been in the study as a case?" This question captures the requirement that controls should be representatives of the source population (*See:* Chap. 6). As a group, the controls should reflect the expected exposure distribution in the source population. Consequently, control selection must be independent of exposure such that exposed persons are not over- or underrepresented (a requirement that is similar to that for case selection). Controls must not be a special group that actively avoids or engages in the exposure. This would exaggerate or underestimate, respectively, the odds ratios estimated in the study.

9.6.2.1 Sources of Controls in Case–Control Studies

Possible sources of controls are equivalent to the above-listed sources of cases:

- Controls sampled in communities
- Controls from national or regional disease registers
- Patient controls identified in care settings
- Neighbors, friends and relatives
- Controls selected from an enumerated cohort or dynamic population under follow-up

For each of these possible sources of controls we can list advantages and disadvantages for feasibility and validity, in a similar fashion as for case selection. For example, identifying controls directly in the communities where the cases occurred is logistically difficult but has the least potential for selection bias. When cases are selected from a hospital, controls are often selected among patients having other diseases in the same hospital. Such hospital controls are easier to find and enroll than community controls and, once enrolled, there may be less danger for recall bias and non-response. However, the danger of selection bias tends to be higher. With hospital controls it is generally more difficult to convincingly argue that they validly represent the true source population. It is also sometimes unclear whether the illness of controls is truly unrelated to the exposures studied. In addition, it may be difficult to convincingly argue that their referral, diagnosis, eligibility assessment and acceptance of participation were also exposure-independent. A better option is often to recruit the controls among clients of doctors who would refer their clients to the hospital where the cases were recruited (if they would acquire the case-defining illness). When identifying such a group one needs to take into account the implications of the definition of source population. For example, clients of a doctor who refers such clients to another hospital cannot be controls.

When neighbors, friends and relatives are chosen as controls, the possibility of selection bias is generally very high. Thus these sources cannot be recommended as a general strategy but can be an option when cases are recruited via snowball sampling. The problem is that neighbors, friends, and relatives of cases often have very similar environmental and behavioral exposure patterns to the cases, not typical for the source population at large.

9.6.2.2 Control Selection Modes in Case–Control Studies

Sampling schemata for controls can be distinguished firstly according to whether the controls are sampled:

- As a group, among non-cases considered to represent the source population ('traditional approach')
- Concurrently with the cases ('concurrent sampling approach')
- From the entire source population regardless of whether they happen to be cases or not ('inclusive approach') (Rodriguez and Kirkwood 1990)

The inclusive approach has regrettably remained very exceptional. It has the advantage that it leads to direct estimation of the incidence rate ratio (*See:* Chap. 22). When the traditional approach is used, a group of eligible non-cases is selected into the study. One considers the date of their inclusion, which is typically identical for all, as the end of their individual exposure and risk period (the zero time-point of negative etiologic time, *See:* Chap. 6). When concurrent sampling is used, one or more controls are sampled each time a case becomes manifest, out of a source of eligible subjects who were at risk for developing the case-defining condition but did not develop it. Here, the zero time-points of etiologic time are spread out over calendar time both for cases and for their selection time-matched controls. If the subjects at risk at the time a case develops are a well-enumerated group, then they are said to form the 'risk set' at that time and the control sampling is then often called 'risk-set sampling'. With this method controls can be sampled more than once and controls can later become cases. Risk-set sampling is often done in nested case–control studies. Figure 9.1 illustrates the method.

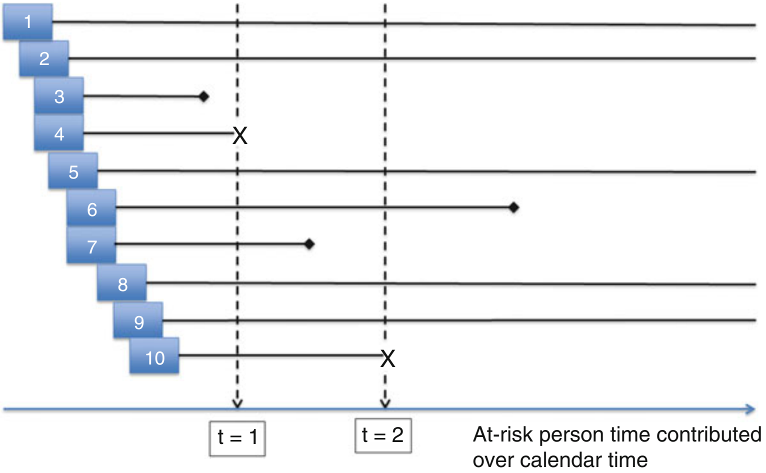


Fig. 9.1 Illustration of the principle of *risk-set sampling* in a nested case–control study with a control-to-case ratio of 1. *Horizontal lines* represent the person time contributed over calendar time by the first ten subjects in a cohort. The position for each subject reflects the time of enrollment. *Lines* represent person time of subjects. Case development is denoted by X, and loss to follow-up is indicated by a diamond. The first subject developing the case-defining condition is subject 4 at t = 1. The second case occurs at t = 2. At t = 1 the risk set is composed of subjects 1, 2, 5, 6, 7, 8, 9, and 10. One control is randomly selected from this risk set. At t = 2 the risk set from which a control may be sampled consists of subjects 1, 2, 5, 6, 8, and 9

A second way to classify sampling schemata for controls is according to how many controls are selected for each case, and if several are selected for each case, whether these are all of the same type or of different types. Multiple 'same type' controls are used to increase the power of the study, but there is little advantage in having more than four controls per case (*See:* Chap. 7). 'Different type' controls may be, for example, one hospital control and one community control, or, two different control diseases. This is sometimes used to study possible biases. When similar findings are obtained with each type of control, this is sometimes interpreted as indicating lack of bias, although it is obviously not a strong argument since bias may be equally big in the two control groups.

9.6.3 Types of Selection Bias in Case Control Studies

In case-control studies the selection of cases and controls is usually done separately. Hence there can be case selection bias, control selection bias, or both.

9.6.3.1 Case Selection Biases

Most case selection biases arise from the cases' survival, referral, diagnosis, eligibility assessment, or acceptance of participation being associated with the exposure(s) under study. In Panel 9.6 we describe the types of case selection bias accordingly.

Panel 9.6 Types of Case Selection Bias in Case–Control Studies

- **Case survival bias** – *See:* discussion about disadvantages of using prevalent cases.
- **Case referral bias** – Cases may have had a higher chance of being referred from lower level facilities to the study hospital/clinic or diagnostic center if exposed (or unexposed). For example, consider a clinic-based case–control study about malnutrition as a possible causal risk factor for persistent diarrhea. Patients with persistent diarrhea may have been more likely to be referred if malnourished than if well-nourished. This would tend to inflate the observed exposure odds among cases which could lead to an overestimation of the odds ratio.
- **Case ascertainment bias** – Diagnosis may be more often made among the exposed, so that the unexposed are less likely to become a case. An example is given in Textbox 9.2. Some epidemiologists classify this type of bias as information bias, although case ascertainment is a necessary step in case selection.
- **Case eligibility assessment bias** – Inclusion as a case in a case–control study also passes through a phase of eligibility screening. This involves more than only diagnosis. It can also involve severity assessment, assessment of time since first manifestation of illness, and assessments of other eligibility criteria. All these steps can theoretically lead to bias if the decisions made are influenced by exposure status.
- **Case non-participation bias** – Refusal can be associated with exposure. Imagine a case–control study on blood transfusion as a risk factor for HIV infection. HIV-positives may be more likely to consent to participation if they think they got HIV through blood transfusion than if they think they got it through sexual contact with commercial sex workers. HIV-negatives' willingness to participate would probably be more independent of the exposure.

As to case ascertainment bias, when misclassification in case ascertainment is *non-differential* (i.e., similar in the exposed and unexposed), the lack of sensitivity and/or specificity tends to bias the estimated odds ratio towards the null value (i.e., towards an odds ratio of 1). To illustrate this further, a scenario is described in Textbox 9.2.

Similarly, still with a causative exposure, a non-differential lack of specificity of case ascertainment among exposed and unexposed will tend to preserve the exposure odds among controls but will decrease the exposure odds in the cases and will thus also underestimate the odds ratio.

When misclassification in case ascertainment is *differential* as to exposure level (sensitivity and/or specificity are different among exposed and unexposed) the effect will not necessarily be an underestimation of the odds ratio, but could be an overestimation of it, depending on how the exposure odds in cases and controls are affected.

Textbox 9.2 Non-differential Misclassification in Case Ascertainment: Effect on the Estimated Crude Odds Ratio in a Case–Control Study

Consider a case control study of the effect of poor housing conditions on the occurrence of asthma, and assume there is a true effect e.g., an odds ratio (OR) of 2.67, with the true odds of exposure among the 100 cases being 4 and the true odds of exposure among 200 controls being 1.5:

		Asthma		True
	Poor housing	+	–	**OR**
	+	80	120	
	–	20	80	
Exposure odds		4.0	1.50	**2.67**

High specificity but poor sensitivity in the diagnosis of asthma implies that a proportion of children with asthma are not diagnosed but nearly all those diagnosed will be true cases. When the low sensitivity is non-differential i.e., equal in the exposed and unexposed, the exposure odds of 4 among the cases (numerator of the odds ratio) will be preserved. However, among the controls, the exposure odds (denominator of the odds ratio) could falsely become higher if there are non-diagnosed children with asthma (who have more frequently been exposed) amongst them. The trend will be one of relative over-representation of the exposed among the controls. The consequence will thus be an underestimation of the crude odds ratio. How much underestimation there will be depends on such factors as exact sensitivity, type of controls used, and the prevalence of asthma in the total source population. A possible observed scenario is:

		Asthma		Observed
	Poor housing	+	–	**OR**
	+	80	140	
	–	20	60	
Exposure odds		4.0	2.33	**1.71**

9.6.3.2 Control Selection Biases

As mentioned, in practice the selection process of controls is usually separate from the selection of cases, which leads us to consider control selection biases as a separate class of bias. We list them in Panel 9.7. Note that case selection biases and control selection biases often co-occur. What the expected overall effect is on the estimate of the odds ratio is in such cases not always clear, but the biases may cancel each other out or be superimposed on each other.

Panel 9.7 Types of Control Selection Bias in Case–Control Studies

- **Control source bias** – the chosen source is inadequate; Subtypes are:
  - **Control sampling frame bias** – The frame from which the controls are sampled may not adequately represent the source population
  - **Exposure-related control illness bias** – For example, in a study about smoking as a risk factor for cardiovascular disease, patients with chronic obstructive pulmonary disease would be poor controls since this is a smoking-related illness. Patient controls can also be a highly medicalized group of people who deliberately avoid a variety of exposures including the exposure of interest
  - **Exposure-related healthy control bias** – *See:* text above about the use of neighbours, friends and relatives of cases (Sect. 9.6.2)
- **Control survival, referral, and diagnosis biases –** These types of control selection biases can occur if patient controls are chosen. The mechanisms are the same as those operating for the corresponding types of case selection bias (*See:* Panel 9.6)
- **Control non-participation bias** – Refusals among controls can be associated with the exposure

9.7 Duration of the Recruitment, Eligibility Screening and Enrollment Periods

The recruitment period is not necessarily the same as the screening and enrollment periods; there may be slight timing differences among the three. Initial enrollment rates are often lower or higher than expected and the recruitment and enrollment periods can often be shortened without too many problems except if enrollment was scheduled to be evenly spread over seasons of the year or another calendar period. This is sometimes planned for studies aiming at estimating a period prevalence or at eliminating seasonality as a confounder. Prolongation of the enrollment period may have influences on study budget and usually requires renewed ethics approval. Issues around faster or slower than expected enrollment rates are discussed in greater detail in Chap. 17 (Accrual, Retention and Adherence).

In follow-up studies the total follow-up phase of the study is approximately the duration of the enrollment period plus the duration of the individual follow-up. When the enrollment period is very long, there is a greater risk of so-called ‘cohort effects’ occurring. This means, in this case, that subgroups enrolled over different calendar periods tend to have or acquire, during the follow-up period, different distribution matrices of determinants and covariates. In other words, a lot may happen over a long enrollment period. The early and the late enrollees may have been exposed to quite different circumstances.

In this chapter we discussed aspects of planning recruitment, sampling, eligibility screening, and enrollment activities. In the course of a prospective study, apart from measurements done in pilot studies and for eligibility screening, the 'real' data collection phase of the study usually starts with the enrollment of the first subject. To guide data collection, a measurement plan is needed as well as a plan for quality assurance. Therefore, in the next chapter we discuss the measurement plan.

References

Armitage P, Berry G (1988) Statistical methods in medical research. Blackwell, Oxford, pp 1–559. ISBN 0632015012

Coogan PF, Rosenberg L (2004) Impact of a financial incentive on case and control participation in a telephone interview. Am J Epidemiol 160:295–298

Council for International Organizations of Medical Sciences (2009) International ethical guidelines for epidemiological studies. CIOMS, Geneva, pp 1–128. ISBN 929036081X

Council for International Organizations of Medical Sciences (2010) International ethical guidelines for biomedical research involving human subjects, CIOMS, Geneva. http://www.cioms.ch. Accessed Sept 2012

Dillman DA (2000) Mail and internet surveys: the tailored design method. Wiley, New York

Galea S, Tracy M (2007) Participation rates in epidemiologic studies. Ann Epidemiol 17:643–653

Herold JM (2008) Surveys and sampling. In: Gregg M (ed) Field epidemiology. Oxford University Press, Oxford, pp 97–117. ISBN 9780195313802

Miettinen OS (1985) Theoretical epidemiology. Delmar, New York, pp 1–359. ISBN 0827343132

Moorman PG et al (1999) Participation rates in a case–control study: the impact of age, race, and race of interviewer. Ann Epidemiol 9:188–195

Rodriguez L, Kirkwood BR (1990) Case–control designs in the study of common disease: updates on the demise of the rare disease assumption and the choice of sampling scheme for controls. Int J Epidemiol 19:205–213

Sommerfelt H et al (2012) Case–control studies with follow-up: constructing the source population to estimate effects of risk factors on development, disease, and survival. Clin Inf Dis. doi:10.1093/cid/cis802

White E, Armstrong BK, Saracci R (2008) Principles of exposure measurement in epidemiology. Collecting, evaluating, and improving measures of disease risk factors, 2nd edn. Oxford University Press, Oxford, pp 1–428. ISBN 9780198509851

The Measurement Plan

10

Jan Van den Broeck, Jonathan R. Brestoff, Ari Friedman, Nora Becker, Michael C. Hoaglin, and Bjarne Robberstad

> *If you can't describe what you are doing as a process, you don't know what you are doing.*
>
> W. Edwards Deming

Abstract

In epidemiological research measurements are carried out most importantly to document data on outcomes, exposures, and third factors, but measurements related to procedural or methodological considerations should not be ignored. At the planning stage, it is crucial to conduct a step-by-step critical analysis of the measurement processes that will be employed in the study and to consider how errors at each step can be avoided. By carefully documenting this process for each planned measurement, one assembles a measurement and standardization protocol that conforms with general epidemiological principles by respecting participants and by enhancing reproducibility, completeness, unbiasedness, and precision. We briefly review planning and standardization issues according to type of attribute. Finally, special sections are devoted to quality of life and cost measurements in order to highlight the increasing importance of these in practice.

J. Van den Broeck, M.D., Ph.D. (✉) • B. Robberstad, M.Sc., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J.R. Brestoff, MPH • N. Becker • M.C. Hoaglin
Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

A. Friedman, M.Sc.
Department of Health Care Management, Wharton School of Business
and Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_10,

10.1 Study Measurements

A critical task in epidemiology is to define attributes in need of measurement. Selecting these attributes is part of a study's general design; therefore, discussions on this topic are found mainly in Chaps. 5 and 6. Recall that there is a general preference for higher measurement levels, non-invasive measurements, attributes with clear definitions, attributes measurable with validated measurement tools, and direct measurement as opposed to vague proxies. In this chapter we proceed from general design to methods design, specifically to the planning of the actual measurement activities. See Panel 10.1 for selected terms and concepts relevant to this chapter.

Panel 10.1 Selected Terms and Concepts Relevant to Measurement Planning

Analog display (of a measurement instrument) A display on which the measurement value is suggested by the position of a movable arrow or line on the graduated scale. The actual value is meant to be estimated ('read') in reference to the values of the nearest graduation mark(s)

Anthropometry Practice and science of measuring bodily dimensions using non-invasive instruments, scoring the measurement values and making inferences about growth and nutritional status or about health risks of individuals or populations

Assessment Determination of the importance, size or value of something

Biometry Branch of statistics that supports research concerning living beings in biology, medicine and agriculture. *Syn.* Biostatistics

Calibration (status) checks Assessment of the accuracy of a technical measurement instrument

Calibration status Degree of accuracy of a technical measurement instrument

Continuous measurement scale Scale for measuring continuous underlying attributes, expressing measurement values as multiples (with any number of decimals) of a measurement unit

Data Recorded information regardless of form

De-calibrated Status of a measurement instrument that lost accuracy to an unacceptable degree

Decimals In the notation of algebraic numbers, any digits that indicate fractions of integers

Digital display A display that does not show (part of) a graduated measurement scale for reading of the measurement value but directly presents the measurement value itself

Duplicate measurements Two measurements repeated independently within an interval considered short enough for no measurable change of the underlying dimension to have occurred

(continued)

Panel 10.1 (continued)

Independent measurements Repeated measurements, with each repeat done under conditions that minimize any influence of the previous measurements on the results

Instrument Question(s) or apparatus that helps in obtaining a key description of an attribute or experience of interest. *Syn.:* Measurement tool

Integer Whole number i.e., a number without a fraction or decimal (In mathematics also comprising numbers with an infinite number of nines as the decimals)

Interview Method of data collection based on asking questions orally (face-to-face or over some communication medium) to persons and recording the elicited responses or their inferred meaning

Measurement An act or process that leads to the recorded description of an attribute or experience of a single observation unit in the form of a value placed on a measurement scale or a brief textual summary

Measurement plan A plan as to who should measure what, when and how during a research study

Measurement schedule Planned timing of the sequence of measurements within measurement sessions and/or planned timing of the measurements during individual follow-up

Measurement value Result of a measurement, expressed as a particular position on a measurement scale

Non-invasive Not involving any direct entry into body tissues nor any potentially painful and damaging mechanical forces on body tissues

Score Position of a measurement value on an ordinal or numerical (discrete or continuous) measurement scale

Scoring Locating an individual measurement value on an ordinal or numerical reference scale

Table 10.1 gives an overview of measurement activities; the types of research staff usually involved in those measurement activities; and the study phase in which these measurement activities are typically planned, performed, or considered.

The planning of the measurement activities involves (1) identifying, for each attribute, the source of data (if any can be found) and a strategy of accessing this source, and (2) developing a well-standardized technical measurement plan that maximizes completeness and optimizes validity and precision in an efficient and ethical fashion. These topics are covered in a general way in the next two sections, after which some planning issues for specific types of measurement are reviewed.

Table 10.1 Overview of measurement activities in epidemiological research

Type of measurement activity	Research staff usually involved	Study phases usually concerned
Measurements to inform study design and operations	Investigators, study coordinators, supervisors, data collection staff, laboratory staff	Systematic literature review, pilot studies
Training and test measurements	Same as preceding row	Study preparation
Eligibility screening measurements	Screening and enrollment staff	Screening and enrollment
Exposure, confounder, and effect modifier measurements	Data collection staff, laboratory staff	Exposure and covariate data collection
Outcome measurements	Same as preceding row	Outcome data collection
Quality control and adherence measurements	Quality control staff, supervisors, study coordinator	Pilot studies, study preparation phase, entire data collection phase

10.2 Data Sources and Collection Strategies

10.2.1 Sources of Data

Epidemiological studies produce primary or use secondary data, or both. This distinction is based on whether the data are collected specifically for the research study at hand or for some other purpose. *Primary data* are collected for principal use by the researcher. *Secondary data* are collected for other purposes but are now used for the study at issue. Common sources of primary and secondary data are listed below (Table 10.2).

Issues of primary or secondary data collection are in principle unrelated to timing of the study base experience, be it retrospective, prospective, or ambispective. In retrospective studies, a question arises as to whether primary data on that past experience will be newly collected in the future (by anamnesis) or if any secondary data collected previously will be used. In prospective studies, the same question may arise: shall the researcher collect new data or use data collected for other purposes? Depending on the particular situation, either collection of primary or secondary data may be more cost-effective or unbiased.

10.2.1.1 Medical Records and Chart Reviews

There are important differences between clinical history taking and standardized questionnaire administration. Similarly there are differences between clinical physical exams and biomedical measurements done for research purposes. In a traditional clinical context, history taking and physical examination of a presenting patient involves a continuous mixture of observations and subjective interpretations in a typically unpredictable subjective sequence. For example, selecting which systems should be subjected to a detailed physical examination depends on an interpretation of the patient's presenting symptoms and clinical history. Since no two patients are identical, examination of *k* number of patients by *m* number of clinicians will most likely proceed in *k*m* number of ways. A scientific approach, as should be taken in research, should attempt to avoid all interpretations until the end of the study. All participants should be examined the same way except where the object design calls for a difference. Standardized questionnaires and bio-measurements are meant to provide complete data on all relevant attributes of all participants, whereas medical records tend to provide scattered, incomplete data on a mix of attributes (aimed at clinical relevance) that may be different among patients.

One must therefore consider the limitations of medical records and verify the completeness and format (especially units and levels of measurement) of the data that need to be extracted from medical records before deciding to use this source of data in a research study. One must also acknowledge the fact that 'no information' about a condition does not imply 'no existence' of that condition. For clinical studies with a prospective approach, it may be possible to add a research component to an existing medical record system by adding, for example, an addendum to a participant's paper or electronic chart.

Table 10.2 Sources of primary and secondary research data

Common sources of primary research data	Common sources of secondary data used for research
Questionnaires	Patient files ('patient charts')
Biological samples	Electronic medical record systems
Bio-measurements (e.g., anthropometry)	Census data and vital statistics
Medical imaging	National health information systems
Information from direct observation	Hospital discharge statistics
Audio-recordings	Health center utilization/service statistics
Videos	Health and safety surveillance programs
	Disease registries
	Public or private archives or research data

Textbox 10.1 Electronic Medical Records as a Secondary Source of Research Data

Electronic medical records (EMR)

EMR, also known as electronic health records, are becoming an integral component of many clinical practices. Many countries have implemented programs to promote the use of EMR systems, and the financial commitments by many governments to support health information technologies continue to grow. The perception is that EMR systems will reduce healthcare costs and improve the quality and efficiency of healthcare in the long run. However, EMR systems will hardly realize their full potential until the data contained within them contribute to future healthcare innovations via research.

(continued)

Textbox 10.1 (continued)

Epidemiologists will thus be challenged to leverage EMR systems as if they were research databases (Frankovich et al. 2011). There are also ethical challenges to this as data obtained for these databases are primarily for clinical purposes.

Structured versus unstructured EMR data

EMR data can be divided into two main types: structured data and unstructured data. Structured data are ready to be directly operated on by a computer system. Most simply, these are alphanumeric fields that the computer recognizes, such as name, phone number, lab values, and vital signs. Structured data are more ready for extraction, data cleaning, transformations, and analysis than are unstructured data. Unstructured data often involve free text fields or images that aren't as immediately meaningful to a computer program. For example, scanned images of lab values or a typed referral letter are not immediately available for processing. However, with optical character recognition, natural language processing (NLP) and careful data mining techniques, useful information can be obtained from unstructured data. Data mining from unstructured data is a relatively new field that is rapidly growing because data sources, such as physicians' progress notes, often contain critical knowledge not captured elsewhere in the patient's electronic chart.

Natural Language Processing (NLP)

Narrative text is unavoidable in the EMR, but NLP technologies offer a solution to convert free text data into structured representations. While not perfect, it can be less error-prone than the laborious, resource-intensive task of manual structuring of data. NLP technologies are based on 'linguistic ontologies' that can be customized for particular projects. While this adaptability is a great strength of NLP technologies, customization requires very careful programming, such that "hypertension" and "high blood pressure" and typographical errors of these terms are classified similarly. But they also must correctly identify acronyms (e.g., HTN for hypertension) that often vary considerably among providers and subtle yet critical words, such as negations, that may exist in a clinician's note on a patient.

Clinical Data Repository (CDR)

All raw EMR data for a patient are stored electronically in a CDR, a database that underlies all of an EMR's applications. A CDR allows the user to run reports and searches, analyze statistics, perform computations, import and export data, and manipulate data sets. The primary use today is for budgeting and internal health system monitoring. Obtaining the schema or data model for a research institution's CDR may be helpful in defining future research questions.

(continued)

Textbox 10.1 (continued)

Identifying relevant records for data extraction

When planning a research project using EMR data, one must realize that an initial query for just the most critical attributes will produce a very long screening list of '*encounters*' i.e., episodes during which information on the attribute(s) may have been recorded. New clinical data for a given patient are virtually always entered in association with an *encounter number*. Encounters are associated with a date, location, provider, and one or more diagnoses and are created in association with an inpatient stay, an outpatient visit, a patient phone call, old imported records, or even an e-mailed question to a physician about a patient. These are considered *encounter-level* data. Without an encounter, the data are considered to be *patient-level*. Patient-level data, such as name, contact information, primary care physician and other, are associated with a patient's medical record number, a unique and anonymous identifier.

EMR-based research projects

One of the primary challenges in this endeavor is the standardization of disparate health data from the nation's many health care organizations and providers. Traditionally, EMR data are stored inconsistently and in multiple silos. Researchers may need to work closely with their local IT department to discern the schema or clinical data model in question. In some cases one may be able to access a secure data mart or a custom dashboard for creating queries, much like a familiar relational database. For some projects it may be appropriate to obtain EMR data provided by a systems administrator in a format as simple as a spreadsheet.

Privacy and Security

A common misconception about health IT is that it excessively exposes protected health information to unauthorized parties. EMR can actually provide more security than paper medical records. In the early days of health information privacy laws, caregivers were taught that it was a violation to look at the records of a patient for whom the caregiver had no clinical responsibility. In an EMR, doing so creates mandatory electronic audit trails making it virtually impossible to do it secretly. However, when a paper record is accessed or altered, there is no automatic audit of who saw what when and for how long. Being familiar with privacy laws and working cooperatively with the institution's legal department and ethics committee will ensure that patient privacy is respected while not creating unnecessary barriers to innovation.

10.2.2 Data Collection Modes

A distinction can be made between direct measurements and staged measurement processes that involve some intermediate storage of material for final measurement and recording. An advantage of the latter is that re-measurement can usually be done

Table 10.3 General advantages and frequently encountered disadvantages of various data collection modes

Mode of data collection	Advantages	Disadvantages
Secondary data look-up	Inexpensive	Cumbersome formatting
	Fast	Incomplete data
		Limited variables
Direct observation of behavior or environment	Inexpensive	Subjects may display unnatural behaviors in research context
	High objectivity possible	Lower order measurement scales
Face-to-face interviews	Personal contact facilitates higher response rates	Expensive
	Independent of literacy	Time consuming
		Interpersonal dynamics can inappropriately interfere
		Interviewer variation
Telephone interview	Inexpensive	No non-verbal cues
	Wide area of coverage possible	Suspicions often aroused
		Questionnaire needs to be short
		Less feasible in many low- and middle-income countries
Mail, email, and Internet questionnaires	No interviewer variation	Low response rates
	Anonymous	Time consuming
		Difficult to elicit detailed responses
		Depends on literacy
		Less feasible in many low- and middle-income countries
Bio-measurements	High validity possible	Can be invasive
		Can be expensive
		Depends on high instrument quality
		High technical operator skills required

easily and by different observers, which offers possibilities for quality assurance and control.

Measurement frequently involves recording values on questionnaires (*See:* Chap. 18) that are later transferred into an electronic database manually or using an optical scanner. Sometimes electronically stored results are an immediate output of the measurement process without a need for intermediate recording on a form. The Ulmer stadiometer, for example, reads a person's standing height automatically and simultaneously transfers it to a database. Results of biochemical analyses on biological samples increasingly involve automated reading and electronic recording as well. Other electronic aids often used in data recording include computer-based questionnaires (that may involve the Internet or mobile devices) and tape- or video-recorded encounters.

General advantages and frequently encountered disadvantages of some common data collection modes are listed in Table 10.3. Space constraints preclude detailed discussions of each here. Environmental and bio-measurements are discussed further in Sect. 10.4. Questionnaire administration modes are discussed in Chap. 18. As to self-administered questionnaires, these can be mailed, transmitted electronically, or be handed out in person (e.g., at a medical consultation). The latter approach tends to increase response rates (Herold 2008), which otherwise tend to be quite low. Mailed or emailed questionnaires are most useful for closed-circle target populations, such as employees or members of organizations (Herold 2008). Not all parts of a questionnaire need to be delivered the same way. For example, questions about eligibility may be asked over the phone, and if the person is eligible, then an invitation is issued for a face-to-face interview.

Chosen data collection mode(s) influence(s) the required type and number of data collection personnel and their training (*See:* Chap. 15). In many studies data collection is an activity of study personnel specially hired and trained for the purpose. In clinical studies it may be the investigator-physician herself who (also) collects data. When making decisions about who collects data, one should take into account that problems of inaccurate reporting can arise if the person collecting data is also the person administering an intervention. Consider, for example, a study of the effect of various modes of repeated postnatal counseling about exclusive breastfeeding on time to cessation of exclusive breastfeeding. The measurement plan may include that, at each contact, the same interviewer-counselor first counsels about the virtues and modalities of breastfeeding and, immediately afterwards, asks questions about whether the child is still exclusively breastfed. This plan would amount to an invitation to misreporting.

10.2.2.1 Mobile Devices as Research Tools

The use of mobile devices as interview aids is on the increase and has several advantages. Modern cell phones can display the question; guide the interviewer/examiner or participant through the data collection process; and facilitate direct protected data entry (Vital Wave Consulting 2009; OpenXdata 2010). This system tends to be cost-effective despite the cost of electronic devices; allows automatic recording of the date, time, and location of an interview/examination; and often enables incorporation of audio or visual media into the interview/exam or database. Mobile devices also facilitate quality control procedures, as members of the study team can rapidly consult each other or important references as issues arise. Even real-time analysis in remote places can be made possible with cell phones. There are also some limitations associated with the use of mobile devices as research tools. An important one is that technical difficulties may compromise data collection processes. If one wants to employ mobile devices in a study, logistical aspects of their use and programming will need to be investigated thoroughly to confirm feasibility and utility for the given scenario. Data management issues are further discussed in Chap. 12.

10.2.3 Exposure History Taking

Exposures are aspects of gene-constitution-environment interaction. Perhaps the most common approach to documenting exposures is history taking. This involves extraction of information from the participant's memory via a structured question-and-answer approach i.e., via structured anamnesis.

10.2.3.1 Fine-Tuning of Design Preceding Anamnesis Planning

Preliminary explorations may be needed to clearly define the exposures of interest and the relevant etiologic period for each of those. Some studies have a broad mission to investigate the causes of an epidemic. In such circumstances it may be unclear what exactly needs to be measured. Only when there is clarity about the nature of the exposures, their range of intensities, and their relevant etiologic period can one design a measurement plan. In other words, *refined object design* may be needed. Defining the object design in a study *must always precede measurement planning.* Preliminary explorations and design efforts may require close collaborations with technical experts and/or basic biomedical researchers. It is thus useful to remain aware of major relevant developments in other disciplines.

10.2.3.2 Reconciling Valid Recall Period and Etiologic Period

The next step in planning questions is to consider carefully what a valid recall period is. How long ago was the exposure period at the time the participant is asked the question? For this particular exposure and for this particular type of participants, what is already known about the accuracy and reproducibility of a question (or set of questions) as a function of recall period? This crucial information helps to verify the compatibility of the etiologic period of interest and the valid recall period. This exercise may lead to the dramatic conclusion that the plans for a study need to be abandoned or that the entire object design needs to be revised from scratch. For example, the initial plan may have been to assess, by history taking in a group of mothers of 1-year-olds, the duration of exclusive breastfeeding as a determinant of a particular health outcome at age 1 year. Careful study of the epidemiologic literature will show that exclusive breastfeeding duration cannot be accurately measured in this way because the recall period is too long to produce accurate information (Bland et al. 2003). Situations arise where no reliable information about the recall period–accuracy relationship can be found in the literature. This may be a reason for doing a methods-oriented pilot study first. Another possible conclusion from the comparison between etiologic period and valid recall period is that only a part of the etiologic period can be addressed in the study, which may or may not be a satisfactory solution sufficiently in line with the general objectives of the study.

There are instances where present exposure can be taken to fairly represent past exposure in the relevant etiologic period. In cross-sectional association-based etiognostic studies (*See:* Chap. 6), this is a necessary assumption that needs to be met strictly. An example is current exposure to environmental factors in a confined setting (e.g., a work setting) that has not changed appreciably over the etiologic period considered.

10.2.3.3 Questions on Timing and Intensity Patterns of Exposures

Within the relevant etiologic time span, the exposure to ask questions about may be:

- A single event, e.g., having eaten a particular food item on a particular occasion
- Repeated events, e.g., instances of exposure to X-rays
- A permanent characteristic e.g., sex

- An episode of interaction between body and environment, e.g., a visit to a particular geographical area
- Repeated similar periods of interaction between body and environment, e.g., periods of working in a particular type of workplace

The type of exposure tends to have a bearing on the statistical analysis methods that can be used in the study. For example, longitudinal analysis methods will tend to be more useful for repeated events exposures than for a single event exposure. Depending on the object design, questions of interest may concern the exact timing, duration, and frequency of exposure; the intensities and patterns of exposure; or the exact nature of an exposure (if subtypes of the exposure are of interest).

With information on exact timing available, one may classify subjects according to mutually exclusive reported exposure time categories. Miettinen gave an example of a simple, correctly classified time exposure history in an etiognostic study (Miettinen 1985): the classifications were (1) never used a contraceptive pill; (2) current use only; (3) past use only, 1–5 years ago; (4) past use only, 5–10 years ago; (5) past use only, 10–20 years ago; or (6) other. In this example it is clear that the 'Other' category is important in that it represents the experience of all those who used a contraceptive pill in more than one of the historical time segments. With such a classification of exposure histories, it is possible to validly contrast any of the categories of past use with the 'never used contraceptive' category, except for the 'Other' category, which is a mixed bag category of little further use in the synthesis of the data. Alternatively, one may wish to use more information from this 'Other' category and treat each historical segment as a separate attribute, each with a separate exposure variable representing it in a regression analysis. In any case, characterization of exposures in historical segments usually requires separate questions concerning each of the segments. For instance, in Miettinen's example, it would be important to ask separate questions about the different time segments before attempting to arrive at the proposed classification. The reason is that it would require the respondents to possess a high capacity for abstraction to be able to select the appropriate options about 'past use only'.

10.2.3.4 Measuring Cumulative Exposure Using Anamnesis

Sometimes there will be only an interest in cumulative exposure over the entire etiologically relevant period, not in any patterns of how this accumulation came about. In this case the 'amount of exposure' may be approximated, for example, by a summation of intensity-weighted exposure durations in pre-defined mutually exclusive time segments. If the exposures are repeated events, the number of events (perhaps intensity-weighed) is sometimes taken to represent cumulative exposure amount (e.g., pack-years of smoking).

10.2.4 Prospective Follow-Up Measurement Strategies

The scheduled *length of follow-up* is an element of the general design of longitudinal studies. This does not mean, however, that all individual observation units are followed for the same length of time. If the study base is a dynamic population, individuals are only followed for the time they fit the inclusion criteria. Units scheduled to be followed for the entire study period may be lost to follow-up for various reasons or acquire an ‘endpoint’ and therefore no longer contribute needed data.

10.2.4.1 Choice of Measurement Intervals and Time of Measurement

In any study, the ideal measurement interval may vary according to the study variable at issue. When cumulative or chronic exposure patterns need to be documented and the attribute shows considerable fluctuations within individuals, then more frequent or even continuous follow-up measurements will better characterize the exposure pattern. For continuous characteristics, the measurement interval must permit the potential for a change larger than the expected measurement error. For example, it would be senseless to measure a child’s height every day over a long observation period; a minimal interval of 6 weeks to 2 months would allow detection of small, true gains in height. Height gains over shorter periods would tend to be indistinguishable from measurement error.

A related problem arises when variables are known to fluctuate normally according to the time of day. For example, the circadian rhythm of blood pressure is well documented. Although the amplitude of this cycle is small at approximately 5 mmHg, taking measurements at different times of day in different groups nearly guarantees irreconcilable bias. Many physiological parameters are known to cycle in a circadian manner, examples of which include various white blood cell counts; hematocrit; some serum electrolytes; and many hormones, such as cortisol, melatonin, and catecholamines. One should turn to the literature for each variable potentially liable to time-of-day effects.

When secondary data are used, more follow-up data may be available than are actually needed. For instance, if the outcome is baseline-adjusted change in weight from start till end of follow-up, then weights obtained in the middle period of follow-up may be irrelevant. When events must be recorded by history taking, concerns about recall bias and desirability bias should guide selection of an appropriate interval between interviews. It may be possible to instruct participants to keep a diary on particular behaviors or of signs and symptoms for later use during structured interviews. This may improve accuracy and precision and may also decrease the number of follow-up visits needed.

10.2.4.2 Cost Saving Strategies in Prospective Follow-Up Studies

When the optimal measurement technique is highly reliable but very expensive, one may be led to consider the cost-reducing strategies at the expense of some accuracy (White et al. 2008). One option is to validate a less expensive (and less reliable) method against the expensive one in a sub-sample only. Another may be to pool samples within each main comparison group to determine a single (mean) level in each pool. Sometimes it is possible to analyze only a selection of the collected samples. This may be useful when the timing of an event needs to be estimated on the basis of a series of samples. For instance, when the timing of HIV seroconversion needs to be estimated, it makes economic sense to analyze only the full series of samples from individuals whose last sample showed seroconversion. Whether it also makes ethical sense to delay analysis of samples until the end of the data collection needs to be judged on a case-by-case basis.

10.3 Measurement Standardization

Measurement standardization is the application of an identical standard to measurement procedures.

10.3.1 Aims of Measurement Standardization

In epidemiological studies it is essential to ensure that all measurers uniformly apply optimal measurement procedures. What constitutes the optimal procedure depends on scientific, ethical, and practical considerations. For example, one might consider standardizing measurement procedures to reduce observer fatigue (an ethical issue) with the aims of improving overall measurement reliability (a scientific issue) and cost-efficiency (a practical issue). Properly executed measurement standardization creates data that are comparable among subjects, populations, or subgroups (Textbox 10.2). Failure to execute measurement standardization properly renders data incomparable and can lead to biased study findings.

More information on how measurement error affects validity of study findings can be found in Chaps. 11 and 27. In this section we deal with important ways of avoiding these problems of measurement bias.

Achieving successful measurement standardization can be complex, especially in studies involving many measurers. The preferred approach is to enroll all measurers in a study-specific tutorial, allowing them to be trained simultaneously or in batches. In small studies, however, a training process might amount to a few meetings of measurers

Textbox 10.2 Types of Standardization in Epidemiology

Not to be confused with **measurement standardization** (the topic of this section) is the **standardization of measurement values** (*See:* Chaps. 12 and 13), which refers to the scoring of measurement values in relation to a 'reference distribution,' thereby making scored values comparable. The **standardization of estimates** (*See:* Chap. 22) involves taking one population as a reference and using its underlying distribution of determinants to calculate adjusted estimates for another population. This process produces expected estimates had the underlying distribution of the determinants been the same as in the reference population.

Panel 10.2 Requisite Elements of Adequate Measurement Standardization

- Uniformity in prescribed procedures to be applied to each participant and by each measurer; it is preferable to use validated formal guidelines
- Detailed descriptions of procedures and the successive measurement processes
- Descriptions of what to do (and why) and also of what not to do (to avoid error); protocols/standard operating procedures should be employed to enhance objectivity and reduce subjectivity
- Training of the technical procedures up to or close to an expert level
- Proof of standardization through documentation of data quality (*See:* Chaps. 11 and 29)

and investigators. A special case arises in epidemiologic studies that involve only questionnaires completed individually and without assistance of research staff. In this case, subject-measurer contacts are indirect, so measurement standardization often takes the form of proper questionnaire development (*See:* Chap. 18).

Measurement standardization is critical to the general epidemiological principles listed in Chap. 1 (Panel 1.1), for example:

- To obtain accurate and precise measurement values
- To enhance data completeness
- To contribute to overall unbiasedness of evidence by making data comparable within subjects and observers (over time) and among observers and studies
- To ensure that measurements are efficient, such that no participant undergoes unnecessary lengthy or otherwise burdensome measurement sessions
- To ensure that measurements are taken in the safest possible way

Consider the example of venous blood sampling in children. Minimally, the measurement standardization plan should limit the number of attempts to find venous access, prescribe a sequence of body sites for those attempts, define conditions for referral to a pediatrician or expert phlebotomist, and instruct on the use of anesthetic skin creams or adhesives. The requisite elements of measurement standardization plans are shown in Panel 10.2.

10.3.2 Sources of Measurement Variation

A measurement plan should be based on what the expected sources of measurement variation are. When considering possible error sources one should keep in mind that each single measurement value is the end-result of a complex interaction between measurement instruments, the environment in which measurement occurs, a subject (and any accompanying persons), and usually also one or several measurers (Fig. 10.1). These interactions potentially make the measurement value inaccurate. As an example, consider measurement of usual alcohol consumption in a 17-year old adolescent boy, by questions asked in a face-to-face interview administered during

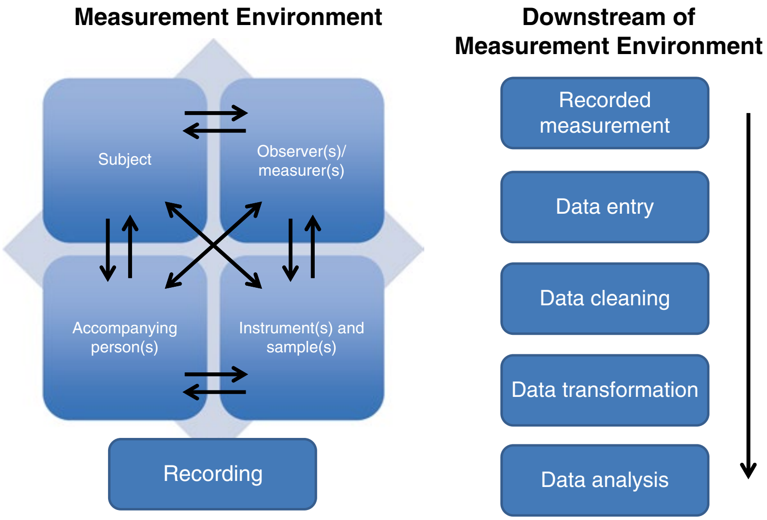


Fig. 10.1 Sources of error in a study variable that depends on a simple direct measurement act. All components of the measurement environment interact to influence the recording. Downstream of the measurement environment, the recorded measurement values are manipulated in a series of steps leading to data analysis. Errors can be introduced at any point from the first contact between the subject and the measurement environment through the completion of data analysis

a home visit. The interviewer visiting the home is a 22-year old female research assistant. The accuracy of the boy's responses may be influenced, among others, by:

- Formulating questions about alcohol consumption in a way that suggests alcohol drinking is for adolescents anything but neutral e.g., bad, unhealthy, or 'cool'
- Preceding questions about condom use; E.g., the boy has become embarrassed and is now very much looking forward to the end of the interview
- The television in the room is showing an exciting soccer match
- The boy's mother, who is strongly opposed to adolescent alcohol consumption, is listening at the other end of the table to what the boy answers
- The interviewer finds the boy attractive and impresses upon him with non-verbal cues; the boy recognizes these cues and, in turn, tries to make a favorable impression upon the interviewer through his answers

The example illustrates how the quality of the instrument, the behavior of the subject, the characteristics of the measurement environment, the measurement skills of the observer, and the behavior of accompanying persons can be sources of error. The example also implies that standardization can avoid these errors.

The interactions shown in Fig. 10.1 are those occurring during a single direct measurement act. Some of these separate measurement acts may be indirect, on sampled materials. For many attributes, several separate measurement acts will be needed to produce the final measurement values. Total error will then be determined by the accumulation of all errors occurring at all stages. With biological sampling, total error variance will be the sum of variance in sampling and storage technique, variance of the actual sample analyses, and variance from data handling after recording of analysis results. What is needed at the planning stage of the study is a step-by-step critical analysis of the measurement process and consideration of how errors at each step can be avoided. All the different measurements and their possible sequences need to be considered. The result should be a measurement and standardization protocol.

10.3.3 The Measurement and Standardization Protocol

A measurement protocol may contain long lists of detailed instructions. Each separate instruction may contribute little to overall data quality, but the combined effect of standardizing a large number of procedural components greatly enhances quality and comparability. Hence, all details are important. Instructions in a measurement protocol may represent a prevailing choice among possible alternatives. This choice may or may not be evidence-based and often represents a consensus among experts. Even for the most common measurements, there is still a need for methods-oriented research.

Panels 10.3, 10.4, and 10.5 provide a checklist of issues to consider when designing a measurement and standardization plan. Protocols for using devices may

Panel 10.3 Overview of Standardization Issues with Instruments

Choice of instrument

- Intrinsic validity
- Precision of the measurement scale
- Design features influencing accuracy or precision
- Applicability to the whole study population
- Usefulness in field studies

Instrument calibration

- Recognized and demonstrated validity
- Use of certificates of calibration
- Assembling; conditions of usage
- Checking of calibration status at start of research data collection
- Frequency of calibration checks during usage
- Equipment needed for calibration checks
- Technical calibration protocol

Maintenance of instrument

- Frequency
- Availability of well-described technical guidelines
- Robustness of instrument
- Cost, affordability
- Cleaning requirements

Instrument storage and transport

- Sensitivity to exposure to sunlight, temperature, humidity, dust, etc.
- Sensitivity to physical impact
- Best mode of transportation

need input from technical experts, such as environmental hygienists, industrial manufacturers, or laboratory technicians. The source of expertise may vary, but an epidemiological researcher must always acquire insight into any elements of the protocol that may affect validity, ethics, or efficiency.

10.3.3.1 Measurement Sessions

The planning of measurement sessions should foresee enough time for all measurements. The session should not be too long in total, and pauses of appropriate lengths should be inserted where appropriate. Sessions should be held at appropriate times

Panel 10.4 Overview of Standardization Issues with Environment, Subjects, and Observers-Measurers

Measurement environment

- Privacy
- Light, room temperature
- Space requirements; necessary furniture; placement of instrument
- Attractiveness; comfort for subject and observer
- Interference; sources of sensory stimuli e.g., radio, TV, loud voices

Subjects and accompanying persons

- Motivation; relationship; encouragement by accompanying persons
- Knowledge of what will happen
- Availability, time of day
- Physical condition; sickness, memory

Observers-measurers and assistants

- Number needed per measurement; number needed for the study
- Training and experience; skills, knowledge and specific training of the technical measurement protocol
- Motivation and attitude; mood; pre-conceptions
- Time; working hours, schedules; remunerations
- Physical condition; sickness, memory; alertness

Panel 10.5 Standard Sections of a Technical Measurement Protocol

- Preparation of subject
- Step-by-step instructions for:
  - Interaction of subject with instrument
  - Handling/application of instrument by observer
- Reading and recording instructions
- Replicate measurements
- Value checks; when to re-measure

of the day and, in follow-up studies, at appropriate intervals. Within sessions, the sequence of measurements and timing of any necessary replicate measurements can have an important influence on data quality. For example, sensitive questions or invasive measurements are best located at the end of a session, and, independent replicates should be made more independent by inserting other measurements in between. As to technical measurement protocols, one can sometimes make use of internationally accepted protocols, such as those available for blood pressure or standing height measurement.

There can also be issues of *standardization over time*. In prospective studies special problems may arise if, in the middle of the data collection period, an improved measurement method becomes accessible. Conversely, at times the optimal method may need to be replaced with a less optimal method. Similarly, in retrospective studies, changes in measurement techniques may have occurred. When measurements are not of the direct type, the solution may be re-analysis of old material. If not, the challenge is to replace values obtained with the sub-optimal method with predicted values under the better method. Regression modeling will require a set of doubly measured items and may lead to a simple conversion factor or a more complex model.

Pre-conceptions can influence the researcher's performance. Researchers may have strong expectations about the existence or direction of an association between risk factor and outcome. This may lead, for example, to an unintentional trend to positively identify expected outcomes among exposed and unintentional mistakes in the analysis. *Blinding* of measurers and investigators as to the exposure status during data collection and analysis can be a useful design decision.

10.4 Measurement Issues According to Type of Attribute

10.4.1 Measurement Issues of Occupational and Environmental Exposures

Data collection about occupational and environmental exposures, including treatments and other interventions, may use:

- Interviews, questionnaires
- Measurements of the workplace environment (e.g., water, noise, air pollution)
- Measurements of individual micro-environments (e.g., office, nearby machines)
- Human tissues (e.g., blood, urine, biological markers of past exposures)

Each of these sources of data has its limitations and will need to be considered on a case-by-case basis. As a brief example, measurements in the general workplace environment may ignore inter-individual exposure variation, and measurements in micro-environments may have unclear long-term relevance and may not reflect true exposure.

A chosen exposure measure tends to have greater validity when measuring 'closer to the physiological impact'. An example is a clinical trial in which the intake of test drug could be measured alternatively as 'dose prescribed', 'dose taken', or 'plasma levels after intake'. Proxies for exposure have often been used in studies in occupational and environmental health, for example an individual's distance to the source of pollution in the workplace. In relation to this example, one should keep in mind that those who do the most dangerous work may tend to also have more dangerous living conditions at home or partake in riskier activities. Such confounders need accurate measurement.

In the measurement of environmental exposures, short-term fluctuations often need to be controlled for by standardization. Continuous or frequent monitoring may be more relevant than a single or small number of scattered measurements. For example, exposure levels within a workplace may fluctuate according to particular circumstances, such as deadline-related workload, atmospheric conditions, technical problems with equipment, and human error. For further reading, *See:* White et al. (2008).

10.4.1.1 The Measurement of Interventions Received

Measuring intervention levels tends to be relevant in etiognostic and prognostic (intervention-prognostic and descriptive-prognostic) research. Attention should be given to the different components of the intervention strategies. The intervention of interest is likely to be accompanied by additional interventions which can be prescribed under the study protocol, initiated by health care workers (biomedical or alternative) or imposed by policy makers. Forms of additional interventions that are particularly frequent but not always accurately measured include (1) various types and intensities of advice and counseling and to what extent they are followed and (2) the use of non-prescribed medications (Bland et al. 2004).

Individual clients, patients, or other units may undergo different levels of actual exposure, adherence, or policy penetration of both the main intervention and the additional interventions. Ideally, all these individual levels of exposure should be captured accurately in all comparison groups during a study. The description of an intervention level as 'usual care' or 'standard care' is problematic; it does not allow clarity about the actual intervention contrast between index groups and reference group. But accurate measurement of the intervention can be difficult. One reason is that misreporting of adherence during interviews is frequent. Secondly, carefully observing intake or exposure at an individual level can be misleading. For example, part of the doses of drugs taken may not be absorbed due to spitting, vomiting, or malabsorption. Repeatedly measuring plasma levels is invasive and expensive, and often unacceptable to participants. Even integrated plasma levels may not adequately reflect what happens at the receptor level. Concentrations of many drugs can be measured in hair or nails, where they accumulate and reflect intake in the last weeks to months. This may be a preferred method in some settings, for example for the monitoring of adherence to antiretroviral drug treatments (Ghandi et al. 2009). The sampling is easy and non-invasive. However, there can be cultural taboos around the collection of hair or nails. Infant hair, for example, is particularly difficult to obtain in some African areas.

10.4.2 Measurement Issues of Constitutional Attributes

The following methods are frequently used to characterize constitutional factors.

10.4.2.1 Traditional Anthropometry

Weight and height measurements are among the most frequently performed of all measurements in medical research. They are variably used to construct indices of general nutritional status, general heath status, and total body adiposity (through the calculation of body mass index). Other frequent measurements are circumferences of head, neck, left mid-upper arm, waist and hip. The exceptional popularity of weight and height derives from their non-invasiveness, perceived simplicity of measurement, and widely accepted usefulness for assessing adiposity and obesity-related morbidities. Issues of economy and efficiency, especially in large studies, have made it common practice to ask subjects to report their body weight and height in lieu of direct measurement by research staff. Unfortunately, self-reported weight and height tend to be highly inaccurate, with increasing degrees of error for both as individuals become heavier. No reliable standardization protocols currently exist to adjust self-reported weight and height, so it is highly recommended to employ direct measurements.

- For an overview of the use of anthropometry, *See:* WHO (1995)
- For anthropometric standardization guidelines, *See:* Lohman et al. (1988) and Growth Analyzer (2009)

10.4.2.2 Medical Imaging

This family of methods is sometimes classified under anthropometry. Medical imaging, like traditional anthropometry, tends to be non-invasive. Yet, there may be concerns about, for example, exposure to radiation during X-rays, and the required preparations for patients who undergo imaging procedures may be burdensome. In research, the main roles of medical imaging are in case diagnosis and case severity assessment, and in assessment of body composition. Similar to traditional anthropometry, a main challenge lies in achieving high enough sensitivity and specificity through standardization of measurement and quality control of observer performance. A common fallacy is to judge the level of standardization on the basis of reproducibility and accuracy of readings of images without taking into account measurement variation attributable to subject preparation and technical aspects to imaging. It is also important to blind image assessors to the exposure level of the participant.

- Imaging techniques are not always within the area of technical expertise of epidemiologists and thus a good collaboration with radiologists and radiographers is often important in the research setup

10.4.2.3 Blood Pressure Measurement

Diastolic and systolic blood pressure measurements are other examples of widely performed non-invasive measurements used in many research projects. They are usually done in the context of measuring cardiovascular or renal health. Accurate and reproducible measurement of cardiovascular health aspects through blood pressure is very challenging, and elaborate standardization plans are required. This standardization plan should explicitly account for the time of blood pressure measurements given the well-characterized normal fluctuations in blood pressure over the course of the day.

- For standardization of sphygmomanometer-based blood pressure measurements *See:* Perloff et al. (2001), Chobanian et al. (2003), Pickering et al. (2005), and Shea et al. (2011)

10.4.2.4 Measurements Involving Laboratory Analyses

Methods sections of study proposals and study reports need to describe any biological sampling and laboratory methods used to measure constitutional characteristics or traces of environmental impact on the body. Panel 10.6 is a list of some major issues regarding laboratory methods that need to be addressed in the study proposal or protocol.

Panel 10.6 Checklist for the Description of Biological Sampling and Lab Methods

- Places and circumstances of biological sampling
- Type of tissue, secretions, excretions
- Method of accessing and collecting samples
  - Body site, timing
  - Preparation of subjects for sampling or direct measurement
  - Equipment used during sampling e.g., syringes, intubation, endoscopic equipment, tubes, rectal swabs
  - Drainage, aspiration, biopsy
  - Special considerations around environment of sampling
- Handling of samples before arrival at lab
  - Splitting in subsamples for different purposes: spare sample, samples for different types of analyses
  - Manipulations: centrifuging, addition of reagents, preserving agents
  - Storage: place, timing, duration, equipment e.g., cooler box, fridge, freezer (including temperature of storage at minimum)
  - Dispatch to laboratory: transport means, route, maximum delays, cold chain issues
  - Good Laboratory Practice guidelines followed
- Lab storage conditions of samples until processing e.g., freezer temperature
- Lab analysis method
- Use of lab analysis results to calculate study variables

10.4.3 Measurement Scales for Mental-Behavioral Characteristics

Accurate measurement of mental and behavioral characteristics can be very challenging because direct measurement is often impossible. The researcher is often be forced to resort to an extensive questions-based measurement tool with a series of questions that all measure 'something of' the attribute of interest. Such measurement tools are developed by a method called '*scaling*'. Their relevance and use is well established in psychiatric research. An excellent example of the development of a set of psychometric 'diagnostic tools' is the Composite International Diagnostic Interview (CIDI and CIDI-SF) for eight psychiatric conditions, including major depressive episodes, general anxiety disorder, and others (*See also:* National Comorbidity Survey website). The study of children's mental health is particularly challenging given the need for proxy information. There is a serious risk of underreporting of risk exposures, such as violence or abuse, due to fear of stigma, legal consequences, or denial.

Subjective experiences (attitudes and perceptions) around health-related phenomena have become increasingly popular topics for investigation. An illustration of this is the wide interest in quality of life measurements (next section) and in the measurement of health state preferences in cost-effectiveness analyses. We therefore further discuss scaling and methods of local adaptation of questions-based measurement scales in this section.

Scaling uses methods that have been developed in psychometrics. We briefly describe the phases of a typical scaling exercise. For further introduction, *See:* Howitt and Cramer (2008) and Streiner and Norman (2008). This type of development and adaptation exercise may require a preparatory sub-study prior to use of the final scale in the main epidemiological study. The phases of development are as follows.

10.4.3.1 Phase-1: Designing Questions for Scale Construction

The construction of a new scale starts with designing a series of questions that are all thought to capture something of the underlying attribute. The questions may be selected from a variety of sources, including personal experiences and questions borrowed from existing questionnaires. This is not an exact science. To cite Howitt and Cramer (2008), "Writing appropriate and insightful items to measure psychological characteristics can be regarded as a skill involving a range of talents and abilities." Patients or potential research subjects are excellent sources for creating questions (Streiner and Norman 2008). Focus group discussions and key informant interviews can be helpful in designing relevant and appropriately worded questions about subjective experiences, attitudes, opinions, and knowledge. While borrowing questions from existing sources allows other researchers to perform secondary analysis across surveys and measurements, a word of caution is in order. First, the mere fact that something has been used by others is insufficient proof of quality. Secondly, as discussed below, there are many possible reasons why an existing tool may need adaptation. There exist publicly accessible databases of questions-based measurement scales in certain domains. Table 10.4 lists some examples.

Table 10.4 Selected publicly available sources of questions-based measurement scales for mental-behavioral characteristics

Type of attribute	Public source
Personality aspects	Goldberg et al. 2006 http://ipip.ori.org
Pain, fatigue, emotional distress, physical functioning	Cella et al. 2007 www.nihpromis.org
Quality of life	Generic scales: SF-36 www.sf-36.org/ and www.proqolid.org
Physical activity	International Physical Activity Questionnaire (IPAQ) https://sites.google.com/site/theipaq/
Work performance and health	Health and Work Performance Questionnaire (HPQ) http://www.hcp.med.harvard.edu/hpq

Some thought must go into whether it makes sense to regard the attribute as one-dimensional or multi-dimensional. If the multidimensional nature of the attribute seems obvious and one can clearly conceive the different aspects of it then it becomes necessary to design questions for each of those aspects. One may want to separately measure one or more of these aspects in addition to the overarching attribute. A larger number of questions will need to be devised for the aspect one chooses to document separately. In general at this stage, the more questions the better, because redundant questions are eliminated in the phase of scale construction. When compiling questions about an underlying attribute, the aim is to get as much variation in responses as possible. For example, if a questionnaire intends to test knowledge, there should be a mix of questions with various degrees of difficulty. One should limit opinion questions on issues that nearly everybody will know or strongly agree with, although this cannot always be anticipated. The set of chosen questions should be developed into a questionnaire that can be piloted. When the attribute is an attitude or a perception one should make sure that about half of the questions gauging agreement or disagreement are worded negatively and half of them positively. This is because some people have a tendency to always agree with statements whereas others have a tendency to always disagree.

10.4.3.2 Phase-2: Selecting Questions for Final Scale Construction

Once the questionnaire is devised it should be tested in a group of people similar to future study participants. One should use the data from this exercise to eliminate questions that tend to get the same answer from everybody and questions that many people do not answer. Questions which participants found unclear or overly intrusive should be reworked or dropped. The list can be made shorter still and also more internally consistent by eliminating questions that do not seem to measure the same attribute as the others. This can be done using item-total correlation analysis or factor analysis.

In item-total correlation analysis, one calculates for each question the correlation of the question's response with the total score from the remainder of the questions. Questions that do not correlate with the total score can be eliminated. This is a matter of judgment. What constitutes a 'good enough' question-total correlation is arbitrary (Howitt and Cramer 2008). Kline (1987) has proposed a correlation coefficient cut-off of 0.2 for deciding which questions to eliminate.

Factor analysis is often made use of for scale construction. It is a statistical method that aims to detect virtual underlying factors in a set of variables such as answers to a series of questions about an attribute. Factor analysis can be useful:

- To explore whether the attribute is one-dimensional or multi-dimensional. The conclusion may be that the attribute was one-dimensional if a single factor was detected or that it was multi-dimensional with the different sub-dimensions represented by meaningful factors
- To detect the nature of the underlying attributes represented by the factors. By looking at all the questions that have a high loading on a particular factor and then looking at all those that do *not* have a loading on the same factor one can usually 'see' what kind of attribute is represented by the factor. This is a matter of insight into psychological processes and common sense
- To drop questions that do not seem to load on any relevant underlying factor
- To calculate a factor score for each participant on a dimension that one wishes to further use as an epidemiological study variable

10.4.3.3 Phase-3: From Multiple Questions to a Measurement Scale and a Normative Range

A frequently used approach to integrating the responses to all questions is to give each question a separate score (e.g., 0 or 1 for each yes/no question) and construct a total score based on the summation of all the question scores. When each question gets the same maximum score and the total score is the sum, it is assumed that all questions have the same importance. At times, however, some questions may seem more important than others. This could be based on tacit expert knowledge about what is crucial and what is accessory. Based on this assessment, a weight can be given to each question, and the score of each question is then multiplied by this weight before the total score is calculated. Another form of weighting gives each question a weight proportional to its standardized beta-coefficient in a regression of total scores on question scores (Streiner and Norman 2008). This method is based on the idea that questions explaining more of the variance in total scores should get proportionally more weight. In practice, this form of weighting has been found to rarely impact the total score's ability to predict clinical outcomes known to be related to the attribute (Streiner and Norman 2008). More research is needed in this area. Weighting may also be needed because:

- The attribute is multi-dimensional but the number of questions for each aspect is not in proportion to the perceived relative importance of the aspect
- Some questions are so highly correlated that they can be considered to measure the same aspect, artificially inflating the importance of this aspect in calculating the total score

The next step in scaling is to standardize the raw scale for purposes of comparability among scales and populations. For this it is useful to study the distribution of raw scores in the target population through examining a representative sample of that target population. Another step may be the selection of normative cut-offs and assessment of discriminatory power.

10.4.3.4 Adapting an Existing Questions-Based Measurement Tool to a Local Context

Adaptations of existing measurement tools, using the same approached as outlined above, are frequently needed because of issues with (1) Locally unacceptable or locally invariant questions (2) Locally poorly understood questions; (3) Outdated terminology, (4) Translation, and (5) Issues of different factor loadings in different contexts. Questions use concepts and terms that can bear different meaning and can have different uses in different cultures and languages. They can also acquire different meanings over time. This can complicate translation and local adaptation of questions-based measurement scales (Herdman et al. 1997). Formal permission from the original developers may be needed to facilitate dissemination and use of the adapted instrument by others. The use of appropriate translation procedures to achieve linguistic, dialectal, and cultural appropriateness is also needed. This will often require the involvement of appropriate translators and/or ethnographic experts.

10.4.4 Physical Activity Measurements

Physical activity is a frequently measured behavioral characteristic. In some studies it is possible to quantify certain aspects of physical activity prospectively using a pedometer or accelerometer. The former measures the number of steps an individual takes but cannot distinguish between different intensities of movement (e.g., walking versus running), whereas the latter generally has a greater degree of freedom and can make this distinction by measuring a person's changes in acceleration. Data generated by both devices may need to be standardized to a person's metabolic rate or a proxy measure thereof. Questions-based assessment of physical activity has been greatly facilitated by the IPAQ questionnaire (*See:* Table 10.1). The IPAQ score allows categorization of an individual's daily life into low, moderate, and high levels of physical activity. New technologies such as small Global Positioning System devices are allowing new types of physical activity measurement. A comprehensive list of measures of physical activity is found in Bauman et al. (2006).

This concludes the first part of the chapter, in which we have discussed several aspects of measurement planning. It is not possible to discuss many types of measurement; however, there are two that we wish to explain in greater detail in the rest of the chapter to highlight their increasing importance in health research. The first is the measurement of health related quality of life, and the second is cost measurement.

10.5 The Measurement of Health Related Quality of Life

There are two main objectives motivating the development of instruments to measure and value health related quality of life. The first is to monitor and compare value-adjusted burdens of disease across settings, space and time. The second objective is to measure value-adjusted health improvements from health interventions. Health improvement is a key element in all economic evaluations in which alternative and competing health interventions are compared.

Health related quality of life, sometimes called health state preferences or health utilities, can be evaluated using monetary and non-monetary approaches. The distinction refers not to whether cost is included but to whether health status is converted into dollar estimates. The monetary techniques are not commonly applied in epidemiology or economic evaluation of health interventions, although they can be useful in some situations. In the remainder of the chapter we will therefore mainly focus on the non-monetary approaches of health valuation.

10.5.1 Requirements for Instruments for Valuation of Health Related Quality of Life

There are some general requirements for instruments to measure and value health state preferences to be useful in economic evaluation such as cost-utility analyses. First, the instrument should be able to capture differences and compare changes across diseases and interventions. Second, they should preferably have "ratio scale" i.e., containing a true zero, and "interval properties." 'Interval property' means that a constant change has the same value across the entire scale. A change from 0.2 to 0.3 should in other words have the same value as a change from 0.8 to 0.9. A final requirement for the usefulness of health indices as input to economic valuation is that they should represent the values and preferences of affected individuals or the society.

The question of whose preferences are elicited is critical not only for theoretical reasons but because the numbers obtained can vary widely. Two aspects of preference elicitation are relevant here. First, who is asked? The individuals participating in the preference elicitation exercise could be medical experts, lay persons, or individuals who suffer from the disease in question. This latter category can be further divided into patients who have recently developed the disease, and patients who have accommodated to their current health state. Due to the remarkable ability of the human spirit to accommodate dramatic changes in physical states, asking the same question of these two different groups often yields very different answers. For instance, the self-assessed health status of recent quadriplegics can be much lower than the self-assessed health status of the same individual even a year later. Second, given a particular type of participant in preference elicitation studies, should a societal or individual perspective be taken? Certain types of measurements induce a frame of mind closer to an objective policymaker, whereas others induce the more personal view of a caregiver in the field. Here, again, results vary widely depending on how the question is framed.

10.5.2 Ethical Considerations for Calculating Quality-of-Life

As a society we feel hesitant about putting a quantitative value on human life. Everyone experiences the world differently, and on a fundamental level comparing one person's life to another is an impossible task. The measures used to calculate quality of life certainly do not pretend to be able to encapsulate everything about the human experience.

Despite ethical concerns listed in Panel 10.7, it is often very useful to put a numerical value on human life. In a world of finite resources, we must have a way to make difficult decisions about where best to spend our money. A valuation of health and quality of life enables us to systematically compare across many different diseases, treatments, and health care delivery settings. These valuation methods put a value on the health related quality of people's life years, creating a unit which is commensurable with the length of their lives. This reflects the important fact that people are willing to make trade-offs between those two aspects. It also enables us to assess the relative values of two very different health states, and thus identify potential resource distributions that could maximize societal welfare. In a world where all resources are finite and scarcity is a fact of life, these tools are imperfect but necessary for societal decision-making.

Panel 10.7 Important Ethical Concerns Around Numerical Quality of Life Measures

- Is it morally right to use these calculations to deny anyone a treatment that could extend her or his life, even for a short time or with considerable impairment? In theory everyone agrees that we need to reduce health care spending, but far fewer people actually want those limits imposed on them.
- Some valuations over-weight the young vs. the elderly. Since the elderly tend to have lower quality-of-life by most metrics, a policy focused on maximizing quality-life-years would disproportionately reward resources to the young. The "fair innings" argument is an opposite position, in which it is considered to be more fair to accrue additional weight to loss of health among the young individuals.
- Similarly, when looking at a measure of "cost per quality-of-life," highly prevalent, low-burden conditions may be favored if rounding error occurs in weighting. For instance, if dental caries is given a weight of 0.01 but this value was not exact but simply meant "very small" in the minds of the interviewed, then since many people have dental carries their burden could be disproportionately large.
- These valuations do not measure objective differences in the need for a treatment, but rather how individuals subjectively value different health states. For example, a value measurement could theoretically be the same for fistula surgery vs. cosmetic surgery, but support for spending public funds on the former would be considerably greater than for the latter.

10.5.3 Non-monetary Valuation of Health

In all the non-monetary techniques for measuring health related quality of life, one tries to capture one or several intrinsic elements of "health-related." Exactly what this intrinsic aspect of "health" should be depends on the objectives of the study. This choice of health outcome measure must be done carefully, as it will influence the commensurability and usefulness of the results in a broader context. Commensurability means that health outcomes are measurable by the same standard, which is a prerequisite for comparison with other diseases and interventions. Many health economists therefore consider commensurable health outcome measures to be the gold standard. In other situations, incommensurable measures of health may be sufficient to meet study objectives. An overview of alternative non-monetary outcome measures is given below.

10.5.4 Incommensurable Measures of Health

Incommensurable health outcome measures are useful foremost for comparison within a single disease, but pose restrictions regarding comparability and usefulness in a broader context. They provide valuable information regarding epidemiology and clinical practice, and many are also commonly used as measures of health improvement in cost-effectiveness studies.

Disease incidence or prevalence are typical examples of incommensurable measures. It is not meaningful to compare a study presenting the cost per averted case of tuberculosis with a study reporting the costs per prevented case of malaria. On the other hand, if the study objective was to compare different malaria prevention strategies, malaria incidence would be an appropriate choice of outcome measure and the cost per prevented case of malaria for the alternatives would be highly relevant information for decision-makers.

Survival rates for fatal or non-fatal outcomes are also common incommensurable outcomes in clinical trials. They are crude measures because they do not distinguish well between survival at different ages and because they only bluntly capture differences in disease severity. Survival rates are not useful as health outcome measures in economic evaluation, but may be very useful as intermediate outcomes or for other purposes.

10.5.5 Commensurable Measures of Health

Commensurable measures of health can be applied to a wide range of diseases, including chronic or acute based disease and somatic or psychiatric conditions. The instruments can be one-dimensional, whereby the health state preferences are directly measured. Alternatively, health state preferences can be measured indirectly, using multi-dimensional instruments. Some of the most common approaches are explained in more detail below.

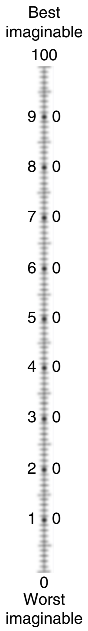


Fig. 10.2 Visual Analogue Scale (VAS). The value 0 is often set to represents the condition "worst imaginable health", whereas 100 represents "best imaginable health." Tha VAS can be employed in many circumstances, as long as answering a question on a 0-to-100 scale makes sense and is appropriate

10.5.5.1 One-Dimensional Health Valuation Instruments

One-dimensional valuation instruments ask participants to report overall health in a single number. Since health is a multidimensional construct, individuals must therefore implicitly weight different aspects of health to provide an answer.

The simplest way to measure health state preferences is to use a Visual Analogue Scale (VAS) (Fig. 10.2). The VAS scale resembles a "thermometer" with values typically from 0 to 100. The value 0 is often set to represent the condition "worst imaginable health," whereas 100 represents "best imaginable health". The respondents, who most commonly are patients, are asked to indicate on the scale how good or bad they consider their health state to be at a specified point of time (e.g., today). The VAS scale is very easy to apply, is usually considered cognitively easy to respond to, and provides results that can be interpreted straightforwardly in the sense that the preference value is given directly from the scale. On the other hand VAS scales are considered to be overly simplistic by many researchers. Because responding does not include any explicit elements of weighting or trade-off, VAS scales tend to result in disease states being weighed as more severe than with other instruments.

With Time Trade-off (TTO) instruments, respondents are asked to hypothetically trade-off a long life with an inferior health state and a shorter life with perfect health. This process starts out with clearly describing the condition in question, including

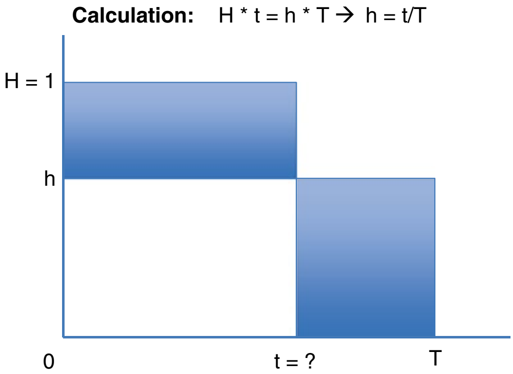


Fig. 10.3 Illustration of how Time Trade-off methods can be used to calculate health related quality of life weights. Respondents are asked to indicate how much of their remaining life expectancy (T) in an inferior health state (h) they would be willing to give up in order to live in the best imaginable health (H) all the time. The size of this time sacrifice is found by equaling the *two shaded areas*, as indicated by the formula

details about different aspects of health such as somatic and psychiatric symptoms, pain and functionality. An example of this type of hypothetical question is:

- "Imagine yourself in the described health state. Given that you would live T years in this health state, how much of the final time would you be willing to give up in order to live in best imaginable state all the time (t)?"

TTO instruments can be well-suited for health state valuation, especially for chronic conditions (Fig. 10.3). Also, respondents are "forced" to weight quality-of-life against duration-of-life; proponents of the method believe this makes the responses more carefully considered. Empirical experience has shown that TTO methods typically rate diseases as less severe than VAS. TTO questions are a bit more cognitively challenging to answer than VAS, which may reduce survey response rates.

With Person Trade-off (PTO) instruments respondents are given hypothetical choices between saving the life of one person and treating N persons with a specified health condition. As for TTO, the health condition in question is first described in detail. An example of a PTO question is:

- "For a given sum of money one may either save the life of one person or prevent N cases of illness X. How great would N have to be for you to consider the two programs equally good?"

Although hypothetical, PTO questions have clear resemblance to policy decisions where limited resources must be prioritized between patient groups. PTO instruments are therefore less appropriate for estimating health state preferences of patients or care takers, but can be used to estimate societal preferences. The health related quality of life weights (h) are given with the formula $h = 1 - 1/N$.

Both TTO and PTO are simplistic in the sense that they ignore a very important aspect of valuation of benefits, namely uncertainty. According to welfare economic theory, people have preferences for risk, which subsequently will influence the value they attach to alternative outcomes.

In Standard Gamble (SG) instruments, risk is included through asking hypothetical questions about preferences for a long life in a specific inferior health state (which must be clearly described) versus a risky intervention that will result in one out of two possible outcomes: the best imaginable health state or death. An example of a SG question is:

- "Imagine yourself in the described health state over T years. You are offered a chance to receive a treatment which will either cure you completely with some probability or lead to your death. What is the lowest probability (p) of a successful outcome you would require in order to choose the intervention rather than living in the described health state?"

The health state preference weights (h) are directly given with the following formula $h = p$, where p is the probability from the standard gamble.

By incorporating uncertainty, SG is the only of these four instruments for direct valuation of health state preferences that is consistent with welfare economic theory. Despite this advantage, the method has not become very popular, primarily because such questions are cognitively rather demanding to answer. In particular, people tend to mis-assess small risks and treat losses and gains differently (Kahneman and Tversky 1979).

10.5.5.2 Multi-dimensional Health Valuation Instruments and Summary Measures of Health

With the one-dimensional instruments described above, health state preferences were directly assessed by asking questions about "health" as such. Although these approaches are computationally simple, we have claimed that they either tend to be overly simplistic (VAS), or cognitively demanding for respondents (TTO, PTO and SG). These are among the reasons why it is common to see multi-dimensional instruments in empiric research on health state preferences. By dividing "health" into several sub-dimensions (e.g., pain, physical functioning, psychological state, etc.), each with independent response alternatives, multi-dimensional instruments are generally easier for patients and others to respond to. To create a summary measure of overall health, a weighting between the dimensions of health must be established.

Summary measures of overall health take two forms: "health gain" and "health gap" measures. Health gain measures assess how much "health" or "quality of life" is added by an intervention or policy. Health gap measures assume a baseline state of "ideal health" and assess how the gap between someone's actual health state and the ideal health state shrinks following an intervention or policy. These are two conceptually different but numerically related ways of measuring health. Figure 10.4 shows the relationship between the two. Health gain measures put no limit on how much health someone can achieve; health gap measures assume a limit (an "ideal" state of health).

The most common health gain measure used is the quality-adjusted life-year (QALY), and the most common health gap measure is the disability-adjusted life year (DALY). The World Health Organization measures the global burden of disease using DALYs, whereas cost-effectiveness studies almost universally use QALYs.

A multi-dimensional quality-of-life metric mathematically represents the overall weighted value of all relevant dimensions of health. For instance, a minor tooth ache might be given a disutility of 0.05 (or 0.95 of full health), whereas a

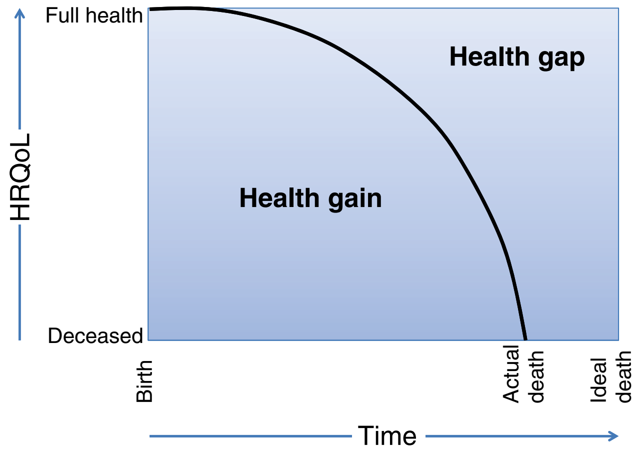


Fig. 10.4 The relationship between health-related quality of life (HRQoL) and life expectancy (Time) can be used to illustrate the difference between health achievements that health systems seek to maximize, and health gaps that one wishes to minimize

painful chronic illness might be given a disutility of 0.7 (0.3 of full health). In order to provide a generic model in which we might investigate the different types of multi-dimensional metrics which exist, we provide the following basic formula for the common (linear) case:

Multidimensional quality of life of individual *i*

$$\mathrm{Q}_i = w^1 c_i^1 + w^2 c_i^2 + \ldots + w^n c_i^n$$

Where:
Superscripts refer to the j^{th} condition
Subscripts refer to the i^{th} individual
c^j_i = a "health profile": a numeric measure of the extent to which individual *i* has the j^{th} condition. This could be an indicator variable (e.g., presence of blindness) or a measurement on a continuous scale (e.g., mobility)
w^j = the linear weight given to the *j*th condition

Weighting or value functions such as those described above are used to transform the health profiles into health indices, a single numerical value representing the health state preference of the patient population in question. Preference basis is a prerequisite for the validity of the estimates as input in for example economic evaluation of health interventions. For an estimate to be valid in economic evaluation – for example as input for a quality-adjusted life year (QALY) calculation – the health profiles and indices should have generic properties, i.e., they should be applicable to all types of health conditions and patients groups. When this is the case, the instruments are commensurable. Non-commensurable instruments cannot be used to calculate QALYs, and are far less applicable in economic evaluation. In the next paragraphs we explain how health profiles and indices are generated for two popular multi-dimensional generic instruments.

Perhaps the most commonly used multi-dimensional instrument is the EQ-5D. It was developed by the Euroqol group, and as the name indicates it has five sub-dimensions of health: Mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension has three severity levels: No problems (level 1), some or moderate problems (level 2), and extreme problems (level 3). The exact wording of each level is customized to each dimension being measured. This makes the instrument very simple for patients to respond to. With five dimensions and three levels, EQ-5D has 243 potential health profiles, although some combinations are not very likely. A patient with moderate anxiety/depression, and no other problems, has the following EQ-5D health profile: (1 1 1 1 2). Someone confined to bed (severe mobility problems), having some problems with washing and dressing (self-care), but otherwise having no problems, has the following profile: (3 2 1 1 1).

EQ-5D has an additive weighting function. The starting point is health profile (1 1 1 1 1), which represents the best possible health state with an index value of 1.0. For each health dimension where the level deviates from 1, a certain value is subtracted from 1. More value is subtracted for level 3 than for level 2 scores, and the more dimensions that deviate from level 1 the lower is the remaining health index value. Value sets have been developed for several countries, and are important for the subsequent size of the health index. Robberstad and Olsen (2010) illustrate how the use of a UK value set produced health state preferences for severe AIDS that were much lower than when a value set from Zimbabwe was applied.

EQ-5D is considered to be easy to use, to be well tested and reliable, and the instrument has a well-defined protocol for adaptation and use. The instrument is criticized by some for being a bit "blunt," with relatively few dimensions and levels. In particular, many consider the instrument to be invalid for assessment of minor health decrements, and that this leads to non-severe health conditions being rated too severely. In the literature, this is referred to as a "ceiling effect." In 2012, a version of EQ-5D was launched with five rather than three levels for each dimension. Potentially, five levels might reduce sensitivity challenges and ceiling effects of EQ-5D, but the empirical evidence regarding this is still limited.

The Health Utilities Index Mark 2 (HUI2) has seven dimensions of health, with 3–5 levels within each health dimension. These are Sensory (four levels), Mobility (five levels), Emotion (five levels), Cognitive (four levels), Self-Care (four levels), Pain (five levels) and Fertility (three levels). With a maximum of 24,000 hypothetical health states, it can be argued that the instrument is more suitable to capture minor health changes than the less detailed EQ-5D. The health index value is calculated using a multiplicative scoring function. The starting point is 1.06 (best possible health), corresponding to health profile (1 1 1 1 1 1 1). For each dimension, this value is multiplied with a score between 0 and 1 depending on the level.

HUI2 has some apparent attractiveness over EQ-5D in terms of level of detail, but the instrument has fewer translations and is not as well validated as the latter. Other commonly used descriptive systems have adopted even more detailed classifications of "health." For example, the Australian AQoL (Assessment of Quality of Life) has five main dimensions of health (illness, independent living, social relationships, physical senses and psychological wellbeing), each with three sub-dimensions and four levels. The Finish 15D has 15 dimensions with five levels for each, and thus a total number of hypothetical health states exceeding 30 million. An overview of available instruments is provided by the Quality of Life Instruments Database (www.proqolid.org).

10.5.6 Monetary Valuation of Health

As mentioned, monetary valuation techniques have been less dominant in practice, partly due to measurement challenges and partly because measuring health benefits in monetary terms has not communicated well with the health care disciplines nor health care decision makers. Briefly, there are two main approaches to monetary valuation of health that both relate to how much people are willing to pay to achieve health improvements. The underlying assumption of the first approach is that people implicitly prioritize and value health together with other commodities they need or desire, and that the value of health thus can be estimated through observing how much they actually pay for various health improvements or how much health they give up in exchange for undertaking risk. Classic studies in this stream of research have used the additional wage required to be paid to employees to take on a riskier job.

The other more commonly applied approach is to ask people questions about how much they would be willing to pay for certain types of health care or health benefits. Several techniques have been developed to elicit willingness to pay, broadly labeled as "contingent valuation." Questions can be open-ended: How much would you be willing to pay? Or questions can be closed-ended: Would you be willing to pay X to achieve Y? Questions can be organized as bidding games, where values are changed upward or downward depending on the previous answer; or through payment cards, where alternative values are suggested ranging from zero to a value assumed to exceed a realistic maximum, and people are asked for the maximum they are willing to pay.

10.6 Cost Measurement

In Chap. 6, you were briefly introduced to two categories of costing studies, namely "cost-of -illness" and "cost-of-intervention" studies. As the names indicate, the former is concerned about the cost imposed by disease on different parties of society, while

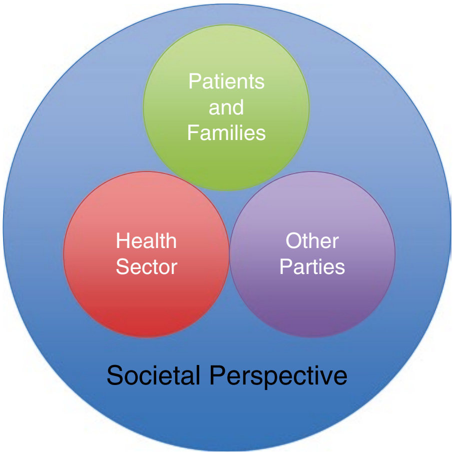


Fig. 10.5 Venn diagram of possible perspectives for economic evaluation

the latter is concerned with the costs related to preventing or treating the illness. While these objectives are clearly distinct, the two types of studies share some methodological issues that are most reasonably dealt with simultaneously. Below we discuss how costing studies should be designed to be useful in planning and decision making processes, or as input into further research such as full economic evaluations.

All costing exercises involve three common steps: (1) Identification, (2) Measurement and (3) Valuation. It is important for transparency and reproducibility that these steps are performed explicitly, and that measurement and valuation results are reported separately. It is a common mistake in costing exercises that cost estimates are not reported disaggregated into quantities (units) and unit prices. This makes it difficult for readers to assess the validity of the results, reduces the usefulness of the results for e.g., budgeting purposes, and precludes translation of the results into different settings (where unit prices may be different). Presenting quantities and unit prices separately is called the *ingredient approach*. Below, we discuss important aspects of the three steps of costing.

10.6.1 Costing Step-1: Identification

Before it is possible to consider the quantities and values of various cost items, we need to consider and justify a list of items that should be included in the analysis. This *identification* exercise is far from trivial, and will strongly affect the downstream results and the range of conclusions that can be drawn from the study. For example, the analyst needs to decide from whose point of view the costs should be 'considered, i.e.' "the perspective" of the study. These can be broadly categorized into *health sector*, *patients and families* and *other parties* (Fig. 10.5). Sometimes *productivity losses* are categorized as a separate perspective, while others will see this as a sub-category of the perspectives previously mentioned. When all these perspectives are combined, the costing exercise is done from a *societal perspective*.

10.6.1.1 Study Perspective

Disease imposes costs on the health sector. Because planning of health care services is a very common motivation for doing cost studies, it is rare to see costing exercises that exclude this perspective. There is, however, great variation in how the health sector is defined in applied studies. The *point of care* is usually included, e.g. the district hospital or primary health services that actually deliver the health care to patients. The degree to which up-stream levels of the health sector are included varies considerably. Horizontal health care programs involve sector-wide planning, co-ordination, training and monitoring at district, regional and national levels. Vertical programs involve the same processes, although differently since they are typically organized through international donor-based activities. It is quite common to see that costing exercises fail to include the full range of up-stream costs. Such studies can be insufficient as input for budgeting of e.g. scale-up exercises. If they are used naively, they will result in budget deficits and subsequent implementation problems for the new activities.

Arguably, illness always entails costs to patients and their families. Typically, illness will impose *direct costs* to pay for treatment, drugs and transport. User fees and drug costs are important determinants for health seeking behavior, especially in low-income situations, and better knowledge of such factors may be of great relevance. In addition, illness typically is associated with *indirect costs* in terms of reduced ability to work for the patient, or because family members must divert efforts from their usual activities to take care of the patient. Malaria is a typical low-income disease resulting in acute illness as well as in high prevalence of chronic anemia and resultant fatigue. Therefore, malaria imposes indirect costs for patients and their families in terms of reduced ability to perform subsistence activities both in the short and longer terms. Generally, chronic illnesses represent substantial costs to patients and families, both direct and indirect, and reducing these costs through illness control efforts can therefore be valuable to society.

In addition to the health sector and patients/families, other sectors are also commonly involved. *Third party payers* have important roles in funding health care, and are thus influenced by the occurrence and management of disease. Examples of third party payers are private or collective insurance schemes, which are common in most high income settings and increasingly important in low and middle income settings. Different types of governmental social insurance schemes may exist outside the health sector.

As mentioned, *productivity losses* include indirect costs incurred on patients and their families. Illness does however affect the society more broadly. It will for example affect employers in the short term through reduced production, and in the longer term through increased costs for recruiting and training of employers to replace those who are ill. It will affect governmental income, through reduced taxes from both the employers and the employed, and subsequently morbidity will affect national income and overall economic development. Other effects are far more difficult to estimate, and they are therefore often pragmatically ignored in costing exercises. Nutritional disorders, such as iodine deficiency, are for example known to affect the cognitive development and learning abilities of school children. But the down-stream effect of this on personal development and national income is extremely difficult to estimate. In such situations, consider presenting sensitivity analyses based on different assumptions about down-stream effects.

The choice of study perspective should be decided by the research question and the objectives of the study. A societal perspective provides the most comprehensive and complete picture of costs related to morbidity and its management. In low-income settings the governmental expenditure on health care typically represents around 50 % of the total health care expenditures, while private money and to a varying degree insurance schemes represent most of the rest. Costing exercises that only focus on the public health sector costs therefore very poorly represent societal costs in such settings. In high income countries the governmental share of total health care expenses is typically much higher, with some important exceptions including the USA. It can be argued that social planners should be concerned about total welfare in society, and that the societal perspective therefore should be applied for all types of social planning. Indeed, prominent guidelines from academics and journals single out the societal perspective as the appropriate one. However, identifying, measuring and valuing all societal costs is a costly and demanding exercise. With a narrower study objective, a more narrow costing perspective can therefore sometimes be defended. If for example the objective of the costing study is to improve internal organization of hospital services, a more narrow health sector perspective might be justified. In this case, since the objective relates to hospital budgets, other societal costs are *ex ante* assumed not to matter for the decision.

10.6.1.2 Cost-of-Illness Versus Cost-of-Intervention

Since health interventions usually are delivered by the health system, the health systems perspective naturally becomes the core element of most *cost-of-intervention* studies. This does not rule out the relevance of other perspectives, since different health interventions may influence stakeholders differently. An example is the choice between facility based and community based services for tuberculosis patients. A facility based service is typically more costly to patients who regularly have to travel to a hospital to receive treatment and follow-up. Community based services on the other hand require an extension service program, and are typically more costly to the health care provider. Inclusion of patient costs may in this situation represent important information for the health care planners. *Cost-of-illness* studies aim at estimating the economic burden of specific morbidities to society and should therefore have a broader perspective than simply focusing on the intervention costs. The exact framing of a study should be based on the nature of the disease.

Cost items can be categorized according to activities (e.g. administration, training, patient treatment), according to input categories (e.g. salaries, medical supplies) or according to organizational level (e.g. facility, district, region, national level). What is more practical depends on the nature of the study question, but it is important that all relevant cost categories are covered to avoid underestimation, and that the categories are not overlapping, which may result in double counting.

10.6.2 Costing Step-2: Measurement

Once all relevant cost-items have been identified, the second step of a costing process is to *measure* the quantities of each. Broadly speaking, measurement can be done *prospectively*, for example alongside a clinical trial, or *retrospectively*, through modeling.

Measuring resources used by a health care activity alongside a clinical trial has several advantages. With this prospective approach it is possible to monitor resource use relatively accurately. The trial situation is a good opportunity to expand existing data collection tools to include information for example about how many hours of work different categories of health personnel use to perform the various activities. One can track patients through the different procedures and accurately take account of the time and resource consumption associated with their treatment. While prospective costing can yield accurate estimates of resource use in a particular setting, thus representing high *internal validity*, the *external validity* is not necessarily good. This is because organizational factors, populations and epidemiology differ substantially between settings, and resource use in one setting therefore is not necessarily representative of another setting. The simple fact that a clinical trial in itself represents a special case, typically with more resources and higher standards of care than what is common in a country, calls for caution. In a study on the costs of breastfeeding promotion in Uganda, Chola et al. (2011) describe how estimates on resource use from a clinical trial were adjusted to better represent resource use in an assumed national scale up of the intervention.

While prospective costing alongside clinical trials has several advantages in terms of accuracy, the most common approach in applied economic evaluation is to measure resource use retrospectively. In many cases this is a consequence of clinical trial designs failing to properly include costing aspects during planning and implementation. In other cases researchers aim for results that are more generic and that can be transferred and used in broader contexts. The wider implementation of a trial's test intervention has organizational and resource use implications at many levels of society, including for the patients, that are not captured through the core design of clinical trials. In both cases modeling of resource use can be helpful.

Modeling, in short, implies that clinical, epidemiological, socio-economic and institutional factors are considered jointly. The relationship between these factors and how they influence resource use is based on assumptions that should be based on best available evidence, which can come from a variety of sources. Robberstad et al. (2011) illustrate how different sources including registry data, evidence from clinical trials and cost databases can be combined to model resource use and societal costs of morbidity preventable by pneumococcal vaccines in Norway. Modeling introduces a great level of flexibility, and is applicable in most situations. But different sources of evidence must be combined cautiously since they may represent situations that are not compatible.

10.6.3 Costing Step-3: Valuation

The third step in costing processes is *valuation*, which should always be based on the concept of *opportunity cost* (sometimes called alternative cost). The opportunity cost concept is fundamental in all economic thinking. It captures the idea that the true cost of a resource is represented by what we have to give up when we choose to use it. It is a simple fact that when we choose to spend a sum of money on an activity, the same amount of money is no longer available for other activities. According to economic theory it is the value of these foregone activities that represent the true cost of an activity, and this value is not necessarily the same as the amount of money paid.

Opportunity costs are straightforward to estimate for commodities or resources that can be bought in well-functioning markets, in which case the opportunity cost is reflected through the market prices. Good examples of this are commodities such as stationary, furniture and fuel. The opportunity cost of capital (money) can likewise be extracted from the financial markets, for example the interest rate one has to pay to lend money to fund the necessary investments. There are two common situations when market prices are no longer good proxies for opportunity cost; the first is for non-market goods, and the second is when the markets are regulated through taxes or subsidies. We will start with a brief discussion about adjustment of taxes and subsidies.

Taxes are important sources of income for both low- and high-income countries, and a sales tax or value-added tax (VAT) increases the price the consumers have to pay for certain goods and services. Whether or not VAT should be included as a cost depends, however, on the perspective of the analysis. From a societal perspective, where all parties in society are taken into account, VAT is but a transfer of resources from one party to another (usually the government). From a societal point of view VAT is thus not a cost, and it should be deducted from the price of goods and services to reflect opportunity cost. If the perspective is narrower, e.g. a pure health care provider perspective, it may be correct to include VAT as a cost since this reflects the situation of e.g. the hospital in question.

Subsidies are also common in the health care sector. In many ways subsidies are negative taxes, but whereas taxation is usually done by governments, subsidies are commonly provided also by non-governmental organizations. For example, the prices of antiretroviral drugs to treat AIDS are usually strongly subsidized in low income countries. Leading pharmaceutical companies now provide their products at 90 % discount to low-income high-prevalence countries, compared to their prices in high-income countries. This represents a kind of cross-subsidization where buyers in the best-off countries subsidize buyers in the worst-off countries. Like taxes, subsidies should be dealt with in economic analyses depending on the perspective. In a societal perspective, subsidies are merely a transfer of resources between different parties and do not represent cost-reduction. From a societal perspective, prices should therefore not be adjusted for subsidies.

While the size of subsidies and taxes is relatively easy to observe, the true opportunity cost of non-market goods is often much more difficult to assess. What is, for example, the opportunity cost of the time a mother uses to bring her sick child to the hospital? What is the value of volunteer time, which is important for the running of many community health services? For patients and care takers who are employed, the valuation of time use can be based on the wage rates. This is called the *human capital approach*, which is considered to be a good approximation in a short time horizon. With a longer time horizon, work absenteeism can be compensated through recruitment and training of new employees, in which case the *friction cost approach* may be more appropriate. For self-employed people, time is often a non-market good, and it may be necessary to consider a *shadow wage rate* that represents the value of the lost production. For e.g. subsistence farmers, the consequences for crop production may be much higher in the wet-season, when planting and weeding require attention, compared to the dry season. A pragmatic approach is to apply governmental minimum wage rates as proxies for the value of time, but the validity of such a proxy will vary with local circumstances.

Above we have reflected on some issues regarding valuation of resource use. The bottom line is that market prices do not always reflect the opportunity costs of resource use, in which case it may be necessary to make price adjustments or to value the resource use with alternative techniques. When this is adequately done, the analysis represents the *economic costs*. It is, however, also common in economic analyses to present unadjusted prices, in which case the results are the *financial costs* of an activity or a disease. While economic costing is appropriate when considering resource allocation (prioritization), financial costing has a role to play for budgeting purposes. In the following paragraphs we will briefly look at a few cross-cutting topics that are important in costing exercises.

10.6.3.1 Cost Items that Are Durable

Some types of cost items are consumed continuously throughout the period of an activity. A good example is pharmaceutical drugs that need continuous purchasing and re-stocking. Such consumables are called *recurrent* cost items, and they should be valued and allocated directly to the point of time in which they occur. Other costs items represent investments that should last for several years. Good examples are building facilities, vehicles and expensive diagnostic equipment. Such durables are labeled *capital* cost items, and they need to be treated differently from recurrent cost items in costing exercises to reflect the true opportunity costs of undertaking the activity.

Capital goods represent two types of costs: (1) the *opportunity cost* of the money invested to purchase the item, and (2) *depreciation*. The opportunity cost of the investment can commonly be reflected by the interest rate for money in the mortgage market. The money invested in e.g. a new car could have been put in a bank account and yielded interest that could have been used for other purposes, such as better coverage of essential drugs. Alternatively, if the money to buy the car needs to be lent, interest will have to be paid to the bank. In any case, the interest represents a stream of cost that must be included in the economic assessment of the activity. Depreciation, on the other hand, represents the tear and wear on the equipment. Capital goods have a limited life time, and during the lifetime the value of the item is gradually reduced until eventually it is zero. The stream of opportunity and depreciation costs represents the capital cost of an item, and the process of calculation is called *annuitization*. When the stream of costs is calculated as constant annual values, the result is called *equivalent annual value* (E), calculated as:

$$E = \frac{K - \frac{S}{(1+r)^n}}{A(n, r)}$$

Where K denotes purchase price, S is the value at the end of the period, A is the annuity factor, r is the discount rate, and n is the useful life of the equipment.

Equivalent annual values can conveniently be calculated by using a spreadsheet:

Equivalent Annual Value (E) Calculation
Excel: =-PMT(r,n,K)
Lotus 1-2-3: @PMT(K, r, n)

10.6.3.2 Costs and Quantities of Output

Above you were introduced to the difference between recurrent and capital costs and how to deal with cost items that last for more than a year. Another important dimension of costing is to consider how the costs are likely to change when the level of output of an activity is increased or decreased. What are for example the cost implications of increasing the number of patients that are treated in a certain program? To answer this, it is necessary to understand the concepts of *fixed* and *variable costs*, and how they can produce information about *marginal costs* that is crucial for budgeting purposes and for consideration about how much a health care provider should offer for a service.

Fixed costs do not depend on the level of output. They are constants that do not change with output, at least not within the limits that are relevant in the context of the analysis. In order to provide a service, a facility is usually required, and this facility needs electricity for lighting and insurance. These costs are the same if the facility is used to treat 10 patients or if 100 patients are treated in a day – they are, in other words, fixed (at least in the short term – a facility that consistently sees 100 patients a day might need to build a new building to accommodate the demand). By increasing from 10 to 100 patients there are, however, other costs that will increase. For example, ten times as many drugs and other consumables are likely to be required. These cost items, which depend on the level of output, represent the variable costs of service provision. The sum of fixed and variable costs is the *total cost.*

When the variable and fixed costs are known for different levels of output, a number of useful calculations and projections can be made. An apparent outcome is total average cost (total costs/total number of output), for example the total cost per patient. This is useful information for budgeting of an activity that will deliver a certain level of output. A related outcome is average variable cost (variable costs/ total number of output). Average unit costs are, however, insufficient information to consider the appropriate level of service provision. Appropriate assessment of level of service provision, i.e. the level of scale-up, requires information about *marginal costs*, defined as the change of total costs when the level of output is increased or decreased by one unit.

Marginal costs bear particular significance for all profit maximizing actors, who according to economic theory will increase production until marginal costs equal marginal revenue. Public health service providers are usually not profit-maximizers but it is still relevant to consider the marginal costs for budgeting purposes. In addition, such evidence is highly relevant as input in full economic evaluations, because marginal costs are likely to affect the cost-effectiveness at different levels of output. Essential childhood vaccines are for example highly cost-effective in most situations, and according to demographic health surveillances the coverage is generally high (70–90 %), including in low income countries. Increasing the immunization coverage towards 100 % is, however, not necessarily cost-effective for two reasons. First, the marginal cost of reaching the last mother-baby pair is likely to be much higher than the average cost since it is increasingly difficult to reach them. This could be due to poor infrastructure in the most remote areas or information challenges to reach the least motivated mothers. Second, the marginal benefits are likely to be smaller for the last individuals due to herd immunity. In sum, cost-effectiveness is generally likely to decrease with immunization coverage.

10.6.3.3 Future Costs

Rather than occurring within one year, most health projects and diseases involve streams of costs ranging over several years, often the entire life-time of patients. This generates some challenges, since costs in one year are not directly comparable to costs in another year. This phenomenon is caused by two factors: (1) inflation, by which the real value of currencies decrease over time, and (2) the time value of money, whereby individuals prefer consumption today to consumption tomorrow and therefore demand a real interest rate as compensation for that deferred consumption (the opportunity cost). In effect, the purchasing power of e.g., one Euro is diminishing with time, and so is the opportunity cost. In order to make costs occurring at different points of time comparable, it is necessary to calculate *present values*, and the technique for doing this is called *discounting*. The following formula calculates the present value (PV) of a stream of future values (FV) for a number of individual years (t = 1, 2, … n), where r denotes the discount rate.

$$\mathrm{PV} = \sum_{t=1}^{n} \frac{\mathrm{FV}}{(1+\mathrm{r})^{\mathrm{t}}}$$

All costs should be discounted at the same rate. In addition, in order to be consistent with economic welfare theory, all health benefits should be discounted at the same rate. In practice, opposition to discounting health benefits can be significant, and several cost-effectiveness guidelines therefore recommend that health benefits be included and presented both discounted and undiscounted. If health benefits are not discounted, considerable care should be exercised to make sure that nonsensical results do not result. For instance, consider the case of the varicella vaccine, which is delivered to children but helps prevent shingles in the elderly. Ignoring contagion effects (which are substantial) we have an intervention which costs money now to prevent health outcomes five to six decades later. If the health benefits are not discounted by time, the vaccine will appear to be much more valuable than is sensible. In practice, the pharmaceutical firm which developed it will price it significantly higher than welfare theory would suggest is reasonable, in order to capture governmental willingness to pay for the vaccine based on the study. Individuals seeking to purchase the vaccine through unsubsidized channels, however, will see too high a price relative to benefits in their own view (which likely discounts future health), and will be unlikely to purchase it.

In this chapter we discussed the planning of measurements. The goal of this planning is to establish optimal measurement procedures and schedules in terms of intrinsic validity, error avoidance, efficiency, and respect for persons. To achieve this goal, one requires a twin plan that focuses on verifying and, if necessary, adjusting the planned measurement procedures and schedules once they become functional. In other words, there is a need for a quality control plan. Chapter 11 deals with this topic, placing quality control (QC) in the wider framework of study quality assurance (QA).

References

Bauman A et al (2006) Physical activity measurement – a primer for health promotion. IUHPE – Promot Educ 13:92–103

Bland RM et al (2003) Maternal recall of exclusive breastfeeding duration. Arch Dis Child 88:778–783

Bland RM et al (2004) The use of non-prescribed medication in the first three months of life in rural South Africa. Trop Med Int Health 9:118–124

Cella D et al (2007) The patient-reported outcomes measurement information system (PROMIS): progress of an NIH roadmap cooperative group during its first two years. Med Care 45(Suppl 1):S3–S11

Chobanian AV et al (2003) JNC 7 complete version. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. Hypertension 42:1206–1252

Chola L et al (2011) Cost of individual peer counseling for the promotion of exclusive breastfeeding in Uganda. BMC Cost-Eff Resour Alloc 9:11

CIDI/Composite International Diagnostic Interview. www.CRUfAD.org. Accessed Sept 2012

Frankovich J, Longhurst CA, Sutherland SM (2011) Evidence-based medicine in the EMR era. NEJM 365:1758–1759

Ghandi M et al (2009) Protease inhibitor levels in hair strongly predict virological response to treatment. AIDS 23:471–478

Goldberg LR et al (2006) The international personality item pool and the future of public-domain personality measures. J Res Personal 40:84–96

Growth Analyser (2009) Version 3.0 (Application). Dutch Growth Foundation, Rotterdam. www.growthanalyser.org. Accessed Sept 2012

Herdman M, Fox-Rushby J, Badia X (1997) 'Equivalence' and the translation and adaptation of health-related quality of life questionnaires. Qual Life Res 6:237–247

Herold JM (2008) Surveys and sampling. In: Gregg M (ed) Field epidemiology. Oxford University Press, Oxford, pp 97–117. ISBN 9780195313802

Howitt D, Cramer D (2008) Introduction to research methods in psychology, 2nd edn. Prentice Hall, Harlow, pp 1–439. ISBN 9780132051637

HPQ/The World Health Organization Health and Work Performance Questionnaire (HPQ). http://www.hcp.med.harvard.edu/hpq. Accessed Sept 2012

IPAQ Group. International Physical Activity Questionnaire. https://sites.google.com/site/theipaq/. Accessed Sept 2012

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:263–292

Kline PA (1987) Handbook of test construction. Routledge, London, pp 1–250. ISBN 9780416394306

Lohman TG, Roche AF, Martorell R (1988) Anthropometric standardization reference manual. Books on Demand. Human Kinetics Books, Champaign, pp 1–183. ISBN 060807070X

Miettinen OS (1985) Theoretical epidemiology. Delmar, New York, pp 1–359. ISBN 0827343132

National Comorbidity Website. http://www.hcp.med.harvard.edu/ncs/scales.php. Accessed Sept 2012

OpenXdata (2010) Open-source software for data collection. www.openxdata.org. Accessed Sept 2012

Perloff D et al (2001) Human blood pressure determination by sphygmomanometry. American Heart Association, Dallas, pp 1–33

Pickering TG et al (2005) Recommendations for blood pressure measurement in humans and experimental animals. Part 1: blood pressure measurement in humans. Hypertension 45:142–161

PROQoLID. Patient reported outcome and quality of life instruments database. Mapi Research Institute, 2001–2012. www.proqolid.org. Accessed Sept 2012

Robberstad B, Olsen JA (2010) The health related quality of life of people living with HIV/AIDS in sub-Saharan Africa – a literature review and focus group study. Cost Eff Resour Alloc 8:5

Robberstad B et al (2011) Economic evaluation of second generation pneumococcal conjugate vaccines in Norway. Vaccine 29:8564–8574

Shea SA et al (2011) Existence of an endogenous circadian blood pressure rhythm in humans that peaks in the evening. Circ Res 108:980–984

Streiner DL, Norman GR (2008) Health measurement scales. A practical guide to their development and use, 4th edn. Oxford University Press, Oxford, pp 1–431. ISBN 9780199231881

Vital Wave Consulting (2009) mHealth for development. The opportunity of mobile technology for healthcare in the developing world. UN Foundation and Vodaphone Foundation Partnership, Washington DC/Berkshire

White E, Armstrong BK, Saracci R (2008) Principles of exposure measurement in epidemiology. Collecting, evaluating, and improving measures of disease risk factors, 2nd edn. Oxford University Press, Oxford, pp 1–428. ISBN 9780198509851

World Health Organization (1995) The use and interpretation of anthropometry. WHO, Geneva, pp 1–452. ISBN 9241208546

The Quality Assurance and Control Plan 11

Jan Van den Broeck and Jonathan R. Brestoff

Only two things are infinite: the universe and human stupidity.
And I am not sure about the former.

Albert Einstein

Abstract

Quality assurance relates to all actions taken to ensure respect for general epidemiological principles. Consequently, quality assurance includes many aspects of study design and conduct, including quality control activities. Quality control (QC) relates to the monitoring and documentation of the validity and efficiency of study procedures and, if necessary, actions for adapting procedures or improving adherence to them. Ultimately, the purpose of QC is to achieve optimal data quality. In this chapter we outline QC methods and tools, prime among them being the monitoring of measurement error and other factors that could bias the statistical results of a study.

11.1 Quality Assurance

11.1.1 Minimum Data Quality

A fundamental question that is asked about every study is this: are the study data of high enough quality to make inferences about the target population? This question

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_11,

is ultimately about the success of *quality assurance* (QA), a broad concept that encompasses all actions taken to increase the internal and external validity of the study as well as to ensure the study's ethical value. High quality data are valid and useful for making inferences. In fact, even slight suspicion that data are *not* of good quality (e.g., by being non-attributable or construed) can prohibit any inferences from being made. In order for data to be of high quality, they must have some minimum characteristics:

- Each data point must be original (not construed), legible, and verifiably attributable to a particular observation unit and source document
- Data used for analysis should be accurate enough and complete enough; Required minimum levels of accuracy and completeness are rather study-specific and need specification in the study planning, as they depend on ambition, availability of tools and other resources, specific aims, and study design

To increase the likelihood that data will meet the specified minimum standards (an ethical obligation), it is necessary to create a quality assurance and control (QA/QC) plan. Such a plan includes QC procedures for monitoring the validity and efficiency of study procedures and, if necessary, for adapting these procedures to improve adherence to them.

The focus of the remainder of this chapter will be on quality control only. Quality control functions and methods will be reviewed, after which the focus will be on (patterns of) measurement error and on how individual and group performance in avoidance of measurement error can be monitored and steered in the ongoing study (*See:* Panel 11.1 for terminology).

Panel 11.1 Selected Terms and Concepts Around Quality Assurance and Control

Accuracy Proximity to true value (mostly used in respect of measurement values or estimates)

Adherence Implementation according to protocol or guideline

Bias Deviation from true value

Data quality The accuracy and completeness of data as well as their confidentiality, and their verifiable originality and attributability (i.e., their belonging to particular measurements of particular observation units)

Error A departure from what is correct (from Latin *errare* = wandering off the correct path)

Gold standard Test or measurement procedure considered to yield results that are valid enough to serve as a measure of truth against which the results of other tests/measurements can be compared to determine their validity#

Instrument error Inaccuracy in measurement value due to a fault inherent in the measuring instrument, not due to the observer or subject#

Misclassification Act or result of an act of attributing a wrong level of a categorical variable to one or more observation units

(continued)

Panel 11.1 (continued)

Observer accuracy Degree to which an observer tends to measure accurately when using an accurate instrument

Observer precision Degree to which an observer, when making independent replicate measurements with an accurate instrument, tends to obtain values that are close to each other

Precision (of measurement) Degree of agreement among a set of replicate measurement values obtained using the same accurate instrument (*Syn*: Reproducibility)

Protocol violation Not executing, or applying a different or altered procedure than the procedure prescribed by the official study protocol

Quality assurance Study activities aimed at optimizing and maintaining data quality

Quality control Screening, diagnosing and, if possible, correcting problems with the performance of individuals, procedures and systems involved in a research study

Replicate measurements Measurements repeated independently within an interval considered short enough to assume that no measurable change of the underlying dimension has occurred

[#] Definition contributed by Douladel Willie

11.1.2 Quality Control Functions and Methods

Quality control has some specific functions, as listed in Panel 11.2. These functions are ultimately aimed at preventing, detecting, and correcting human error caused by factors affecting personal performance or inadequacies and failures of systems and procedures. Because QC is ultimately about the sensitive issue of human error caused by study staff, QC planning and implementation is a very delicate task. 'People skills' are therefore important for those who plan and implement QC activities. Panel 11.3 offers some selected general recommendations on approaching QC issues.

11.1.2.1 Quality Control Tools

For the screening and diagnostic function of QC, a number of tools are available to the designer of the QA/QC plan (Table 11.1). All these tools are commonly employed, as they tend to complement each other well. What needs to be screened and diagnosed is triple: (1) individual errors occurring in routine study procedures, (2) overall performance of systems, and (3) individuals' problem solving capacities in special (not routine) circumstances. Self-checks and supervisor checks are usually the most important field methods. For example, interviewers themselves should check the completeness and correctness of recorded data before submission to the supervisors. Preferably, such self-checks should be done in the measurement setting, so that immediate corrective action can be taken. Supervisors must then check the

Panel 11.2 Functions of Quality Control

Screening function

- Check to what extent prescribed procedures are followed
- Check reliability and efficiency of systems put in place
- Check whether personnel attempt to respect general epidemiological principles in situations for which no detailed written guidelines were made available

Diagnostic function

- If prescribed procedures are not followed, is this random or systematic? What could be the cause of non-adherence to procedures and what could be the remedy? What is the possible effect on study results?
- If a procedure is not as efficient as hoped for, where is the weakness and its origins? What is the projected effect on study results? Are the causes of the inefficiency changeable? At what cost? How urgent is a change?
- If ad hoc decisions, made by personnel in unforeseen situations, are sub-optimal in respect of general epidemiological principles, is there a remediable cause?

Corrective function

- Create conditions that better allow study personnel to comply with prescribed procedures, have low error rates and respect general epidemiological principles. This may include developing new guidelines, retraining, and actions to increase motivation
- Change quality assurance systems or delete and install new ones; If necessary suspend the study

Panel 11.3 General Advice for Approaching Quality Control

- Find an acceptable balance between controlling people and trusting people
- Encourage all study personnel and investigators to be self-critical and open to change and improvement; Do not in any way punish personnel who admit making errors; Reward/praise those who show that they are self-critical
- Intensive communication is the rule; One must involve personnel in diagnostic processes and in decision making processes
- Re-start the study's implementation phase if necessary, redo the analysis if necessary, publish error corrections about a published paper if necessary
- Act before it is too late; but do not always act if the foreseen effect on study results is negligible
- If time permits, retrain people, enhance systems instead of using a tabula rasa approach

Table 11.1 Screening and diagnostic tools of quality control and their general usefulness

	Usefulness of the tool for screening/diagnosis of:		
Quality control tool	Individual errors in routine study procedure	Poor general performance of systems and persons	Poor individual problem solving skills in special circumstances
Self-checks	++	±	–
Supervisor checks	++	++	–
Witnessing	±	++	±
Repeats	++	++	–
Scenario plays	–	+	+
Group discussion	±	++	–
Data cleaning checks	++	++	–
Special data collection for validation	±	+	–

completeness and internal consistency of all data collection forms before sending them to data-entry staff. This screening for data abnormalities must be done within a few days, preferably on the same day of the original data collection.

A percentage of routine questionnaire interviews and bio-measurements conducted by each interviewer should be directly witnessed by supervisors, peers, the study coordinator, quality control persons and/or invited experts. The same witnesses should also independently repeat a percentage of the routine procedures so that results can be compared on a case-by-case basis. These comparisons may also be used to compute observer performance statistics (discussed below). Such repeat measurements should be completed within a few hours of the original measurement. Scenario plays (e.g., mock interviews) can help detect areas of poor performance before they occur during the actual data collection.

In addition to the tactics discussed above, regular group discussions among members of the research team regarding problems of study implementation should be part of the QC plan. Such discussions are particularly useful to identify systematic problems that can be addressed during the study. Another useful strategy to identify systematic problems (as well as individual personnel issues) is to conduct error screening during and after data entry. This may reveal, for example, an unusual frequency of outliers among the values obtained by a particular measurer. These error screens should ideally be done within a few days after initial recording for the early detection of individual and system weaknesses (*See:* Chap. 12: The Data Management Plan, and Chap. 20: Data Cleaning). Lastly, special validation data may be collected in a subset of observation units using a 'gold standard method.' This allows one to make accuracy assessments for individual performance.

As to the corrective function of QC, the tools generally available include retraining, re-motivation, error editing, procedural adjustments, temporarily halting the study, and re-starting the study. The screening and diagnostic efforts of QC often lead to the conclusion that data quality problems are more frequent or severe at the extremes of the range of the measured attribute or among participants with extreme values for important study variables. For example, the quality of ultrasound-based fetal measurements may be lower at the extremes of gestational age. Another example would be the finding that non-response to certain questions is found to be more frequent among very old participants. The consequence is that efforts of correction may have to focus on ensuring more uniform data quality over sub-domains.

11.2 Patterns of Measurement Error

Errors during routine measurements have the potential to create bias in outcome parameter estimates. This implies that a major QC effort must be undertaken to assess the performance of measurers. To appreciate the foundation for QC strategies around measurers' performance, it is helpful to consider possible patterns of error that may be discovered for a single observer using an accurate instrument (Table 11.2).

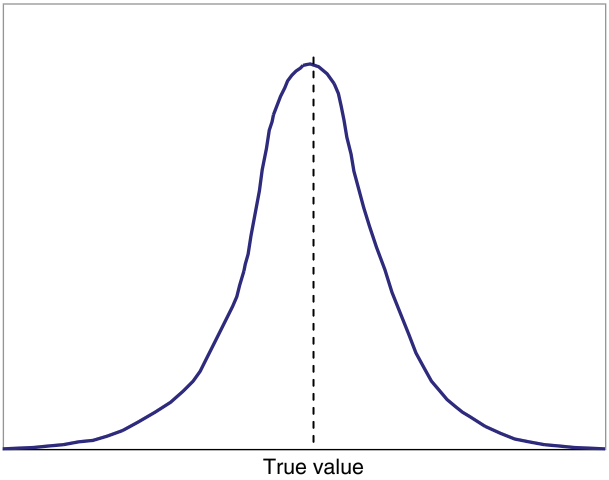
A measurement on a categorical, ordinal, or discrete numerical scale yields a value that is either correct or incorrect (misclassified). In contrast, a measurement value on a continuous measurement scale is always incorrect to some degree, as determined by the limitations of the instrument and the observer. This implies that, for measurements of continuous variables, the frequency of error can only concern errors of a specified size or direction. Errors can be randomly distributed (random error) or their distribution may depend on relevant factors (systematic error). Random error in an observer's measurement values means that the error distribution is unrelated to the levels of the variable itself and to other study variables in the

Table 11.2 Theoretical patterns of measurement error (of a single observer using an accurate instrument)

Measurement level	Patterns of measurement error that can be examined
Continuous	How are sizes and directions of error distributed? Randomly?
	How do sizes and directions of error relate to the magnitude of the attribute?
	How do sizes and directions of error relate to other study variables?
Multi-rank **ordinal** or **discrete numerical**	How frequent is misclassification?
	How are sizes and direction of misclassification distributed? Randomly?
	Does frequency, size or direction of misclassification depend on attribute level?
	Does frequency, size or direction of misclassification relate to other study variables?
Multi-level categorical	How frequent is misclassification?
	Is there a level to which the misclassification is preferentially directed ('direction')?
	Does frequency or 'direction' of misclassification depend on the attribute level?
	Does frequency, 'direction' or level-relatedness relate to other study variables?
Binary categorical	How frequent is misclassification?
	Is it random or is one level misclassified more often than the other?
	Does frequency or level-relatedness of misclassification relate to other study variables?

occurrence relation. A random error distribution for a continuous variable is illustrated in Fig. 11.1 (Panel A). The observer's errors tend to be approximately normally distributed around the true values. Systematic error in an observer's measurement values occurs when the error distribution tends to be skewed in one direction away

Panel A: Example of a distribution of random measurement errors on a *continuous measurement scale*:



Panel B: Example of a distribution of random measurement errors on an *ordinal measurement scale*:

Classified level by observer	True rank		
	1	2	3
1		5% of subjects with true 2	2% of subjects with true 3
2	6% of subjects with true 1		6% of subjects with true 3
3	1% of subjects with true 1	5% of subjects with true 2	

Fig. 11.1 Illustration of the distribution of random errors around the true values for measurements by the same observer using an accurate instrument. *Dashed lines* are lines of symmetry of error distribution. Random errors in continuous variables (*Panel A*) and in ordinal variables (*Panel B*) tend to be symmetrically distributed around the true values. This distribution tends to be Normal for continuous variables and tends to be more frequent in adjacent ranks for ordinal variables. Random errors in binary variables (*Panel C*) tend to be approximately equally frequent for both levels (equal sensitivity and specificity)

Panel C: Example of a distribution of random measurement error on a *binary measurement scale*:

Classified level by observer	True level	
	A	B
A	╲ ╲	9% of subjects with B
B	11% of subjects with A	╲ ╲

Sensitivity for A: 89%
Specificity for A: 91%

Fig. 11.1 (continued)

from the true values, i.e., when there is a correlation between errors and true values, or when there is a dependence of the error pattern upon other study variables.

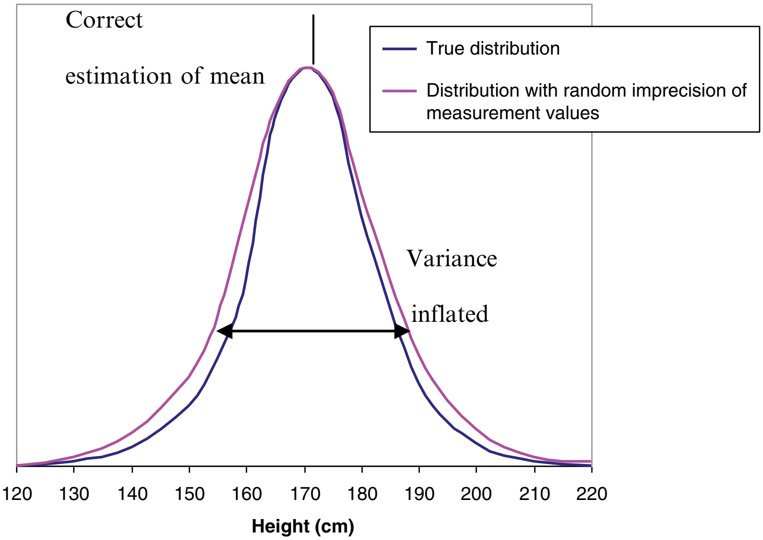
Errors in ordinal and categorical variables can also be randomly distributed (examples illustrated in Fig. 11.1, Panels B and C). Similarly, in ordinal and categorical variables, a deviation from the random distribution pattern or an unexpected relatedness of errors to attribute levels or other study variables raise concerns about systematic error. For example, when the variable is dichotomous, as is common in epidemiology, systematic error may consist of a trend for one level to be misclassified more frequently than the other.

Measurement error is not only an issue of recorded values. There is an additional issue of missing values, i.e., non-recorded values. *Missingness* too can be random or systematic.

When errors are random, they affect the 'precision of the variable'. For example, a larger amount of random error in the measurements for a continuous variable will add more error variance to the true variance of the variable ('lower precision'). It will inflate the observed variance without affecting estimates of the mean, which therefore remain unbiased, as shown in Fig. 11.2, Panel A. Note that, if a continuous variable is measured but then categorized to estimate the proportion falling within an *extreme* category (e.g., the prevalence of stunting, wasting, obesity, anemia, etc.), the effect of random measurement error with inflated variance in the continuous variable will lead to overestimation of that proportion.

Figure 11.2, Panel B, illustrates that a *systematic* error pattern can bias the estimate of a mean *and* variance. When an exposure variable is measured imprecisely, effect estimates (odds ratios, rate ratios) will be attenuated (For illustrations, *See:* Chap. 27). The power of statistical testing will also be diminished. Systematic errors usually cause biased estimates of means, rates, odds ratios, and rate ratios. Thus, both random and systematic errors can bias outcome parameter estimates in epidemiology. Given these effects of measurement error, the QC of observer measurement performance is crucial in every epidemiological investigation.

Panel A: Illustration of the effect of random errors on estimates of mean and variance:



Panel B: Illustration of the possible effect of systematic errors on estimates of mean and variance:

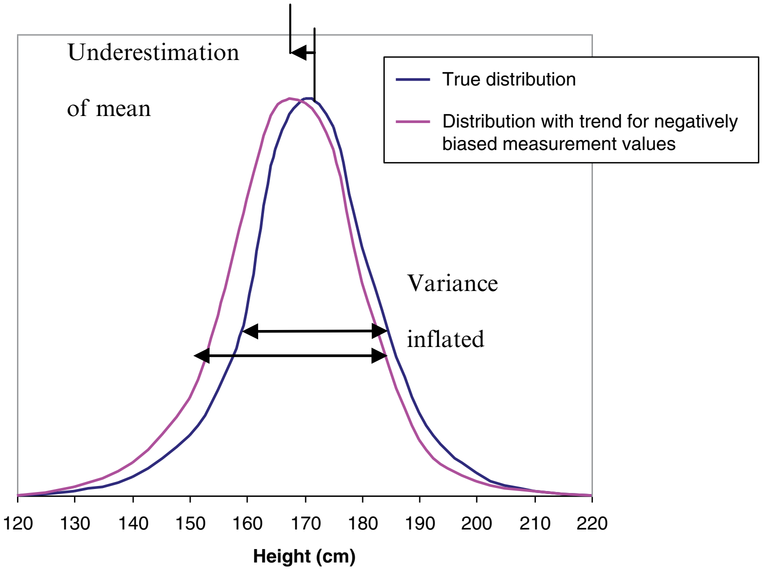


Fig. 11.2 Illustrations of the effects of random and systematic error on the estimation of mean and variance of a continuous variable (height). *Panel A* shows the (vertically aligned) probability density curves for the scenario in which the error pattern was random, resulting in an unbiased estimate of the mean but an inflated variance. *Panel B* Shows a scenario where a systematic error pattern, consisting of a tendency for negatively biased measurement values, resulted in underestimation of the mean and an inflated variance

11.3 Observer Performance

This section focuses on quantifying performance of 'observers.' Observers are understood to be measurers of the personal attributes or living conditions of others who are often called 'subjects.' They can also be measurers of characteristics of other observation units. The quality of the measurements of observers can be influenced by:

- Problems with the instrument
- Inadequate measurement environment; interference with measurement
- Poor subject preparation
- Un-cooperative subject
- Poor general measurement skills
- Particular observer problems during particular measurements

Out of all the sources of error listed, all except the first can result in what is commonly called 'observer error.' The term 'observer error' is thus a misnomer in the sense that it results from a combination of non-instrument sources of error, including errors that are due to the measurement environment and the measured subjects' behavior. It is, however, implicit that it should be part of an observer's skills to adequately prepare the measurement environment and solicit adequate cooperation from subjects.

The performance of an observer can be documented by several types of performance statistics. Discussed below are statistics of error frequency, observer accuracy statistics, observer precision statistics, and terminal digit preference statistics.

11.3.1 Single-Observer Error Frequency Statistics

During a study, error frequency statistics can be monitored over time for each observer. Such statistics describe the frequency of missing, erroneous, or otherwise unacceptable values (according to some hard criterion). One can consider calculating these frequencies for variables separately as well as jointly for groups of related variables. One can also examine whether the error frequencies are related to the level of the variable or whether they are related to other study variables. Frequency statistics are usually calculated separately for outliers and missing values.

As to outliers, sometimes it is clear that a value is erroneous or has an unacceptable degree of error even if the true value is unknown. This is the case when the value does not pass a hard validation criterion (e.g., a height value of 300 cm). We can document the frequency with which recorded values exceed some expected 'normal range' of possible values. The limitation of error frequency statistics is thus that they only allow a rough estimate of the frequency of *severe* errors. Less severe errors are more frequent but they often lead to values well within the 'normal range' ('*erroneous inliers*') of which the frequency cannot easily be estimated.

11.3.2 Single-Observer Accuracy Statistics

The aim of this type of statistic is to document the observer's tendency to contribute to systematic error in important study variables. There are several aspects to observer accuracy because deviance from a 'symmetric' random error pattern and relatedness of the errors to true values and to other variables can take several forms. Observer accuracy statistics can thus differ in what they capture exactly. Common approaches include:

- Calculation of the average of deviations from the true dimension (on a continuous or discrete numerical scale) called *average bias*. According to this criterion an observer is considered to be accurate if, on average, (s)he measures the real dimension when using an accurate instrument in a series of independent replicate measurements. Oppositely, an inaccurate observer tends to record values that are higher (*positively biased*) or lower (*negatively biased*) on average. The observer is thus considered accurate if her/his errors are random.
- Monitoring whether *sensitivity* tends to be different from *specificity* for dichotomous variables. Random error tends to lead to equal sensitivity and specificity. The meanings and calculations of specificity and sensitivity are discussed in Chap. 6 (General Study Designs).
- *Bland-Altman plots* can be used to examine whether errors in continuous variables are related to the magnitudes of true values (Bland and Altman 1986). These are scatter-plots of the difference between paired measurement values (e.g., weight measured by two independent observers) against the mean of the paired measurement values. Each set of paired values must represent the same attribute and must be obtained (1) by two different observers or (2) by two different techniques (by one observer).
- *Sign tests* can be used to capture asymmetry in error distribution in continuous variables (WHO 1983). They are based on the expectation that, if there is no reason why the first of a pair of measurement values would be greater than the second, then the probability of having a particular number of first values being greater follows a binomial distribution.

The listed approaches all require that one has some approximate knowledge of what the true value is behind each recorded value. They are based on a comparison between observed values and so-called *gold standard values*. The understanding is that gold standard values for continuous or discrete numerical variables are not necessarily exact true values, but values considered close enough to the truth to serve as a yardstick for observer accuracy. Gold standard values represent those obtained by a highly accurate *expert observer*, or using a different, *more accurate method*. Gold standard data can be obtained at different study stages (Panel 11.4).

One needs to decide what size of average bias (or other statistic) should be a reason for concern. Availability of standards for this decision is highly dependent on the research domain, measures, procedures, and systems used. If no such standard exists, then deciding the threshold for concern is a matter of judgment.

Panel 11.4 Strategies for Obtaining Gold Standard Data for Assessing Observer Accuracy

Strategies based on independent replicate measurements by experts

- Expert replicates during training and piloting phases
- Expert replicates a proportion of routine study data
  - Random spot-checks
  - Planned quality control re-measurements
- Expert replicates in specially organized sessions outside routine data collection

Strategies based on control measurement with a more accurate gold standard technique

- Validation data collected during a pilot validation study or training session
- Validation measurements for a proportion of routine data
- Special sessions for quality control, with use of the gold standard method

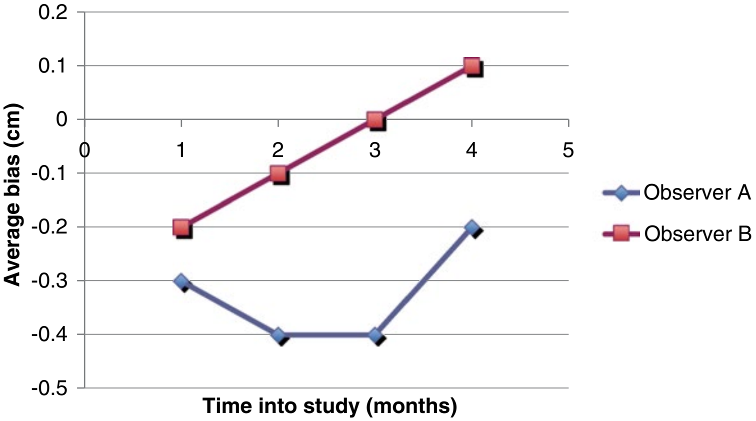


Fig. 11.3 Example of an average bias monitoring plot for height measurements

For average bias, it may be reasonable in most cases to be concerned only when it is greater than the technical error of measurement (TEM) of an expert, a statistic that is discussed further below.

Plots of error rates or average bias over time are used for monitoring individual observer accuracies. Figure 11.3 is an example of a plot of the average bias of two observers during a data collection period. This is not only useful for follow-up studies, but also in cross-sectional studies if the data collection period is longer than a month. Observer accuracy statistics can then be calculated every fortnight or every month, for example. If the observer's accuracy seems to drift, this information may be used as feedback to the observer and may alert the investigative team to individuals in need of re-training.

11.3.3 Single-Observer Precision Statistics

The aim of this type of statistic is to document the observer's tendency to contribute to inflating the variance of important study variables. Such increases are mostly due to random error, but systematic errors can also increase variance. The common approach to calculating single-observer precision statistics consists of quantifying the consistency of the observer's independent replicate measurements on the same subject or item. An observer operating on a multiple-ranks ordinal, discrete numerical or continuous scale is said to be precise if her/his measures values tend to be close to each other and not widely dispersed across the scale. These repeat measurements must be of the same item during an interval in which there was no change in the measured attribute, and the observer must have been using an accurate instrument. For a categorical scale, the total agreement among the observer's replicate assessments (related to total error variance) expresses the observer's precision.

The measures of dispersion or disagreement can be obtained after doing a lot of re-measurements on a single observation unit, or, by doing just one or two repeats on a series of observation units. The first option (many replicates, same unit), tends to place higher burdens on measured individuals and it is also difficult to make the replicate measurements truly independent. When replicate values are obtained from multiple observation units, these may be obtained at different study stages (*See:* Panel 11.5).

The general preference is to obtain replicate data during routine data collection because this usually allows for replicates on a large number of individuals and is directly relevant to the study results. On the other hand, collecting replicates during training or during special quality control sessions allows for easier and more immediate feedback on reasons for the lack of precision. Organization of such sessions ('test-retest exercises') is discussed separately in the next sub-section.

Single-observer precision statistics are often calculated on the basis of the replicate values include:

- For categorical variables:
  - Kappa coefficient
- For continuous variables:
  - Technical error of measurement
  - Coefficient of variation
  - Reliability coefficient (Intra-class correlation coefficient)

Panel 11.5 Strategies for Obtaining Replicate Data on Single Observer Precision

- Observer replicates during training and piloting
- Observer replicates during routine data collection
  - Systematic duplicates or triplicates on the variables e.g., blood pressure
  - Duplicates or triplicates in a proportion of observation units only
- Observer replicates during special quality control sessions outside routine data collection

The *Kappa coefficient* is a measure of agreement in assigning the levels of a categorical variable to a series of observation units, i.e., agreement among k observers or agreement over k different occasions. In the case of monitoring individual performance, it is the agreement among k different replicates which is applicable. The Kappa coefficient is a number between 0 (no agreement) and 1 (perfect agreement) calculated as the ratio of the proportion of times that there is agreement to the proportion of times there could theoretically be agreement (where both proportions are corrected for chance agreement):

$$\textbf{Kappa coefficient} = \frac{P_A - P_E}{1 - P_E}$$

Where:
P_A = proportion of times there is agreement between independent replicates
P_E = expected proportion of chance agreement

For continuous variables in which two or more repeat measurements are taken, the *technical error of measurement* (TEM) can be calculated. The simplest form of the TEM equation involves duplicate measurements and is shown below. If there are more than two replicates, then the calculation of TEM is more complicated (*See:* Uliaszek 1999 for the appropriate TEM equation).

$$\textbf{Technical Error of Measurement (TEM)} = \sqrt{\frac{D^2}{2N}}$$

Where:
D = differences between two independent replicates
N = number of subjects measured

TEM of an observer should approach that of an expert measurer or a TEM typically reached in high quality studies. For example, the ranges of acceptable TEM values for most anthropometric measures are fairly well known (e.g., Chumlea et al. 1990). For variables with an unknown expert TEM, a reference TEM may be obtained by involving an expert in training sessions in the study preparation phase or in the early data collection period. This can be part of the QA/QC plan. Formal comparison of observer TEM with expert TEM is done using an F-test of TEM-squared (Mueller and Martorell 1988).

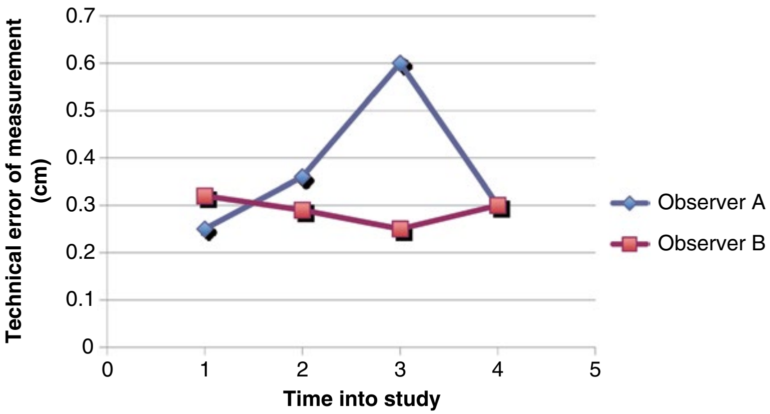


Fig. 11.4 Example of a monitoring plot of technical error of measurement in height measurements

Plots of TEM over time are used to monitor individual observer precision during a study (Fig. 11.4). An intuitive way of using these plots is to determine if the observer's TEM starts exceeding the expert TEM by more than a chosen threshold percentage. Repeatedly deteriorating TEM values, even when they do not reach the threshold of concern, warrant feedback because such a trend could indicate suboptimal performance. The TEM-based monitoring strategy outlined above has an equivalent for categorical variables under the form of repeated assessments of kappa statistics and appropriate comparisons with kappa statistics from experts.

The *coefficient of variation* (CV) of an observer expresses the amount of dispersion in her/his replicate measurement values as a fraction of the mean value. The adjustment for mean value renders this statistic more comparable among continuous variables for different attributes and for different mean sizes.

$$\textbf{Observer coefficient of variation} = \frac{TEM}{\mu}$$

Where:
TEM = technical error of measurement
µ = mean

The *observer reliability coefficient* (ORC or RC), also known as the *intra-class correlation coefficient* (ICC), is the ratio of true subject variance over the observed variance (which is the sum of true subject variance and measurement error variance). An RC of zero means that all observed variation is due to error. Assuming the observed variance reflects that of the target population and that errors are nearly uncorrelated with true values, one can use a formula to calculate RC using observed variance of the variable and TEM obtained from the repeat measurement values:

$$\textbf{Observer reliability coefficient (RC)} = \frac{\sigma_o^2 - TEM^2}{\sigma_o^2}$$

σ^2 = observed variance
TEM = technical error of measurement

TEM tends to be more useful than CV or RC for interpreting individual performances. This is, first of all, because TEM has the same units as the relevant measurements and is therefore more intuitively interpretable. For example, an observer has an intra-observer TEM of 0.25 cm for height measurement. The interpretation is that duplicates of this observer will be within about ± 2 * TEM (i.e., within 1 cm) 95 % of the time. This can be readily understood when explained to trainees. On the other hand, CV and RC/ICC are more valid and useful than TEM for comparing observer precision among several measures, for example for knowing if systolic blood pressure is measured as precisely as diastolic blood pressure. They are also useful for comparing an observer's precision across the range of magnitudes of the attribute. Such comparisons, however, are not always of prime interest for individual performance monitoring in the context of quality control during study implementation.

11.3.4 Test-Retest Exercises

Test-retest exercises are special measurement sessions during which independent replicate data are obtained from a minimum of 10 but preferably 15 or 20 subjects. The replicates are obtained by the observer(s) to document observer precision. The sessions are usually supplemented with additional replicates by the trainer-expert to enable the evaluation of observer accuracy. Often, 2 or 3 replicates are done by the observer and 1 or 2 by the trainer. The subjects are re-measured after a short interval during which no measurable change in the measured attribute is expected so that any change in measurement values is due to instrument or observer error. When an accurate instrument is used, the changes seen are attributable to observer error only. Test-retest exercises can be useful as a part of training and quality control.

When setting up a test-retest exercise, a few issues should be kept in mind. Most importantly, the goal should be to identify the effect of the observers' performance. Therefore, one must try to minimize all other sources of measurement variation. For example, an anthropometric test-retest exercise in children may achieve this goal by aiming for the following setup:

- All instruments are highly accurate and precise
- Instruments are calibrated at the beginning and in the middle of the exercise

- All observers measure the same subjects, or, each observer measures a set of subjects with a similar distribution of key variables, e.g., age
- Every subject is measured in the same room and in the same part of the day by all
- Every subject is measured with the same instrument by all
- Every observer works with the same assistant(s)
- Assistants (e.g., holding the head during a length measurement) perform in a highly standardized way
- Errors in data processing are avoided by double data entry

A second goal is to ensure that the replicate measurements are truly independent. Achieving this goal is difficult if the observer knows the exact value of the previous measurement. Continuing with the example of an anthropometric test-retest study in children, the following strategies can help to increase the likelihood of obtaining independent replicate measurements:

- There should be enough time between replicates (at least 20 min)
- Between replicates the observer should be busy measuring other subjects and measuring other dimensions on the same subject
- The observers should have no access to the written result of previous measurements (patient files or record forms) taken by her/himself or by others
- The observers should be encouraged to not attempt to remember the results of previous measurements

A third goal is to simulate the real study conditions in the test-retest scenario. In other words, the special test-retest study should not be too different from the real study in terms of characteristics of the subjects (e.g., the same age category), the instruments that will be used in the real study, and the type of assistance that will be available in the real study, etc. By meeting this goal, the performance statistics from the test-retest study are more likely to reflect the reliability that may be achieved in the real study.

Finally, a fourth goal is to limit the total number of measurements that each subject will undergo to minimize the burden of participation and the total workload of the participating observers. For example, for an anthropometric test-retest study this usually means that:

- The total daily participation time of each subject should preferably not exceed 15 min for newborns, 30 min for infants, 45 min for young children, and 60 min for older children
- There should be enough time between measurements
- The total duration for observers should never exceed 4–6 h of effective measuring per day

11.3.5 Single Observer Terminal Digit Preference

Inaccuracies of individual measurement values may have a frequency of occurrence, an average magnitude, or a spread. But they can also lead to strange patterns identifiable in the recorded data values. For example, particular end-digits of recorded values of continuous variables may be recorded more frequently than expected, particularly the end-digits 0 and 5. The frequency of certain end-digits can be used for monitoring individual measurement performance for continuous variables. For example, one can plot for each observer the percentage of digits ending with 0 or 5 among all measurement values of a variable taken by that observer over a fortnight and repeat this for successive fortnights. The usefulness of this monitoring method is restricted to continuous variables and is greater for continuous variables measured using instruments with an analogue display than for instruments with a digital display. Chi-square tests (with one degree of freedom) can be used to test whether a particular observer's data show terminal digit preference. In addition, the chi-square test statistic can be used as a measure of *degree of* terminal digit preference (Altman 1991). Terminal digit preference is discussed in more detail in Chap. 29 (Reporting Data Quality).

11.4 Team Performance

Multi-center studies have site teams whose performance is of separate interest. Teams within a single site can also be of interest, as some teams tend to operate quite differently than others. In analogy with single observer performance, the main parameters of team measurement performance concern completeness of data collection, outliers, team accuracy, team precision, and terminal digit preference.

11.4.1 Team Accuracy

Team accuracy in measuring numerical variables is conveniently quantified as *team average bias*, a statistic based on pairs of observed and gold standard values. This statistic is analogous to that of single observer accuracy, but in this context the paired values of all observers of the team in question are pooled. The strategies for collecting the paired data at different stages of a study are also completely similar to the strategies used for single observer accuracy; therefore, monitoring a team's accuracy over time occurs in parallel with monitoring single observer accuracy. The accuracy of one team can be compared with the accuracy of another team or with all remaining teams together using a Student's t-test. For categorical variables, a common approach is to monitor *sensitivity and specificity* of data collected by the entire team.

One can check if team average bias or sensitivity/specificity tend to differ across levels of other important study variables. Based on this, one can assess how they could potentially bias outcome parameter estimates.

11.4.2 Team Precision

The question at issue is how much of the total measurement variation in data produced by a team of observers is attributable to the observer errors of all observers combined. For numerical variables this can be evaluated by calculating a team's

TEM, which is calculated using the same formula as for single observer TEM, but with the replicate values of *all* observers in the calculation. Team TEM naturally has an intra-observer and an inter-observer component. For team TEM to be maximally relevant, each observer in the team should contribute a number of replicates proportionally to her/his contribution to the routine data collected in the relevant study period. Otherwise, the strategies for collecting the replicate data at different stages of a study are similar to the strategies used for single observer accuracy.

For categorical variables a *team Kappa coefficient* can be calculated. The monitoring of the team TEM and Kappa coefficients over time runs in parallel with the monitoring of single observer precision. F-tests can be used to check if a single observer's precision for a numerical variable can be considered as different from the joint precision of the other members of the team. Similarly, the precision of one team can be compared with another team or with all other teams together, using F-tests.

11.5 Instrument Performance

Recall that total measurement error is composed of instrument error and observer error. When the quantification of observer error was discussed in the previous sections, it was assumed that the instrument – each time it was used – was highly accurate and precise. This may be a reasonable assumption over the short term if a quality technical device was used. Yet, it is wise to check the calibration status of instruments frequently. The cost of not monitoring instrument performance can be very high, especially after the sudden detection of a problematic instrument without any idea when the problem started and how many previous measurement values were affected or to what degree. Thus, each measurement value in a study should be linked not only to information on the observer-measurer but also to information on the specific device or devices that were used for its measurement.

The main aspect of performance worthy of monitoring in a technical device is *instrument accuracy*. An instrument for measuring a discrete numerical or numerical variable is accurate ('*well calibrated*') if, on average, it measures the true value when applied correctly. An instrument is called inaccurate ('*de-calibrated*') if it has an inherent tendency to yield values that are too low or too high in comparison with the true values. The quantification of instrument accuracy requires an observer to make measurements of high technical quality on subjects or items for whom/which the true value is exactly known. This is called a *calibration (status)* check, and it yields differences between true values and (average) high-quality measurement values. The items measured can be special calibration materials, e.g., calibration blocks of known weight to check a weight scale. A *calibration log* should be kept during the study with the results of the calibration checks. De-calibrations of a magnitude raising concern should lead to appropriate corrective action, the success of which must be verified by new checks.

A technical instrument is considered to be imprecise if it has properties that contribute to inflating error variance, even if it is accurate on average. *Instrument precision* mainly depends on the graduations of the measurement scale. Widely spaced graduations lead to readings that are widely distributed around the true values and thus contribute to error variance. There may be other intrinsic characteristics ('stability') of an instrument leading to some random error variation. This is not always monitored during quality control but should always be a concern when purchasing instruments.

Validated measurement procedures with a robust QA/QC plan (Chaps. 10 and 11) allow one to obtain high quality data during the data collection phase of a study. The next chapter focuses on what should be done with these data to make them available for statistical analyses.

References

Altman DG (1991) Practical statistics in medical research. Chapman and Hall, London, pp 1–611. ISBN 0412276305

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet i:307–310

Chumlea WC et al (1990) Reliability for anthropometric measurements in the Hispanic Health and Nutrition Examination Survey (HHANES 1982–1984). Am J Clin Nutr 51:902S–907S

Mueller WH, Martorell R (1988) Reliability and accuracy of measurement. In: Lohman TG, Roche AF, Martorell R (eds) Anthropometric standardization reference manual. Human Kinetics Books, Champaign, pp 83–86

Ulijaszek SJ, Kerr DA (1999) Anthropometric measurement error and the assessment of nutritional status. Br J Nutr 82:165–177

World Health Organization (1983) Measuring change in nutritional status. WHO, Geneva, pp 1–101. ISBN 9241541660

The Data Management Plan

12

Meera Chhagan, Shuaib Kauchali,
and Jan Van den Broeck

Data need tender loving care.

Abstract

Data management in research is a process geared towards making recorded information available for use in analyses. This process involves a computerized data system that structures and stores electronic data. The main purposes of a computerized data system are to archive, retrieve, and extract data, and these processes must maintain the integrity of original data. In support of these purposes, data systems should be set up to make input, retrieval, and extraction as efficient (fast, easy) as possible. Moreover, privacy and confidentiality concerns should be of primary consideration when creating and limiting the range of possibilities for data retrieval and extraction. Thus, the principles of validity, efficiency, and ethics apply to the way data systems are set up and managed.

12.1 Data Handling Capacity

The management of research data requires software, hardware, infrastructure, human resources, and a data management protocol. The extent of resources needed for data management can range considerably according to the scale and particularities of a study. On one end of the spectrum, a small study may require only a single person delegated to handle all the data management tasks. For example, a hospital

M. Chhagan, Ph.D., FCPaed (✉) • S. Kauchali, M.Phil., FCPaed
Department of Paediatrics, University of KwaZulu-Natal, Durban, South Africa
e-mail: Chhagan@ukzn.ac.za; kauchalis@ukzn.ac.za

J. Van den Broeck, M.D., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_12,

nurse may wish to conduct a small cross-sectional study on a series of her patients. Any personal computer with a spreadsheet program, together with some printed questionnaires to be stored in a drawer may be all that is needed for data management (in addition to some basic skills of spreadsheet use and a brief data management plan). On the other end of the spectrum, very large studies may require many well trained data management staff, a suite of computers, multiple data servers, expensive software, outside contractors, and a detailed data management plan. Indeed, data handling is as crucial for the validity of a study as is study design. In this chapter, therefore, we provide a rough outline of the data handling capacity typically needed in a medium-to-large size study and provide practical hints for developing and implementing a data management plan. It should be noted that this chapter is mainly targeted to researchers and students who have limited experience with data management. Panel 12.1 introduces terminology used in this chapter.

Panel 12.1 Selected Terms and Concepts Relevant to Designing a Data Management Plan

Access control Restriction and monitoring of access to data

Audit trail (of data processing) Documentation that allows reconstruction of the course of data processing events in a study

Analysis dataset Selection of fields and records extracted from a database, used for a particular statistical analysis

Backup (of a database or analysis dataset) Time-stamped duplicate, separately and safely stored as a security measure against loss of the original

Barcode Sequence of variably spaced bars of variable width, used mostly as a code that uniquely identifies a source document or biological sample and its origin

Code list List of possible values of a categorical variable

Coding Designing a code list

Data capture Process of transforming raw data into electronic data

Data dictionary List of metadata

Data entry Transfer of information into an electronic format suitable for inclusion into a database

Data handling *See:* Data processing

Data handling protocol A plan as to how data should be recorded, stored, cleaned, and prepared for analysis. *Syn.* Data management plan

Data management The organization of data processing

Data manager Member of study personnel to whom the investigator has delegated responsibility for data management

Data processing Recording, storing and extracting data, and cleaning and preparing data for analysis (*q.v.,* data handling)

Data system Infrastructure, logistics and official procedures of data handling

(continued)

Panel 12.1 (continued)

Database Organized set of data collected in the study, kept as a source for extracting analysis datasets

Dataset *See:* Analysis dataset

Good Clinical Practice guidelines A standard for all stages of conduct of clinical trials aimed at: (1) optimizing validity and credibility of data and results, and (2) ensuring that the rights, integrity, and confidentiality of data of trial subjects are protected

Metadata Data about data

Pre-coding Designing a list of possible values of a categorical variable to be used for first recording of measurement values

Standard Operating Procedures (SOP) Detailed written instructions to achieve uniformity of the performance of a specific function

12.1.1 Human Resources for Data Management

The first and most important element of data handling capacity is human resources. Recruitment and training of appropriate personnel can take up considerable study time, and it is notably difficult to attract experienced information technology (IT) and data management staff to data centers located in resource-poor areas (Van den Broeck et al. 2007). In medium-size and large studies the data management team may include:

- Investigators
- Data managers
- Software programmer-developers and IT personnel
- Data collection personnel
- Data entry clerks
- Data quality control personnel
- Archivists
- Transport, communication, administrative, and maintenance personnel

Listed first, the investigator must always be involved in data management even if most of the responsibilities are delegated to a data manager and to software programmers/developers.

In the study preparation phase there will be the needs to train relevant staff in data management procedures and to gradually refine standard operating procedures (SOP) for data handling. Pilot runs are helpful in these tasks and should help staff to standardize their data collection, transfer, entry, quality control, and maintenance procedures. Pilots are also helpful to identify problems that need reconciliation or clarification, to re-enforce responsibilities of each team member, and to establish clear lines of communication.

12.1.2 Data Management Infrastructure

The second element of data handling capacity is appropriate data management infrastructure. Typically the main elements of data management infrastructure consist of transport facilities, a data centre, and an archive. In large multi-centre studies with a decentralized data management model (e.g., Onyango et al. 2004), these three elements are often present in each centre. Below we briefly review the requirements for transport and for setting up a data centre.

12.1.2.1 Transport Facilities for Data Management

Transport capacity is typically a crucial aspect of data management in population-based studies. More transport capacity is needed if the study uses paper source documents as opposed to direct electronic data recording. If there are paper source documents, the required transport capacity depends on such factors as:

- Volume of data
- Data collection schedule
- Batching method and schedule for archiving
- Querying frequency
- Number of data collection sites
- Distances between data collection sites and the data centre
- Maximum allowable delays between data collection and arrival at a data centre
- Distance between data centers and the archive
- Retrieval frequency

The needs may vary over the course of a study. In prospective follow-up studies, for example, there may be a gradual increase in needs in the beginning of the data collection phase and a gradual decrease near the end.

12.1.2.2 Data Centre

The main function of a data centre is to serve as a central location where the peripherally collected data arrive, are initially verified, and entered or consolidated into a database, and cleaned. It is also a centre of communication and feedback about data-related issues. When planning a data centre, one should consider the requirements for space and equipment, again keeping in mind the projected changes over the course of the study.

When planning for physical space, one may consider that the needs are determined by space needs for:

- Staff
- Computers, printers, and telecommunication
- Data storage (e.g., filing cabinets, shelves)
- Storage of consumables (e.g., printer cartridges)

Data centers require both hardware (e.g., computers, external data storage devices) and software. They may also need special equipment arising from the method of data collection. For example, reliable communication networks and battery recharging options are needed for mobile devices.

12.1.3 Software for Data Management

When setting up a data system, consider that multiple types of software may be needed. This is certainly the case if one chooses to set up a data system ‘from scratch’ with the help of programmers-developers. The software needed may include, for example:

- Questionnaire development software
- Barcode printing software
- Data entry and query generating software
- Optical scanning software
- Electronic filing software
- Database software
- Post-entry data cleaning software
- Data extraction software
- Networking software
- Web interface software

Software systems combining database with data entry facilities, data analysis facilities, and data reporting facilities do exist and there is a natural attraction to them. However, when in considering whether or not to use aggregate software systems, beware of a possible drawback: their apparent gain in overall functionality is sometimes offset by a lack of functional flexibility in their components. For example, some combined data entry and database systems have data entry facilities that do not include the possibility of multiple validated data entry that one would expect from dedicated data entry software. Large volumes of data and decentralized multi-centre studies necessitate a system for simultaneous data entry by multiple data entry clerks (*See:* Textbox 12.1). Special software needs arise with use of mobile technology or with use of GIS.

Textbox 12.1 Efficiency of Large Data Systems

It is the perception of many investigators that, in spite of claims to the contrary by data managers and data system developers, the efficiency of a data system, defined as the resources required per volume of quality data available for analysis within a defined time span, does not usually improve with increasing size of a study. The impression is rather that, in line with the well-known ideas of the philosopher Ivan Illich, efficiency rapidly levels off and then decreases with increasing study size and with the increasing refinement and complexity of the data system that goes along. Which view is closer to the truth requires further operational research, as do possible methods to increase cost-effectiveness of data management in health research. Common sense would suggest, however, that the use of mainstream user-friendly database systems and other ready available and simple to use software packages will often be more efficient than the programming of data systems ‘from scratch’ or than using systems that require high technical expertise for development and maintenance.

12.2 The Data Management Protocol

The data management protocol (*Syn.* data handling protocol) is a plan for how data should be recorded, stored, cleaned, and prepared for analysis.

12.2.1 Elements of the Data Management Protocol

The data management protocol commonly includes the following elements:

- Procedures for retrieval of source documents from an inventory for printing
- Data collection procedures
- Source data accumulation and transportation plan
- Database construction and accessibility plan
- Data entry plan
- Data cleaning plan
- Change control and backup plan
- Archiving and retrieval plan

Data collection and data cleaning are aspects that are covered in other chapters. The other aspects of the data management plan are discussed in this chapter.

It is recommended that standard operating procedures, i.e., detailed written instructions to achieve uniformity in the performance of a specific function, be developed for each plan. Guidelines for GCP in data management are available from various sources and can be used to guide the development of a study-specific data management protocol. Special requirements exist for data handling in clinical trials, some of which are discussed in the next sub-sections.

12.2.2 Guidelines for Computerized Systems for Clinical Trials

Guidelines for good practice in the use of computerized systems have been developed for clinical trials (FDA 2007) but they have some relevance to all health research and are therefore mentioned here. Below is a list of selected guidelines [with minor edits] with general relevance to health research:

- “The Study Protocol needs a specification on how the computerized system will be used to create, modify, maintain, archive, retrieve or transmit data”
- “For each study, documentation should identify what software and, if known, what hardware is to be used in computerized systems that create, modify, maintain, archive, retrieve, or transmit data. This documentation should be retained as part of study records”
- “Clinical investigators should retain either the original or a certified copy of all source documents sent to a sponsor or contract research organization, including query resolution correspondence”
- “Computerized systems should be designed to preclude errors in data creation, modification, maintenance, archiving, retrieval, or transmission”
- “Security measures should be in place to prevent unauthorized access to the data and to the computerized system”

- "Data should be retrievable in such a fashion that all information regarding each individual subject is attributable to that subject"
- "The computerized system should comply with study protocol specifications about metric units and blinding"
- "Changes to electronic records require an 'audit trail' and should never obscure the original information"
- "All software, operating systems, development tools should be available in a format that allows review"
- "Preclude unintended interaction with non-study software: isolate study software logically and physically as much as possible"
- "Controls must be in place to prevent, detect and mitigate effects of computer viruses"
- "Contingency plan: there must be an SOP in place describing how to continue the study in the event the computerized system fails"
- "Back-up and recovery SOP for regular backups, storage in a different building. Keep detailed backup and recovery logs"

12.2.3 The Change Control Plan

Change control refers to the process of monitoring and documenting changes in versions of software, hardware and other equipment used in a research study. One should be able to trace the flow of information in a study, and this is a strict requirement for clinical trials. The practical implication is that investigators must document and justify any changes made to data or the data system. In addition to the audit trail discussed separately below, change control also includes software upgrades and version control of data collection forms or questionnaires. Any changes to the computerized system such as software upgrades or replacement of equipment must be validated and documented. While newer versions of software usually have enhanced built-in functions and solutions to technical bugs, one may wish to retain the possibility (if necessary by contractual arrangement with the supplier) to continue running old versions for the purpose of performing study audits.

12.2.4 The Plans for Backups, Archiving, and Retrieval

The structure and location of data archives are essential for good data management. For example, paper records should have a well-organized system of filing and linking that allows easy retrieval. Electronic data archiving systems should differentiate between the current updated database and its previous iterations. There may be a large number of iterations, and as the database progresses from untouched to final, there should be regularly scheduled backups to ensure that work is not lost if an unforeseen computer or technical problem arises. Storage of backups in different physical locations may be part of the plan. The frequency of backups is predetermined and depends on the volume of data entry and edits per time unit. An encrypted dataset that has all identifying data should be separately maintained and properly secured to avoid leaking confidential information. The principal investigator will specify occasions that warrant re-linking identifiers and plan ahead for how the process will be managed.

The retrieval SOP should specify the process whereby investigators gain access to source documents. Retrieval may be needed for data verification or addressing data or protocol queries. Named individuals, usually a data manager, are responsible for data retrieval. The SOP will ideally specify how the request for retrieval is handled and what feedback is needed from investigators once they have viewed retrieved information. Finally, the SOP should also specify the processes around safe transmission of data, for example via copying rules and rules for electronic transmissions.

12.3 Database Systems

A database is an organized set of data collected in the study, kept as a source for extracting analysis datasets. It consists of one or more database tables which are matrices of database fields and records. The database fields are the ‘columns’ in a database table which contain or are intended to contain data on a particular ‘variable’. Database records are the rows in the database table which contain data on particular observation units.

12.3.1 Database Software and Hardware

The requirements for database software and hardware are determined by the study size, design, and type of quantitative and qualitative information. While software and hardware are generally specified in institutional policies and practices, there may be specific study situations that warrant special attention. For example, special measures to secure electronic transfer of scanned forms or the use of secure servers may be necessary in studies in remote field sites in resource poor locations, where transport of data forms may be irregular or power outages frequent.

Software systems may combine database facilities with data entry, data analysis and data reporting facilities, and within these systems there are several options that researchers may find useful. Popular examples are Epidata (2011) and Epi Info™ (2011) software. Some of these allow double entry verification. Unless there is easy and timely access to database programmers, most researchers prefer to use such proprietary software for database-related functions. However, these database systems can be expensive, prompting some developers to publish open source (free) database software. When open-source software is chosen, it may be ideal to have database programmers included in the study team should consultation or adaptations be necessary. Special software for qualitative data or for multiple media files should be planned by or with experts in this area. Finally, with all systems and components of a database, ease of exporting and compatibility with existing statistical software is an important concern.

12.3.2 Structure of the Database

The structure of the database is selected based on considerations of several factors, the most important of which are the study design and planned statistical analyses. Many studies are able to employ the simplest database structure, known as a *horizontal data table* (or *short format database*). In a horizontal structure, there is a single record (row) for each participant irrespective of the number of waves of data collection. Columns represent study variables for each record. Horizontal structures are usually suitable for simple descriptive studies involving only one wave of data collection (e.g., surveys).

Sometimes, however, a more complex structure – the *vertical data table (or long format database*) – is needed. In a vertical structure, there is a new record (row) for each of *n* waves of data collection for each participant. Again, columns represent study variables. Vertical structures are routinely helpful for studies involving more than one wave of data collection (e.g., longitudinal studies), though it is possible to employ a horizontal structure if necessary.

The type of database structure selected has implications for how variables are named and for how waves of data collection are identified and extracted for analyses. In a horizontal database, the variable name from one wave must be different from that of the next wave. In this case, the *waves are identified* in the names of each variable (e.g., height-1, height-2, height-3, where each number represents a separate wave). In a vertical database, the same variable names can be retained for each wave because separate records identify separate waves. Analysts may change the data table structure at the analysis stage depending on what is easier for the specific analysis.

Both horizontal and vertical structures can be compounded to create more complex database structures. For example, imagine a study in which the design calls for relating child-level data to adult- and school-level data. In this circumstance, the database structure can be designed to capture, manipulate, and analyze hierarchical or multi-level data. Compound databases (also called relational databases) are usually required for such studies and necessitate consultation with database designers with more experience in such designs.

Hint

Reporting requirements from oversight bodies may strongly influence the chosen database structure. It can be helpful to consult with oversight bodies during the study planning stages to ensure that the correct database structure has been chosen.

12.3.3 Database Variables and Coding

Data managers should keep a code book with a list of variables and their value codes.

Table 12.1 Coding of multiple response type items: an example

Question/item a	Question/item	Variable in database	Value for variable
D.1.	What are the current treatment needs of the child? (*Tick* all *the options that apply*)		
	□ None	D.1.1	0 or 1
	□ Community rehabilitation	D.1.2	0 or 1
	□ Medication	D.1.3	0 or 1

Table 12.2 Coding of single response type items: an example

Question/item a	Data collection tool	Variable in database	Values for variable
D.1.	D.1. What are the current treatment needs of the child? *(Tick* only one *option)*	D.1.	
	○ None		0
	○ Community rehabilitation		1
	○ Medication		2
	○ Community rehabilitation and medication		3

The paragraphs below illustrate that the development of a questionnaire and the development of a database and code book go hand in hand. Code books need to be updated each time form versions change and/or new variables are created.

12.3.3.1 Coding of Single- and Multiple-Response Items

If multiple-response items are planned (e.g., categorical data collection containing fully or partially non-mutually exclusive response options), care must be taken to have separate variables assigned for each possible option (Table 12.1). The response will then be coded as 0 (not selected) or 1 (selected) for each variable. The example in Table 12.1 shows a single question (item) with three answer options. Each answer option will become its own variable in the database and contain a 0 or 1 to represent whether or not the answer option was selected. If the subject selects 'none,' then community rehabilitation and medication must not be selected. If the subject selects 'community rehabilitation,' then the subject is free to select 'medication' also but cannot logically select 'none.' Such a pattern (selecting 'none' *and* one or both treatment options) would indicate a problem with data quality; therefore, questions/items with *partial* mutual exclusivity can be useful QA/QC tools (*See:* Chap. 11).

If there is a single-response item with fully mutually exclusive answer options, then the database will contain only one variable for that item (Table 12.2). The example in Table 12.2 shows an alternative version of the question asked in Table 12.1. In this alternative version, there are four possible answer options that are fully mutually exclusive; therefore, the subject is only allowed to select a single response. In the dataset, the item will be indicated with a single variable, and the answer will be indicated with a 0, 1, 2, or 3 to indicate the selected answer.

12.3.3.2 Avoidance of Derived Database Variables

Only variables that cannot be derived from other variables should be entered into the database. Derived variables can be added to the database *at the analysis stage* after applying QA/QC measures. The following is an example of the type of problem that can arise if derived variables are entered into the database. An investigator collected data on each subject's date of birth (DOB) and date of measurement (DOM). Instead of entering this raw data into the database, (s)he has included the variable *age*, derived as DOM minus DOB. If errors in the DOM or DOB are corrected at a later stage (usually during the application of QC measures), this error will not be automatically corrected in the age variable, and analyses will have reduced validity.

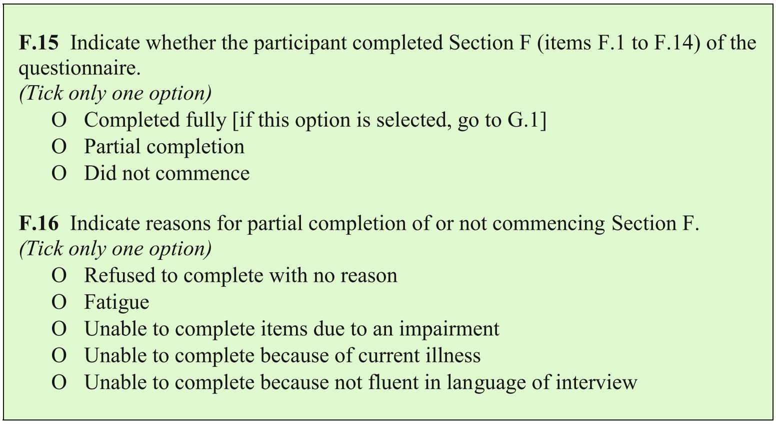
One of the few reasons to record a derived variable at the time of data collection or data entry is for 'live' databases, where such a variable is needed to trigger an essential action. For example, knowing the age of a participant might automatically help to apply inclusion/exclusion criteria.

12.3.3.3 Coding of Non-responses

Sometimes questions do not provide subjects with every plausible answer, and in these circumstances it is important to provide the option to select responses such as 'Don't know,' 'Unknown,' 'Not applicable,' and 'Other.' These response options prevent unanswered questions and forced answers. The early identification of questions/items that are prone to non-responses (or that have incomplete or illogical response codes) is an important function of data collection piloting so that timely improvements can be made.

An extension of this point concerns scenarios where questionnaires have embedded series of questions that are contingent on prior responses. For example, a list of 14 questions might be part of a validated diagnostic tool to identify individuals with a high probability of having pathologic anxiety. The first question in this list might be 'Have you had anxieties of any type in the past 12 months?' Mutually exclusive response options are 'Yes,' 'No,' and 'Not sure.' If the subject answers 'Yes' or 'Not sure,' then they progress through the 13 other questions, but if the subject answers 'No,' then they skip the next 13 questions. This is called a *skip pattern*. In the latter scenario, responses on forms and their corresponding database variables should be left blank.

If a subject willingly decides not to respond this will often lead to a blank item on the questionnaire and a blank corresponding variable in the database. Some questionnaires, however, foresee a response option 'Prefers not to respond' in the options list which can be coded. More generally, whenever special reasons for incomplete questionnaires are anticipated, then an option list can be created for interviewers to capture these reasons. This situation may arise if participants are reluctant to provide sensitive information, or the participant has 'interview fatigue.' In the case of assessments of children, it may be important to differentiate between children who do not have the capacity to complete a questionnaire or activity and those who simply got distracted or tired. This distinction may have implications for study validity. Whether an entire section of the form was not completed during an



F.15 Indicate whether the participant completed Section F (items F.1 to F.14) of the questionnaire.
(Tick only one option)

- O Completed fully [if this option is selected, go to G.1]
- O Partial completion
- O Did not commence

F.16 Indicate reasons for partial completion of or not commencing Section F.
(Tick only one option)

- O Refused to complete with no reason
- O Fatigue
- O Unable to complete items due to an impairment
- O Unable to complete because of current illness
- O Unable to complete because not fluent in language of interview

Fig. 12.1 Items to indicate reasons why a section in a questionnaire was not completed. A simple skip pattern is also included in item **F.15** for individuals who fully completed *Section F*

interview can be identified by having a question in the form about the success of its completion (Fig. 12.1).

All blank responses will have to be carefully scrutinized during pilot studies to develop optimal skip patterns and lists of response options. Final skip patterns require automated coding of variables to appropriately reflect skipping as the reason for missing responses. This means that even though the item is left blank on the questionnaire in line with skip instructions, the database program creates an automatic 'not-applicable' type response.

12.3.3.4 Pre-coding or Post-coding

During their initial design, questionnaires may have items that are captured as *strings or 'free text'* and that are later recognized to be amenable to multiple option coding. For example, an item may be aimed at recording medications being taken. In this case, the capacity of the interviewer to code this type of data should be determined before deciding whether coding should occur at data collection (pre-coding), at data entry (pre-coding), or during data analysis (post-coding). In our example, a lay-interviewer may find it easier to record as text the name of the medication rather than assigning a category such as 'anti-hypertensive' or 'anti-depressant.' The latter categorization is more easily done by a professional interviewer with clinical training. Thus, in the presence of a lay interviewer the coding of the database variable will have to be performed at the data editing stage (post-coding).

12.3.3.5 Variable Names and Attributes

Variable names should be carefully constructed. Considerations include that some software truncates variable names when they exceed a specified length limit, and that certain characters are allowed in variable names in some software but not in others. For longitudinal and compound (relational) databases, it is important to carefully test uniformity of variable names' lengths and other properties across tables so that 'merging' or 'concatenation' becomes hassle-free. Merging and appending data from multiple data files will not have accurate results if the variable attributes and arrangement of the unique identifiers chosen for merger have not been synchronized across all data files involved in such merging. Thus, to achieve uniformity of data, questionnaires should include consistent instructions about variable properties and units of measurement. For example, instructions should clarify that a particular continuous variable always be recorded as a 3-digit number with one decimal and in a specified unit of measurement; or that dates are always recorded in the DD-MMM-YYYY format (e.g., 07-JUN-2012).

Hint

Date and time variables should be formatted to allow calculations and extraction of time intervals.

12.3.4 The Database Inventory

A database inventory provides an overview of collected data and tracks data management activities. Each individual data table is listed in the inventory and organized to reflect the hierarchy of measurement and relational nature between tables. The database inventory should indicate the version of the data collection form that was used at the time of creation of a data snapshot so that associated changes in structure of the database can be monitored and/or accommodated in analyses. The inventory should further have a rolling summary of the number of records captured to date within each sub-database. An example of a spreadsheet that displays a database inventory is shown in Table 12.3.

12.4 Data Entry

As discussed in previous chapters (*See:* Chaps. 1 and 11), data entry errors cause information loss and can lead to biased study findings. Guarding against information loss and the introduction of bias are scientific and ethical imperatives in epidemiology; therefore, even the seemingly simple task of entering observations into a database is a matter of good scientific and ethical practice. In this section, we describe data entry systems and guidelines that help to reduce data entry errors.

12.4.1 General Guidelines for Data Entry Systems

In general, data entry screens should mirror the structure of paper data collection forms. This reduces data entry fatigue by limiting the amount of effort required to enter data. The same rationale is used to justify the implementation of simplified data entry

Table 12.3 Example of a database inventory created for a study involving children and their mothers. The inventory is updated on a weekly basis

Data table and form version	Unique identifier	Number of records expected	Number of unique records in electronic database	Number of paper records to be transcribed	Missing records	Number of variables	Action needed	Action performed & date
A. **Parent data table**								
Questionnaire1_ VERSION_2_2DEC2010	Mother_ID	100	95	0	5	10	Find missing forms; Mis-filed?	
Questionnaire2_ VERSION_1_1DEC2009	Mother_ID	200	200	0	0	12	None	Verified, archived 2-NOV-2010
B. **Child data table**								
Questionnaire3_ VERSION_1_2DEC2010	Child_ID	310	310	0	0	20	None	Archived 2-JUN-2011

processes that minimize typing, e.g., by using simple codes and 'drop-down' lists. The following are a list of approaches that can decrease the burden of data entry and therefore increase efficiency and accuracy:

- Create an ergonomic workspace for data entry clerks, e.g., providing them with a stand to mount paper forms side-by-side with the screen
- Implement data entry software functions that facilitate accurate data entry, e.g., clicking on a data entry field to trigger a selectable drop-down list displaying all possible answers for the relevant item
- Employ data error checks to immediately flag data for verification
- Use double data-entry methods (Day et al. 1998)
- Check all or a proportion (e.g., 10 %) of electronic entries against source forms
- Use logic checks, e.g., fields associated with a skip pattern are left blank if the leading question indicates that a series of questions should be skipped

In addition, the following guidelines may be useful for all research studies, even though they were developed for clinical trials (FDA 2007; only a selection is shown, with minor edits):

- Use electronic signatures for authority to proceed, i.e., a series of symbols authorized by an individual to be the legally binding equivalent of the handwritten signature
- Design a computerized system so that every entry can be attributed to an electronic signature
- Display the printed name of the subject on the screen during whole entry session
- Use a log-off system when a data entry clerk leaves a workstation, or, an automatic screensaver that prevents entry until a password is entered
- Change passwords regularly
- Put in place controls to ensure that the system's date and time are correct (e.g., synchronized with a trusted third party) and cannot be modified by entry personnel
- Include and describe help features for data entry, such as range and consistency checks to reduce errors or lists of codes that are compatible with the analysis plan

Textbox 12.2 Data Entry with Mobile Devices

Some steps in data entry are removed when mobile devices are used for data collection. An advantage is that absence of transcription eliminates transcription error. It is important to recognize that other sources of error still remain and that error rates may differ for different types of mobile technology (Patnaik et al. 2009). These considerations influence the choice of mobile device, the structure and format of data collection tools and the specific steps for training and monitoring of data entry.

12.5 Dataset Creation and Data Export

A well-designed database system should be able to extract data in the form of datasets and export these to statistical software packages. The exported datasets should preserve data definitions, variable labels, and entered values, and the exported datasets should be readable in standard statistical software programs, such as R, SPSS, and Stata. To ensure that database information can be exported and read by statistical software, it is recommended to include in an SOP routine checks for database-to-statistical software compatibility (this can be done on a monthly basis for large studies). With each export it is also important to verify that date and time variables and values have been exported properly.

12.5.1 Merging Data from Multiple Databases or Datasets

Merging procedures should be tested early, and for compound/relational databases the results of merging should be carefully reviewed to ensure that the correct units of analyses are used and that the number of records in the merged dataset is as anticipated. This is especially critical for datasets that share anything more complex than a one-to-one relationship (e.g., a single record in the parent table links to two or more records in the child table).

Anonymity of the data is to be maintained at all times through removal of participant identifiers. This includes names, addresses, and other identifying features (e.g., unique tattoos or markings) as well as GPS coordinates. If the latter is necessary (e.g., for geospatial analyses), then GPS coordinates should be exported as a separate file and then encrypted and secured with a password. All exported data should be safely stored and time-stamped.

12.6 Metadata

12.6.1 The Data Dictionary

A data dictionary is a central table that provides information about each table in the database as well as the data contained in each table. These metadata (i.e., data about data) will include names of individual tables and, within each table, many of the following:

- Variable names and labels
- Type of data (e.g., numeric, text, date)
- Format of each variable (e.g., 0,1,2…; yes, no, …; DD-MMM-YYYY)
- Some descriptive information on data contained in each field
- Codes
- Value labels

Table 12.4 is an example. For tables that contain derived variables or fields (age derived from date of interview and date of birth), the data dictionary will have to document the process of derivation.

Table 12.4 Excerpt of a data dictionary

Table name	Variable name	Variable label	Variable type	Variable format	Value labels
Parent	PARENT_ID	Unique parent ID	Numeric	String	None
	HOUSE_ID	House ID	Numeric	String	None
	SURNAME	Surname	Text		
	DATE_INT	Date of interview	Date	DD-MMM-YYYY	
	EMPLOY	Employed	Numeric	Numeric	1: No
					2: Yes
Child	CHILD_ID	Unique child ID	Numeric	String	None

Software applications like Epidata automatically create a data dictionary during the database design. The other option is to manually create the dictionary. The data dictionary makes it easier to achieve consistency in analytic and reporting stages. This 'memory' is especially useful for large or multi-site projects and those spanning a long duration.

12.6.2 File Names and Labels

Paper forms should be clearly labeled and named. The use of footers and headers with filename and version number and date modified should be standard procedure, and all study staff and investigators should follow computer file-naming rules. An example of a file-naming rule is:

<ProjectName>_<FileName>_<VersionNumber>_<DateModified>_<LastModifiedBy>_<FileExtension>
E.g., ProjectXXX_Clinical_Diagnostic_Questionnaire_V1.2_2011may31_SK.docx

12.6.3 The Form Inventory

As forms are being developed, there may be many versions and drafts of each version. The number of files can become overwhelming without an inventory list to track the different forms. Indeed, deploying the wrong form can be a major, potentially irreconcilable mistake. An example of form inventory is illustrated in Table 12.5. Only authorized personnel should be tasked with monitoring and maintaining such inventory. This should be centralized to prevent duplication and ensure better control of forms.

12.6.4 The Audit Trail

Despite great efforts during study design and piloting, form and database changes are unavoidable and in many cases indicate that there is a well functioning feedback loop among data collectors, data managers, and investigators. Especially when data are

Table 12.5 Example of a form inventory

Level of information	Construct	Name of form/ questionnaire	Translated version tracker	English version tracker
Demo-graphic	Child demo-graphic profile	Child demo-graphic form	ProjectXX_Child_Demographic_Form_Zulu_20110531_V1.2_SK.docx	ProjectXX_Child_Demographic_Form_Eng_20110431_V1.1_SK.docx
	Adult demo-graphic profile	Adult demo-graphic profile	ProjectXX_Adult_Demographic_Form_Zulu_20110531_V1.2_SK.docx	ProjectXX_Adult_Demographic_Form_Eng_20110431_V1.1_SK.docx
	Household profile, socio-environment factors	Household demo-graphic profile	ProjectXX_House_Demographic_Form_Zulu_20110531_V1.2_SK.docx	ProjectXX_House_Demographic_Form_Eng_20110531_V1.2_SK.docx

collected for novel hypotheses or in novel settings and populations, many potential database changes are likely to need consideration in the course of the study. An audit trail is a system that helps to keep track of changes made to the electronic forms and records. Audit trails should describe when, by whom, and the reasons changes were made to the database or other documents (e.g., forms). In order to maintain data integrity from collection to dataset export, only authorized edits (additions, deletions, and modifications) of the design and data elements should be permitted. Computer-generated time-stamped audit trails can facilitate tracking these changes especially if there are numerous edits that need to be made; however, sometimes it is necessary to supplement audit trails with diaries documenting changes and comments from colleagues. With recent developments in technology, the project diary can be shared among investigators and include logs of all project-related correspondence.

Audit trails must be secure and, when possible, computer-generated and time-stamped. They must also be readily accessible in a chronological format that allows immediate auditing. Personnel creating, modifying or deleting electronic data should not be allowed to change audit trails, but rather to add records whenever necessary. WORM (Write Once, Read Many) computer data storage systems allow the user to write data to such a storage system only a single time, but to read any number of times. This prevents the user from accidentally or intentionally altering or erasing data. Ideally, any changes noted in the audit trail should be attached to a tag in the database/document alerting the user that a change was made at some point (as well as when and by whom).

In this chapter we discussed issues related to data management. In making a data management plan, the goal is to facilitate making high quality data available for analyses. Analyses also require planning and will be the subject of the next chapter.

References

Day S, Fayers P, Harvey D (1998) Double data entry: what value, what price? Contr Clin Trials 19:15–24

Epi Info™ Centers for Disease Control and Prevention (2011) http://wwwn.cdc.gov/epiinfo/. Accessed Sept 2012

Epidata Software (2011) http://www.epidata.dk. Accessed Sept 2012

FDA/Food and Drug Administration (2007) Guidance for industry. Computerized systems used in clinical investigations. USDHHS-FDA, Rockville

Onyango AW et al (2004) Managing data for a multicountry longitudinal study: experience from the WHO multicentre growth reference study. Food Nutr Bull 25:S46–S52

Patnaik S, Brunskill E, Thies W (2009) Evaluating the accuracy of data collection on mobile phones: a study of forms, SMS, and voice. International Conference on Information and Communication Technologies and Development (ICTD), 2009: 74-84. Available at: http://dspace.mit.edu/handle/1721.1/60077. Accessed Feb 2013

Van den Broeck J et al (2007) Maintaining data integrity in a rural clinical trial. Clin Trials 4:572–582

The Analysis Plan

13

Jan Van den Broeck and Jonathan R. Brestoff

Plan A.

Abstract

Carefully designing an analysis plan is an important part of study protocol development. Analysis plans specify the chosen outcome parameters, the analysis procedures, and the way in which the statistical findings will be reported. Planned analysis procedures can include data transformations to prepare study variables, descriptions of sample characteristics, methods of statistical estimation, and methods of statistical testing. However, one cannot foresee every detail of how the analysis will proceed. Indeed, particularities of the data, unknown at the study's planning stage, will guide many decisions during the actual data analysis. This chapter therefore deals with general issues that arise in the preparation of an analysis plan and in the setup and approach to analysis, and provides a broad framework for analysis planning applicable to most epidemiological studies.

13.1 The Usefulness of an Analysis Plan

An analysis plan can be useful for the following purposes:

- To develop an overall analysis strategy that will be applicable if the collected data have anticipated distributional characteristics and quality

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_13,

- To help potential collaborators, research ethics committee, sponsors, and other stakeholders judge the proposal. A written plan allows readers to scrutinize statistical aspects of the study and to relate these aspects to the data collection, data handling, and even the object design (Miettinen 1985)
- To simulate outcomes under different scenarios, with foci on precision and power (*See:* Chap. 7). Attention can focus on the ideal and expected scenarios, but it is sometimes helpful to consider less likely scenarios as well. These simulation exercises can also provide insight into required study size and study feasibility.

See Panel 13.1 for a list of key terms and concepts relevant to analysis plans.

Panel 13.1 Key Terminology Used in the Description of Analysis Plans

Analysis dataset Selection of fields and records extracted from a database, used for a particular statistical analysis

Analysis plan Plan specifying the study outcome parameters, their calculation procedures, and how they will be reported

Analysis variables Variables representing determinants, outcomes, confounders and effect modifiers used in a analysis

Confidence interval A range of values within which the true population value is expected to fall, given the evidence in the sample data

Data Recorded information regardless of form

Data analysis Activities done to maximize and summarize relevant information contained in datasets, usually including data transformations and statistical analyses

Database Organized set of data kept as a source for extracting datasets

Data transformations Creation of derived variables in a database or dataset needed to facilitate analysis

Derived variable Variable whose values are not created on the basis of measurement but by transforming and/or combining values from existing variables

Interval estimates Confidence intervals: a range of values within which the true population value is expected to fall, given the evidence in the sample data

Null hypothesis A statistical hypothesis stating that two or more variables are expected to be statistically unrelated

P-value (of a null hypothesis) Probability of finding a value for the test statistic at least as extreme as the value obtained in the study, in a situation where the null hypothesis is true

Point estimates 'Best guess' estimates of the population value, inferred from the sample

Primary analysis Analysis carried out to produce evidence about the most important specific aim

Score Position of a measurement value on an ordinal or numerical measurement scale

(continued)

Panel 13.1 (continued)

Scoring Locating an individual measurement value on an ordinal or numerical reference scale

Significance level (of a test) A particular a priori P-value α used to label obtained P-values as 'significant' if the obtained P-value is smaller than α or 'non-significant' if the obtained P value is greater or equal to α

Statistical analysis The calculation of statistics to summarize some aspect of the data

Statistical estimation The calculation of point and interval estimates

Statistical methods Methods for sampling from sampling frames, summarizing and presenting data, estimation of population parameters and hypothesis testing

Statistical package A computer program especially designed to facilitate the use of statistical methods

Statistical testing Computation of a P-value (Miettinen 1985)

Stratified analysis Separate calculation of outcome parameters for different levels of suspected modifiers or confounders (usually but not necessarily accompanied with calculation of a pooled estimate)

Subgroup analysis Analysis done separately for one or for several levels of a presumed effect modifier

Syntax A set of instructions (statements) written by a computer user in a syntax screen within a statistical software package, submitted to trigger the execution of chosen procedures with the data

Table Matrix of rows and columns used to summarize relevant information about categories of interest

13.2 General Structure of the Analysis Plan

Figure 13.1 shows the general process by which one typically carries out an analysis and, accordingly, constructs an analysis plan. It is most efficient to construct the analysis plan in a linear manner, starting first with 'data extraction' and ending with 'statistical analyses.' This order is recommended because each major step tends to influence the next one. The structure of an analysis plan can recapitulate the order of this process.

13.3 Planned Data Transformations

Study variables represent determinants, outcomes, confounders, and effect modifiers used in analyses. Some variables are directly available from the database (e.g., reported sex or gender), but others may require derivation (e.g., age based on

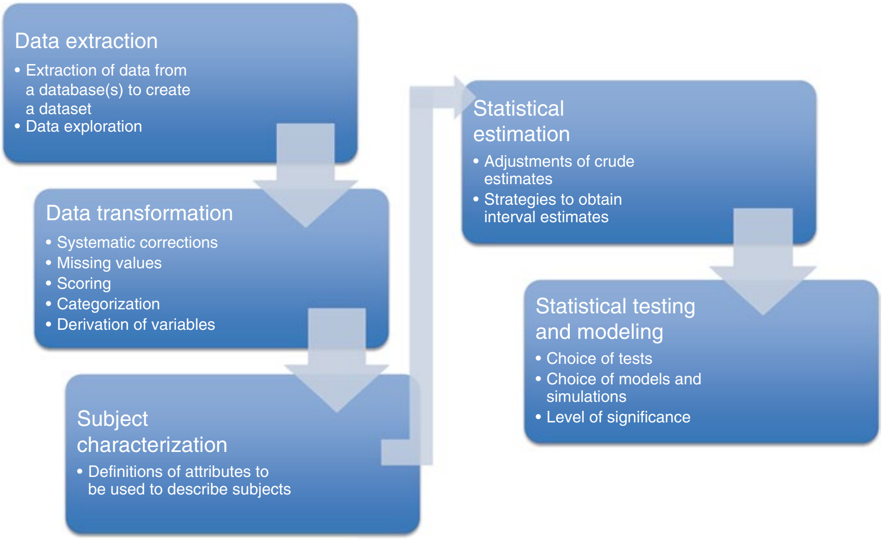

Data extraction
• Extraction of data from a database(s) to create a dataset
• Data exploration
Data transformation
• Systematic corrections
• Missing values
• Scoring
• Categorization
• Derivation of variables
Subject characterization
• Definitions of attributes to be used to describe subjects
Statistical estimation
• Adjustments of crude estimates
• Strategies to obtain interval estimates
Statistical testing and modeling
• Choice of tests
• Choice of models and simulations
• Level of significance

Fig. 13.1 General analysis process and structure of an analysis plan. This order is recommended because each major step tends to influence the next one

date-of-birth and date-of-measurement). The derivation of variables may involve re-arrangements of information, scoring, categorization, collapsing, and normalizing transformations.

13.3.1 Data Extraction and Re-arrangements of Information

The first step of every analysis plan relates to the extraction of an *analysis dataset* (also referred to as a *dataset*) from a database (i.e., source master-file). The dataset must include all information needed for the calculation of planned study variables. As mentioned in the previous chapter, if information from several databases needs to be combined, correct match-merging by unique subject identifiers (or, more generally, unique observation unit identifiers) is essential to avoid duplication of data. Before embarking on any systematic transformations one should explore all 'fields' (i.e., columns with data on a specific variable) as to the appropriate length, number of decimals, formats, allowable values, outliers, and inconsistencies. This exploration should give the analyst an indication of the general qualities of the data and raise any data management issues that need to be addressed before further work with the dataset.

13.3.1.1 Derived Variables

As mentioned above, some variables are directly available from the dataset but others need to be derived using a process known as *data transformation*. Perhaps the most common data transformations are the calculations of age from date-of-birth and date-of-measurement/enrollment, and 'time into study' from date-of-enrollment and date-of-measurement. In performing these or other transformations, statistical packages may require the analyst to recode missing values, such that all missing data have a uniform notation.

One should keep in mind that the precision of a derived variable cannot be better than the least precise element in the transformation used to make that derivation. For example, in a self-administered questionnaire, self-reported height may be recorded alternatively as number of centimeters (cm) or as number of feet and inches (ft, in.), according to the preference of the respondent. When these data are merged into a single variable, a precision problem arises because inches are less precise than centimeters; thus, the overall precision for height cannot exceed that of inches. As another example, if in the calculation of body mass index, weight was recorded with high precision, to the nearest 10 g, but height was recorded only roughly to the nearest 5 cm, the precision of the resulting body mass index variable would be very poor in spite of the precise weight measurements. One should be attentive to such problems whenever different measurement units were used during data collection.

Hint

The precision of a derived variable cannot be better than the least precise element contributing to that variable

13.3.1.2 Systematic Corrections

In some instances, there is a known systematic measurement error, and data can be transformed using a validated correction factor. For example, measurements during a certain period may have been made using a spare instrument that was discovered to be slightly decalibrated in comparison with the routinely used instrument. If the amount of inaccuracy is known and constant, a single correction value can be added to the relevant subset of values during analysis. Corrections may become more complex than this example, especially if part of the data were collected using a totally different method altogether (e.g., a study in which self-reported body weight is collected for all subjects and directly measured in only a sample thereof). In such instances, the relationship between values obtained with both methods needs to be the subject of careful regression modeling. Put another way, a prediction model must be constructed to replace the values obtained with the less reliable method with imputed values virtually obtained with the better method. Acceptable prediction may or may not be achieved with a simple conversion factor.

The issue of systematic corrections can also arise in situations where surrogate measures need to be used, such as when current exposure is to be used as a proxy for past exposure. Feasibility studies may have shed some light on historical exposure changes and may allow some corrections to be applied to the current exposure data (Esmen 1979; Cherrie et al. 1987; White et al. 2008). If such information exists, it is very helpful to reference it in the analysis plan and to develop a strategy for obtaining an acceptable correction factor.

13.3.1.3 Dealing with Missing Values

After missing values have been appropriately dealt with in data cleaning (*See:* Chap. 20), a question may arise as to whether, for analysis, one should remove from the dataset records with missing values for an important analysis variable. Statistical packages do this automatically for standard regression analyses. However, deleting records can introduce bias and reduce precision of the outcome parameter estimates. It may therefore be preferable to find an alternative solution, if a valid one exists, under the form of imputation, weighted regression, or adjustment for predictors of missingness (*See:* Textbox 13.1).

Textbox 13.1 Approaches to the Handling of Missing Data

The most appropriate solution to handling missing data depends on the process that led to the missing data, especially whether that process was random or systematic in some way. *Accordingly, the analysis plan may specify how patterns of missing values will be assessed and may foresee alternatives to complete case analysis* (*See:* Donders et al. 2006). With **complete case analysis,** one only uses data from observation units with complete data on the variables needed to calculate the outcome parameter estimate. The alternative may be some form of imputation, i.e., replacement of missing values with an estimated value. **Single imputation** based on regression modeling tends to overestimate precision, whereas **multiple imputation** performs better in this regard, as it takes into account the imprecision of multiple imputations (Little and Rubin 2002; Sterne et al. 2009).

Complete case analysis and imputations tend to lead to unbiased estimates only if the missingness is unrelated to the study variables. When missingness is systematically associated with the outcome event and data on other study variables are near-complete, as is often the case in prospective longitudinal studies with losses to follow-up, complete case analysis using multiple regression can be as valid as multiple imputation if proper statistical adjustments are made (Lewis 1999; Committee for Proprietary Medicinal Products 2004; Groenwold et al. 2011). Alternatively, current multiple imputation methods allow specification of various patterns of non-random missingness and can yield valid results. In **weighted regression**, one gives more weight to data from subject categories that are underrepresented due to missing data.

Non-recommended approaches include the use of missing data indicator variables and treating missing data as a level of a variable (Greenland and Finkle 1995). One should not create categories or indicator variables for missing values unless one is examining whether missing values for a variable are more likely in some groups than in others.

13.3.2 Scoring

Scoring is the location of individual measurement values on an ordinal or continuous measurement scale (i.e., on an *index*). It allows expression of a value's magnitude in reference to an expected distribution. Scoring is also helpful to make magnitudes of one variable comparable across levels of another variable, such as age and sex, as the reference distribution can be made age- and sex-specific (e.g., anthropometric scoring in children, discussed below). Scoring can be based on known indices (*external scoring*) or without pre-existing indices for that variable (*internal scoring*).

External scoring systems use an accepted reference distribution for the variable in question. A typical example of external scoring is anthropometric scoring, in which measurement values of, for example, height and weight are scored using accepted anthropometric indices, such as height-for-age or weight-for-height reference distributions. When the reference distribution is continuous, as is the case in anthropometric scoring, the scores can take the form of a centile position, or, more commonly, of a Standard Deviation Score (also called a Z score). The latter expresses the position of the measurement value within the reference distribution as the number of standard deviations away from the reference mean:

$$\text{Z score} = \frac{\text{Measurement value} - \text{Reference mean}}{\text{Reference standard deviation}}$$

Mean and standard deviation (SD) adequately describe a Normal or a Normalized reference distribution. When the reference distribution is non-Normal and without kurtosis, the Z score can be described adequately by *three* parameters: an L-value (skewness parameter), S-value (a parameter of dispersion), and M-value (median) (Cole and Green 1992). For example, the WHO child growth standards mainly consist of age- and sex-specific L, M and S values (WHO 2006, 2007). With such reference values the Z scores can in principle be calculated as follows:

$$\text{Z score} = \frac{\left(\dfrac{\text{Measurement value}}{\text{M}}\right)^{\text{L}} - 1}{\text{SL}}$$

In internal scoring, it is assumed that the measurement values have a particular underlying distribution – a distribution for which no valid external reference exists – that can serve as an internal reference distribution. To perform internal scoring, one often calculates an internal Z score (after a normalizing transformation, if necessary) for each measurement value based on the mean and standard deviation of the internal reference. Internal scoring is commonly employed to develop in-study scoring systems for latent variables based on multi-item questionnaires, though there are other uses. As with external scoring, internal scoring can also include adjustments for other variables (for an example, *See:* Van den Broeck et al. 1998; Francis et al. 2009).

13.3.3 Categorizing Variables and Collapsing Categories

Measured variables and scores are often continuous variables, but they will not necessarily be used in analyses as continuous variables. There are several good reasons to categorize continuous variables, and the analysis plan can be explicit about these:

- To create contingency tables that show how variables are distributed across levels of another variable
- To make a histogram (e.g., during data exploration)
- To prepare determinant variables with a non-linear relation to the outcome (e.g., a J–shape or U-shape) for analysis
- To prepare for stratified analyses aimed at controlling for confounding or to demonstrating effect modification
- To create indicator categories (e.g., hypertension or obesity)
- To prepare for subgroup analyses

13.3.3.1 What Number of Categories Is Optimal?

The answer to this question is context-dependent. For indicator categories the optimal number is usually two, as implied by the definition of the indicator (e.g., obese vs. non-obese), but more can be chosen if the object design calls for it (e.g., morbidly obese vs. obese vs. non-obese). For histograms, seven categories are often enough. For adequate control of confounding using stratified analyses and for evaluating dose–response relationships, four-to-five categories are usually sufficient.

Irrespective of context, though, there are two general rules to keep in mind. First, if more data are available, then more categories can be made (though this does not mean that more categories is better). And second, one should avoid having categories with sparse data if possible. If sparely populated categories exist in an analysis, the only viable solution may be to reduce the number of categories, perhaps by collapsing the spare category with a neighboring category.

13.3.3.2 Where to Place the Cut-Offs?

There is no generally accepted method to define cut-offs for a categorical variable, making this task prone to manipulation to obtain expected or statistically significant results. One of the most common methods to defining cut-offs is to use accepted indicator definitions (e.g., body mass indices of 25.0–29.9 and 30.0–34.9 are categorized as ‘overweight’ and ‘obese,’ respectively). Alternatively, if there is an unusually shaped distribution (e.g., peaks and gaps), natural cut-offs may become apparent. If neither accepted nor natural cut-offs exist, a common approach is to categorize data into centiles (e.g., tertiles, quartiles, quintiles, etc.).

By creating categories, one raises an additional issue: is it okay to have extreme categories that are uncensored (e.g., age 65+)? This creates a heterogeneous category somewhat incomparable with other categories. For example, age may be categorized by decades from 25 through 64.9 years with an uppermost category of 65+ years. The uppermost category will include subjects ranging in age from 65 to the oldest person in the study, whereas all other categories will range from 0 to 10 years only. The uppermost category can create confounding in analytical studies; consequently, in analytical studies one prefers closed extreme boundaries, even if it results in a category with small numbers.

13.3.4 Transforming the Distribution Shape of Analysis Variables

Knowledge or anticipation of the distributional characteristics of important study variables is essential for planning statistical estimations and testing. The analysis plan may specify how distribution shapes will be investigated and how any transformations of shapes will be done. By far the most frequent type of transformation is the Normalizing transformation, usually successfully done by replacing data values by their logarithm or by raising them to some power. Checking Normality can be done by a combination of approaches that may include:

- Histogram inspection
- Shapiro-Wilk test
- Kolmogorov-Smirnoff test
- Calculation of kurtosis and skewness statistics
- Q-Q plots

13.4 Description of Subject Characteristics

After having planned the data transformations, the usual next step is to plan the description of subject characteristics. The importance of this task is highlighted by its position in most epidemiologic papers: at the beginning of the results section. These descriptions can involve both tables and graphical displays and usually focus on relevant variables for the total sample and for determinant levels of interest (e.g., treatment groups), perhaps further stratified by levels of major effect modifiers, such as biological sex.

The normality of continuous variables must be examined to assess which measure of central tendency is appropriate to report. If the distribution is Normal or near-Normal, one traditionally reports the mean and standard deviation. Non-Normal distributions are often reported using the median, interquartile range (P25–75), P10–90, or range (max-min). To assess data distributions, the usual starting point is to graphically depict the data. Different graphical styles are preferred for different types of data:

- Histograms and box-plots are popular for displaying the distribution of continuous variables and to compare those distributions across subgroups
- Bar charts and pie charts are popular for displaying distributions of categorical variables with three or more categories

- The frequency of a single category (e.g., females) is popular for describing frequency distributions of dichotomous variables (e.g., males/females), as the frequency of the remaining category is easily implied

However, one should not use graphs to display data from two-by-two contingency tables.

This description of subject characteristics may include a description of the frequency distribution of the outcome variable (e.g., hypertension: yes or no) or of variables used to derive the outcome variable (systolic and diastolic blood pressure). One or both of these approaches are commonly taken when the actual outcome parameter (e.g., the prevalence odds ratio of hypertension in males vs. females) is not simply a distributional characteristic of the outcome variable. Thus, when reporting the findings of a cross-sectional study about determinants of hypertension, for example, a table could be planned to describe systolic and diastolic blood pressure by age and sex.

13.5 Statistical Analysis Strategy for Outcome Parameters

The next major step in planning an analysis is to describe the strategy that will be taken to compute the desired outcome statistics.

13.5.1 Primary and Secondary Analyses

It is usually recommended to specify a primary analysis and one or more secondary analyses. The former addresses what is seen as the main research question; study design and implementation are geared towards optimal validity and efficiency in creating empirical evidence about the question addressed in the primary analysis. Secondary analyses may address additional research questions of a different nature, or they may concern interesting sub-group analyses.

13.5.2 Estimation, Testing, or Both

There is a close link between statistical estimation and testing (Miettinen 1985; Rice 1988). For example, if the difference in the effect of two treatment levels is significant at the 5 % level (a matter of testing), then the associated 95 % confidence interval of the difference will exclude zero (a matter of estimation). Thus, the significance of a hypothesis test can often be inferred simply from inspection of two confidence intervals. Only estimation, however, can provide clear insight into what the *magnitude* of the parameter could be.

The choice between estimation, testing, or the use of both depends commonly on what the objective of the study is:

- If aiming to create evidence about the possible *existence* of a determinant-outcome relationship, with no ambition to actually quantify the magnitude or precise shape of such a relationship, then the choice for statistical testing of a null hypothesis is logical

- If aiming to create evidence about the *magnitude or shape* of a relationship whose existence is considered highly probable or certain already, then it may be possible to perform statistical estimation only, although there may be an additional perceived need to actually address the existence of the relationship with testing
- If aiming to create evidence about *both the existence and the magnitude or shape*, then it is logical to choose both estimation and testing

In the analysis plan, both estimation and testing can be the basis for sample size calculations (*See:* Chap. 7). Consider a study of the difference in systolic blood pressure between diabetics and non-diabetics. One may choose the sample size to ensure that each estimate of mean blood pressure is surrounded by a margin of error of a certain width. One may equally choose the sample size to ensure a certain power and significance level for a t-test.

Hint

In addition, one must decide whether the analyses will be performed using Bayesian or frequentist approaches. There is a great divide between these two statistical approaches, and no consensus exists in the field regarding one's blanket superiority over the other. This chapter deals only with analysis plans employing a frequentist approach.

13.5.3 Simulation of Potential Scenarios

Sometimes analyses can be simulated using hypothetical data; this process can be useful for estimating precision of the outcome parameters under a range of circumstances (such as expected distributions of confounders and modifiers), including extreme circumstances that could become realities. On the basis of this exercise, a refinement of the analysis plan may be possible, as one typically gains insight into how categorizations should be done and may also realize the need to adjust the planned study size.

13.6 Basic Choices in Statistical Estimation

This section gives an overview of basic choices in estimation that deserve mentioning in the analysis plan, without explaining the actual methods listed: Statistical estimation is more extensively discussed in Chap. 22.

13.6.1 Crude and Adjusted Estimates

Crude (unadjusted) estimates can often be obtained without resorting to regression analysis or other modeling approaches. However, because evidence in epidemiology is very often properly presented under the form of probability functions, regression modeling has become a predominant method in statistical estimation in this discipline.

It conveniently allows for the estimation of both crude and adjusted estimates and is commonly applicable to various diagnostic, etiognostic, prognostic, and methods-oriented research projects, as will be discussed in Chap. 24.

Crude estimates may need adjustment for a variety of reasons. The analysis plan may describe which adjustments will be considered and how. Examples of adjustments are:

- Stratifications, with or without pooled estimates
- Age standardization
- Adjustment for confounding
- Adjustment for measurement bias or imprecision
- Adjustment of one variable for another, by creating a composite variable incorporating information from both variables (e.g., Disability Adjusted Life Years lost and cost-of-intervention estimates adjusted for intervention efficiency)
- Calculation of robust estimators (e.g., down-weighting of outliers)
- Adjustment for clustering
- Adjustment for missing information

13.6.2 Strategies to Obtain Interval Estimates

Each crude or adjusted estimate needs to be composed of a point estimate and at least one interval estimate. There are three main options available for the calculation of interval estimates:

- Classical standard-error-based interval estimates
- Bootstrapping: estimation of the standard error and confidence interval of an outcome parameter based on the distribution of parameter values obtained in a large number of random samples with replacement of size n drawn from the original sample (of size n)
- Likelihood-ratio-based interval estimates

The aims of the study and the type of statistical analysis will greatly inform which interval estimation strategy is best for a given study.

13.7 Basic Choices in Statistical Testing

This section gives an overview of basic choices in statistical testing that should be mentioned in the analysis plan. Statistical testing is more fully discussed in Chap. 23. In epidemiology, null hypotheses are usually tested. Analysis plans tend to specify that null hypothesis testing will be performed; the chosen test(s), conditional on distributional characteristics; the choice of one- or two-sided P-values; and the level of significance.

The most important assumption underlying any statistical test is *full stochasticity*. Only if the null-hypothesis testing concerns an occurrence relation that is fully stochastic (as opposed to partly or fully deterministic by structure) does testing make sense. An example of flagrant violation of the assumption of full stochasticity – where null hypothesis testing would be meaningless – would be testing for the existence of a difference in body mass index between obese and non-obese persons, with obesity defined on the basis of body mass index. Before considering testing, it is wise to check if the determinant variable, or any variable from which it is derived, is computationally incorporated in the outcome variable. When that is the case, stochasticity may be compromised and null hypothesis testing meaningless.

13.7.1 Choice of Test

If the assumption of full stochasticity is not violated, then the planning of null hypothesis testing can move ahead. In Chap. 23 advice can be found on the choice of null hypothesis tests. This choice commonly requires determining or anticipating the following:

- The measurement scale of the outcome variable (categorical vs. ordinal vs. numerical)
- The distributional characteristics of an outcome variable if it is numerical (Normal vs. non-Normal distribution)
- Whether or not the determinant variable will be dealt with as a continuous variable
- The number of determinant categories/groups to be compared (single group comparison against a theoretically expected frequency distribution; two groups; or *k* groups)
- Whether observations in comparison groups are unrelated (unpaired or independent) or related (paired or interdependent)

13.7.2 One-Sided or Two-Sided P-Values

A P-value is the probability of finding a value for a statistic at least as extreme as the value obtained in a situation where the null hypothesis is in fact true. It is customary to carry out two-sided tests. If a one-sided test is used, this decision needs to be justified by showing that the expected difference between comparison groups can only go in one specific direction. For example, in a disaster area with a high burden of acute starvation, a study was done to look at whether young children were still growing in length during a 6-months observation period. The chosen statistical test was a one-sided paired t-test of length measured at baseline and after 6 months. A one-sided test was appropriate because children do not shrink in length.

13.7.3 Level of Significance

The concept of significance level will be discussed extensively in Chap. 23. In brief, it is a P-value threshold used for interpretation of the test result. A $P=0.05$ cut-off is usually chosen as a rough guide to evaluate how likely it is that the null hypothesis holds (with P-values lower than 0.05 considered to indicate that the null hypothesis is unlikely to hold), but this interpretation also depends on sample size, prior credibility of the null hypothesis, the number of tests that are being done, and biases. For example, in very large studies, a P-value of 0.001 can be found for a difference of a magnitude that is irrelevant or unimportant and can be easily caused by a small bias.

An important issue is when to do adjustments of the habitual $P < 0.05$ criterion. When the sample size is very large, the prior credibility of the null hypothesis very high, or many tests are done, most researchers tend to use lower levels of significance for interpretation, for example $P = 0.01$ or $P = 0.005$. Adaptations towards lower P-values are often advocated for repeated interim analyses, subgroup analyses, and multiple comparisons in the same study. The Bonferroni method adjusts the critical value for statistical significance (The Bonferroni method provides a Q value, which is an adjusted P-value) when multiple comparisons are being performed. For a critical discussion of this controversial topic, *See:* Chap. 23.

13.8 The Statistical Methods Section of a Study Proposal

Many sponsors and journals only require minimal information on the statistical analysis plan. The essential and required information almost always includes a clear description of which analysis is primary and which are secondary, and for each specific aim of the study an outline of:

- Any scoring systems to be used (and whether they have been validated)
- The chosen descriptive summary statistics
- Outcome parameters
  - Point and interval estimates
  - Adjustments
  - What tests will be used and why? How will the assumptions underlying the tests be verified? What is the level of significance?
- Target sample size and power of each analysis
- The chosen statistical software packages that will be used and whether the use of any software will involve any particular macros or syntaxes

In this chapter we sketched the usual main steps in data analysis and outlined a corresponding structure for the analysis plan. The essence of this plan needs to be described in a special section of the study protocol. In the study protocol, most sections on study design focus mainly on scientific aspects. However, important ethical issues also require adequate planning, and study protocols must contain a special section on relevant ethical issues. The next chapter introduces some important ethical issues that deserve special consideration in a study proposal and protocol.

References

Cherrie J et al (1987) An experimental simulation of an early rock wool/slag wool production process. Ann Occup Hyg 31:583–593

Cole TJ, Green PJ (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. Stat Med 11:1305–1319

Committee for Proprietary Medicinal Products (2004) Points to consider on adjustment for baseline covariates. Stat Med 23:701–709

Donders AR et al (2006) Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 59:1087–1091

Esmen N (1979) Retrospective industrial hygiene surveys. Am Ind Hyg Assoc J 40:58–65

Francis D et al (2009) Fast-food and sweetened beverage consumption: association with overweight and high waist circumference in Jamaican adolescents. Public Health Nutr 12:1106–1114

Greenland S, Finkle WD (1995) A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 142:1255–1264

Groenwold RHH et al (2011) Dealing with missing outcome data in randomized trials and observational studies. Am J Epidemiol 175:210–217

Lewis JA (1999) Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. Stat Med 18:1903–1942

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York

Miettinen OS (1985) Theoretical epidemiology. Delmar, New York, pp 1–359. ISBN 0827343132

Rice JA (1988) Mathematical statistics and data analysis. Wadsworth, Belmont, pp 1–595. ISBN 0534082475

Sterne JA et al (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Brit Med J 338:b2393

Van den Broeck J et al (1998) Fatness and muscularity as risk indicators of child mortality in rural Zaire. Int J Epidemiol 27:840–844

White E, Armstrong BK, Saracci R (2008) Principles of exposure measurement in epidemiology: collecting, evaluating, and improving measures of disease risk factors, 2nd edn. Oxford University Press, Oxford, pp 1–428. ISBN 9780198509851

World Health Organization (2006) WHO child growth standards. Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age. WHO, Geneva, pp 1–312. ISBN 924154693X

World Health Organization (2007) WHO child growth standards: methods and development. Head circumference-for-age, arm circumference-for-age, triceps skinfold-for age and sub-scapular skinfold-for-age. WHO, Geneva, pp 1–217. ISBN 9789241547185

Ethics Support

14

Emma A. Meagher, Tracy S. Ziolek, and Jan Van den Broeck

Ethics and Science need to shake hands.

Richard Clarke Cabot

Abstract

In order to facilitate the conducting of successful epidemiological research, with the utilization of human subjects, several elements of protections for the subjects, as well as an internal resource for the investigator, make up the framework of a research support program. The program must include an Ethics Committee and may have several other ancillary Committees (e.g., a Conflicts of Interest Committee). The elements of research support programs, and the role of the investigator and their research team, are described within this chapter, along with how the two entities may interact. Finally, the expectations for investigators to facilitate ongoing oversight while research is being conducted, via monitoring, is outlined, along with the overall benefits of an organized program for providing ethics support to those conducting research with human subjects.

E.A. Meagher, M.D. (✉)
Department of Medicine, Institute for Translational Medicine and Therapeutics,
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: emma@exchange.upenn.edu

T.S. Ziolek, M.Sc., CIP
Institutional Review Board, University of Pennsylvania, Philadelphia, PA, USA

J. Van den Broeck, M.D., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_14,

14.1 Research Participant Protection Frameworks

The ethical conduct of ethical epidemiological research involving humans is a primary obligation shared by all members of the research community including, the research team, the institution, the Ethics Committee (EC) or Institutional Review Board (IRB), and, the subject, the funding agency, the sponsor and national or federal agencies (Fig. 14.1 and Panel 14.1). This shared responsibility can only be achieved when all constituencies have an adequate understanding of the ethical principles that underpin the research process.

In addition to application of ethical principles, the EC/IRB will consider the science of the proposal to verify that the research objectives appear attainable and that the plan for evaluation of the data provides the best potential for achieving generalizable information in the intended area of science. Some of these scientific elements will align with the research regulations (e.g., enrolling the proper population of subjects or using the appropriate data to answer the research question), but the overall assessment of the science relates to whether an EC/IRB should allow a study to expose a subject to any element of risk, for research that doesn't seem well supported by current needs of the scientific community.

Panel 14.1 Selected Terms and Concepts Relating to the Organization of Ethics Support

Clinical monitor Person designated by the trial sponsor to check if the actual trial procedures conform with study protocol, standard operating procedures and regulatory requirements

Conflict of interest (in research) Study personnel's personal, financial or other interests that may intentionally or unintentionally influence the procedures and outcomes of a research activity

DSMB Data and Safety Monitoring Board. Independent oversight committee installed by a research sponsor in support of a particular ongoing study, charged with the regular review of data quality and participant safety, and advising investigators and sponsor on these

Ethics Committee Independent committee that decides, based on independent ethical review, about initiation or continuation of a research study

Expedited review A procedure for documents submitted for review to an ethics committee to be reviewed not by the full committee during its scheduled meetings, but either by a smaller number of committee members outside the scheduled full committee meetings, or by appointed designee(s) of the IRB Chairperson. The criteria for approval under an expedited mechanism are the same as the criteria for approval by the convened IRB.

(continued)

Panel 14.1 (continued)

Independent ethics review Evaluation of planned or ongoing research in terms of respect for dignity, rights, safety and wellbeing of individuals and communities and fairness of distribution of the benefits and burdens of the research among all groups and classes in society

Informed consent process Process of fully informing potential study subjects about the study and of obtaining their voluntary agreement to participate or to continue participation

Investigator A person responsible for the conduct of a research study

IRB Institutional Review Board. *See:* Ethics Committee

Minimal risk study Study in which the harm to participants anticipated on the basis of the apparent nature of interventions, contacts and measurements, is judged by the ethics committee to be not greater than risks run ordinarily in routine medical checkups or psychological tests.

Protocol amendment Written description of a change made to an earlier version of the official study protocol

Sponsor An individual, company, institution, or organization that takes responsibility for the initiation, management, and/or financing of a study

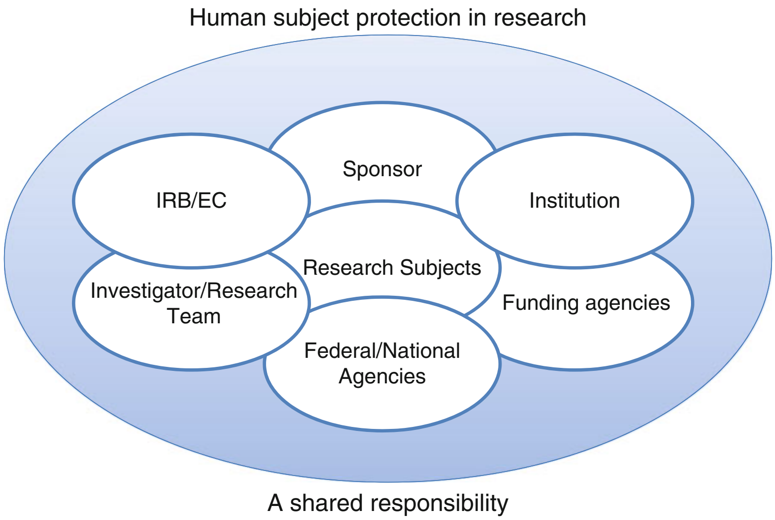

Human subject protection in research
IRB/EC
Sponsor
Institution
Research Subjects
Investigator/Research Team
Funding agencies
Federal/National Agencies
A shared responsibility

Fig. 14.1 Human subject protection: a shared responsibility

Various organizations around the world have created guidelines for ethical conduct in human subject research. The World Medical Association's Declaration of Helsinki and the US Belmont Report are the most commonly referenced and comprehensively describe the three originating fundamental principles of ethical research conduct: i.e. respect for persons, beneficence, and justice (World Medical Association 2010; The Belmont Report 1979). Derived from these fundamental principles are codes of conduct such as, for example, the International Ethical Guidelines for Epidemiological Studies from The Council for International Organizations of Medical Sciences (2009), and Good Clinical Practice guidelines for trials.

The codes of conduct and the general philosophy that has contributed to the evolution of these codes have resulted in the research community adopting an expanded evaluable framework from which one can determine that human subjects' research is ethical (Emanuel et al. 2000). This framework, if adhered to by all members of the research community, as previously characterized, significantly enhances the likelihood of human research being carried out in an ethical manner. Among the major structural elements in the framework are the informed consent process (*See:* Chap. 16) and independent review, the latter being one of the main topics of the present chapter.

Adherence to general codes of conduct by all constituencies is enforced by numerous regulatory agencies throughout the world. For example in the US, the Department of Health and Human Services, Office of Human Research Protections enforces regulations to ensure ethical conduct in research. In addition, the Food and Drug Administration also imposes regulations concerning human subjects' protection as it relates to the evaluation of therapeutics and devices. One approach to formalizing the protection of subjects (or their affiliated data) is to develop Human Research Protections Programs (HRPPs), which can function as the official frameworks created at research sites/institutions where human subjects research is carried out. They are designed to facilitate an understanding of and compliance with the necessary regulations. Below we expand on this mode of organizing ethical support.

14.2 Elements of a Human Research Protection Program

The fundamental components of a HRPP and their individual responsibilities are listed in Table 14.1 and are described in greater detail below. The components of an HRPP that exist at any research site will be determined by many factors including the type of research being conducted. Additional components, other than those listed in Table 14.1, may be an office that handles processing of grants and contracts that fund the research protocol (these may be referred to as the Office of Research Services, Office of Research Support Services or Contract/Grants Office). For institutions that also conduct complex biomedical research there are typically several additional elements to the HRPP including: a scientific committee responsible for oversight of high risk studies such as "first time in man" research, or those involving gene transfer or cellular therapy or committees that oversees studies that use novel advanced diagnostics. Finally, an additional element may be a Community Advisory Board. This is a committee of individuals representing the views of a community in which the study takes place. The common goal is to maintain informed participation with research activities perceived as relevant and feasible by the community. This committee is often established in large community-based studies.

Table 14.1 Important elements of a human research protections program

Element	Responsibility
Ethics committee/IRB	Independent review of proposed research
	Ongoing oversight of research conduct
	Service to the research community: subjects, research team and the institution
Conflict of Interest Committee	Independent evaluation of possible individual and institutional conflicts of interest
	Creation of a conflict of interest management plan
Monitoring and auditing (Clinical monitor, Data and Safety Monitoring Board)	Independent evaluation of patient level or aggregate data to ensure human subject safety, data integrity and appropriate research conduct
Research team (Investigator, research nurses and support staff)	Propose scientifically and ethically sound research
	Conduct the research in keeping with the approved protocol
	Uphold the highest ethical standard to maximize safety and respect participants
	Respect the opinions of the community of subjects expected to participate
Research subjects	Be fully informed about what they are agreeing to do
	Adhering to the requirements of the study
	Asking questions when information is missing or unclear

Within an HRPP the responsibility of the research subject is to ensure that (s)he (1) understands what participation entails (2) has a clear appreciation of the possible risks, benefits and alternatives, (3) proactively seeks clarification and maintains and open dialogue with the research team and (4) adheres to what is required of them as a subject.

Below we further discuss the other fundamental elements of an HRPP.

14.3 The Ethics Committee

The one common element to all HRPPs is the independent Ethics Committee (EC), in some countries referred to as the Institutional Review Board (IRB). The EC/IRB structure may vary greatly between institutions and may be designed to support the type of research conducted. The EC/IRB is typically supported by a dedicated office that is staffed with administrators who are career professionals with expertise in human subjects research. These individuals are responsible for a preliminary review of all ethics committee submissions and making a determination of what level of review is required (see below). In addition the staff organizes the convened committee meetings. A review committee is typically composed of members who are physicians, scientists, nonscientists, and individuals who are not affiliated with the institution conducting the research. EC/IRB staff and committee members adhere to common ethical principles and apply the codes of conduct referred to above. ECs/IRBs are required to review all human research; however the level of review will be determined by the risk of the proposed research. Whilst the initial emphasis is placed, by the research community, on achieving approval for a research protocols from the review committees, many of the compliance and other oversight issues don't arise until research activities are underway. Thus ECs/IRBs are required to (1) establish that the initial criteria for approval are met and (2) provide ongoing assessments of ethical research conduct throughout the life cycle of the research study to include review at least annually of research activity, review of modifications to the research protocol, review of deviations from the approved protocol that may increase the risks of harm to research subjects or may negatively impact the integrity of the data, and review of unanticipated events that might impact human subject safety.

14.3.1 How Does the Independent Ethics Committee Review Process Work?

Once an investigator has established an idea for a research proposal, (s)he has to consider what approvals will be required from the relevant ethics review committees and any other human subjects' research oversight office(s) prior to initiating any research-specific activities with human subjects. It is important to distinguish activities that are considered preparatory to research (i.e. assessing records for feasibility) versus actually beginning the research. An easy way to differentiate these activities is to consider whether information you derive from your assessment will be solely for verification that the site has a good representation of the available population needed for enrollment into the research or whether you will be collecting specific information relating to these potential subjects for inclusion into the research dataset for analysis.

When the research proposal has been written, the corresponding EC/IRB application materials should be completed. Many offices now have electronic submission options and templates available to assist the investigators with their submissions. Templates are typically very useful because they will incorporate any "required" language agreed to by the institution and may provide guidance text for the investigator/research staff to use when completing the document.

The overall expectation of risk based on the description of research procedures is the catalyst driving the level of review required. Research that is considered to be no greater than minimal risk and falls within one of the categories of research that is defined by the regulations can be reviewed in an expedited manner by an EC/IRB staff member, whereas research that is greater than minimal risk and minimal research not falling within the defined categories eligible for expedited review will require review by a convened committee. For example, research questions answered by querying existing data sets or bio-repositories are generally considered minimal risk activities, in contract to research questions that are answered by comparing two interventions would typically be considered greater than minimal risk. Common types of procedures are listed in Table 14.2 together with an example of how their associated level of risk and required level of initial review could be defined by an ethics committee. When compiling the application, the investigator will need to have an idea of what level of ethical review their application will initially require. This level of review may define the processing time needed by the EC/IRB to complete of the review.

Table 14.2 Examples of research procedures and corresponding levels of risk and initial review

Procedure	Risk	Level of initial review	Average time to review (Note: Time will vary from site to site)
MRI (without contrast)	Minimal	Expedited	1 week
Urine collection, nail clipping	Minimal	Expedited	1 week
EEG/ECG/ECHO	Minimal	Expedited	1 week
Collection/analysis of identifiable data	Minimal	Expedited	1 week
Blood draw ≤ 50 mL in *patients* in an 8-week period and ≤ 2 sticks per week	Minimal	Expedited	1 week
X-ray/CT/PET	Potentially minimal	Convened	4 weeks
Skin biopsy	Potentially minimal	Convened	4 weeks
Investigational Intervention (drug/device/biologic)	Greater than minimal	Convened	4 weeks
Invasive procedures solely for research purposes	Greater than minimal	Convened	4 weeks

The regulations that are used to establish if the criteria for approval are met will be determined by the research type and design. If a clinical trial is being conducted to evaluate the safety and/or efficacy of a drug or a device an additional set of international standards, Good Clinical Practice guidelines, will be applied (*See:* Chap. 21: Good Clinical Practice). Recognizing that data derived from studies that assess investigational agents are used to support approval of those agents, trial data must be verifiable and reproducible, and most importantly must have been derived from an ethically rigorous protocol. Once the research team has submitted the proposal to the ethics committee the formal review will occur. This process is systematic and requires that the proposal is assessed to determine if the criteria for initial approval have been met.

14.3.2 The Initial Criteria for Ethics Approval

1. Risks to subjects are minimized: The ECs evaluate the proposal to ensure that the procedures/interventions (a) are consistent with sound research design, (b) do not unnecessarily expose subjects to risk, (c) are necessary to achieve the proposed objectives of the research and (d) whenever appropriate, are also required for diagnostic or treatment purposes as predicated by the provision of standard medical care.
2. Risks to subjects are reasonable in relation to anticipated benefits to subjects, and the importance of the derived knowledge that may be reasonably expected: In evaluating risks and benefits, the ECs consider only those risks and benefits that may result from the research (as distinguished from risks and benefits of therapies subjects would receive even if not participating in the research).

3. Selection of subjects is equitable: In making this assessment the ECs evaluate whether the subject population that will be invited to participate is ethically appropriate for the study question; i.e. that only those subjects who will be necessary to support meeting the objectives of the research will be enrolled and no subjects will be denied enrollment. When some or all of the subjects are likely to be vulnerable to coercion or undue influence, such as children, prisoners, pregnant women, mentally disabled persons, or economically or educationally disadvantaged persons, additional safeguards must be included in the study to protect the rights and welfare of these subjects.
4. Informed consent will be sought from each prospective subject or the subject's legally authorized representative and documented. The ECs evaluate the informed consent document to ensure that all required information is included and subject autonomy is preserved. The elements of informed consent are described in Chap. 16. The EC also evaluate the proposed consenting process recognizing that the existence of an informed consent document, and any other related informational materials that are provided for subjects when considering participation, must be coupled with an appropriate process for obtaining and maintaining voluntary informed consent at the point of enrollment and throughout the subjects' participation in the research study.
5. Respect for enrolled subjects: The ECs evaluate whether adequate plans for monitoring research subjects and the data collected exist to ensure the safety of subjects during the execution of the research in studies that are considered to be greater than minimal risk. That the proposed research allows for subject withdrawal without compromising the rights and welfare of the participant.
6. Privacy and confidentiality: The ECs assess whether there are adequate provisions to protect the privacy of subjects and to maintain the confidentiality of data. In certain types of research, such as retrospective chart reviews and epidemiological studies, the risk of loss of confidentiality exists. Indeed in many situations this may be the only anticipated risk of the research. In these situations, having special protections in place for storing and analyzing data mitigates the likelihood of a breach of confidentiality. Additionally, the ECs review the research procedures and the consenting process to ensure that the study team is striving to protect the privacy of the research participants.

14.3.3 EC/IRB Decisions

The ECs/IRBs may make the following decisions upon initial review of a research protocol:

1. Approve
2. Withheld approval pending the research team making minor changes and clarifications to the protocol or consent document
3. Table with requests for significant changes and/or explanations that affect human safety
4. Disapprove

Typically, the most common decision is withheld approval. When this decision is made the submission is returned to the research team to address the deficiencies and resubmitted to the EC for final approval. Tabled submissions typically require more in-depth revision by the research team. The most common deficiency in tabled protocols is a failure on the part of the investigator to adequately address two key ethical principles: Respect for Persons (appropriate targeted population of subjects and appropriate process for initial and ongoing consent of subjects) and Respect for Enrolled Subjects (an appropriately defined research plan that balances risks to benefits successfully). The EC staff or the chair of the committee will communicate directly with the investigator to facilitate the revisions. When the EC receives the revised submission it is returned to the committee for further review. It is unusual to disapprove a submission. Most protocols can be revised so that they ultimately meet the criteria for approval.

14.3.4 Review of Protocol Amendments

Protocol amendments proposed by investigators and sponsors need approval by all ethics committees involved. Some examples of important amendments:

- Any change to the informed consent form or process
- Request for adding ancillary study components
- Change in treatment strategy imposed by new laws and regulations, by compelling changes in standard of care, or by the discovery of unexpected rates of side-effects
- Expansion of sample size, study area, recruitment and enrollment strategy and period, frequency of contacts, number or wording of potentially sensitive questions and volume of biological samples
- Correction of major previously undetected errors in the study protocol
- Some amendments entail suspension of enrollment e.g. major changes proposed for enrollment procedures. Ethics committees should have a system for rapid decision if enrolment needs to be suspended or not

14.3.5 Protection of Participants After Study Cessation

Ethics committees also pay attention to the safety and rights of participants after study cessation. For patients, study cessation, whether early or foreseen and for whatever reason, may entail safety issues. Firstly, sudden cessation of certain types of medication is known to have unwanted, possibly serious, side-effects. To avoid this, there should be a transition period of gradual dosage decrease and intensive treatment monitoring around the time of cessation of individual follow-up. The participants' health care providers must be involved in this process. Secondly, study cessation may imply a return to local standard of care. This switch may again carry safety risks that need to be handled appropriately. The ethics committee may decide that the best solution includes a period of continued treatment with the test regimen, an extra window of study follow-up, or a referral to the patient's doctor for a gradual monitored change to another regimen. If several ethics committees are involved in a study the proposed solutions can differ or even be contradictory among committees.

14.3.6 Collaboration with Other Ethics Committees

In multi-country studies and multi-center studies there may be an issue of collaboration between several ethics committees, each linked to a different data collection site. It may also arise in single-site studies when there is a local ethics committee in the country of data collection and another committee in the country of the investigator's institution. The ethics committees involved may take contradictory decisions and impede progress with study preparations. Their level of insight into local study setting and local standards of care can vary substantially. They may also have widely different review times, ranging from a few weeks to a few years. At this stage no guidelines seem to be available for collaboration and harmonization but, clearly, ethics committees have some responsibility towards each other, if only in respecting and recognizing each other's areas of special competence.

14.4 The Conflicts of Interest Committee

There has been an increasing focus on the potential for conflict of interest in biomedical research, particularly as it relates to the research involving drugs and devices. The ethical principles applied to the conduct of research require that the semblance of a conflict, whether perceived or real, is disclosed to the research participant prior to them agreeing to participate in the research study. This information is included in the informed consent document. The Ethics Committee will review the description of the conflict and will opine on whether the description is adequate. In addition, a Conflict of Interest Committee will conduct a formal assessment of the conflict and will determine whether the conflict is manageable. A manageable conflict is one where the conduct of the study and the analysis of the data will be managed in such a way that the outcome of the study will not be influenced by the participation of the conflicted member of the research team. The Conflict of Interest Committee will determine the most appropriate management plan to mitigate the conflict. In some centers or institutions the conflict of interest assessment may be incorporated into the ethical review as conducted by the EC/IRB and no separate committee will exist.

14.5 The Role of Investigator and Research Team

Throughout this epidemiology book, ethical responsibilities of epidemiologists relevant to particular stages of the research process are mentioned alongside the scientific validity and practical guidelines relevant to the same stages. Chap. 1

Panel 14.2 Elements of a Research Proposal Important for Ethics Review

- Background and Rationale
- Objectives and Specific Aims
- Study Design
- Study procedures
- Study Population (inclusion and exclusion criteria)
- Assessments (safety and efficacy)
- Monitoring plan (for more than minimal risk studies)
- Ethical justification (*See:* Panel 14.3)
- Finances
- Publication plan
- References

Panel 14.3 Elements for Inclusion in the Ethics Section of a Research Proposal

- Relevance, feasibility, and value
- Justification for study population and recruitment process
- Assessment of risk and benefit and how risks will be minimized
- A description of the informed consent process
- How privacy and confidentiality will be maintained
- Plans to protect subject safety
- Research oversight

gave an overview of major responsibilities of epidemiologists and their link with the originating fundamental ethical principles and with codes of conduct was pointed out.

The investigator is held to the highest standards of ethical research conduct. (S) he is responsible for the conduct of the other members of the research team. Human research protection begins here. The investigator's responsibilities include designing the research protocol to include all elements listed in Panels 14.2 and 14.3; submitting it to all necessary review bodies and securing approval prior to initiating any research activities. Following receipt of approval and study initiation the investigator is required to:

- Adhere to the approved protocol, actively monitor the safety of participants and adhering to the approved monitoring plan (more details in the next section)
- Seek permission from oversight bodies (EC, monitors, etc.) whenever possible prior to deviating from the approved protocol
- Report to oversight bodies when unanticipated events occur that can impact human subject safety

- Modify the study procedures in a timely fashion when necessary to ensure ongoing subject safety
- Provide an annual progress report to the EC/IRB
- Keep subjects informed of any changes to study procedures or safety profile inclusive of any changes that may affect their willingness to participate in the study
- Maintain data integrity and to ensure that all research-related documentation is protected, yet accessible for review by monitors and auditors when necessary

14.6 The Monitoring Plan

Oversight of human subject safety is of paramount importance and all clinical trials require monitoring. Different levels of monitoring can occur and the degree of monitoring is typically driven by the type of research being conducted. For example, in a study that involves the comparison of two 'standard of care' treatment regimens, it may be appropriate that the investigator her/himself monitors the safety of enrolled subjects. In Phase-1 and early Phase-2 clinical trials it may be sufficient to identify a single clinical monitor, independent of the study team, who is willing to evaluate ongoing oversight and provide independent safety assessments when requested to do so either by the investigator, the sponsor or the review bodies such as the EC/IRB. This common type of monitoring involves site monitoring visits, which will be further discussed in Sect. 21.6.

For more complex Phase-3 clinical trials the monitoring plan may include the existence of an independent Data Monitoring Committee, also called a Data and Safety Monitoring Board (DSMB). This is a group of individuals with pertinent expertise that, on a regular basis, reviews accumulating data from an ongoing clinical trial. The DSMB advises regarding the continuing safety of current participants and those yet to be recruited, as well as the continuing validity and scientific merit of the trial. The committee may decide that the clinical trial (1) may continue without interruption, (2) requires modifications, (3) be suspended or terminated.

14.6.1 DSMB Composition

The selection of DSMB members is extremely important as the DSMB is assigned critical responsibilities in protecting the safety and well-being of trial participants. Membership should include individuals with expertise in the medical management of the condition under study, biostatistics, clinical trial conduct, ethics and clinical pharmacology. All members must be devoid of serious conflicts of interest and must agree to maintain confidentiality of the interim results they have reviewed. In some circumstances a representative from the intended community of subjects will be also be included. A DSMB may have as few as 3 members, but may need to be larger when representation of multiple scientific and other disciplines or a wider range of perspectives is desirable. For logistical reasons it is sensible to keep the DSMB as small as possible, while still having representation of all needed skills and experience.

14.6.2 The DSMB Monitoring Plan

All DSMBs should have a well-defined monitoring plan that is documented in the form of a DSMB charter. The topics that are typically addressed by the DSMB include (1) schedule and format for meetings, (2) interim analysis plans and format for presentation of data, (3) specification of who will have access to interim data and who may attend all or part of DSMB meetings, (4) procedures for assessing conflict of interest of potential DSMB members and (5) the method and timing of providing interim reports to the DSMB.

14.7 The Benefits of a Human Research Protection Program

One of the greatest advantages of a functional HRPP is that all components work together to optimize the research enterprise and maximize human subject safety. In addition, it provides every member the opportunity to understand the contribution and role each plays in this endeavor. While the primary responsibility of the HRPP is to verify that protection of human subjects is being adequately addressed, a secondary goal of the HRPP should is to facilitate the research process. For example, if a clinical trial is being conducted at multiple sites, the regulations allow for review of the multi-site trial by a single IRB. However, the practice at most institutions is to require the investigator to seek approval from her institution's EC/IRB. Not too surprisingly, multiple reviews leads to multiple opinions and often contradicting decisions. In a effort to address, many institutions are beginning to allow alternative and collaborative review models that will markedly reduce effort on the part of the research team and time to achieve approval from reviewing entities. An additional example of an attempt to reduce burden is the creation of a shared electronic submission system, which with appropriate alerts facilitates communication between all individuals engaged in the research process. Going forward the focus will remain the same – human subjects' protection, but the manner in which it is achieved will continue to become more efficient.

This chapter was the final chapter of Part II: Study Design. The cross-cutting topic of Part II was the development of a study protocol, which has important implications for planning projects and developing proposals to obtain financial support for them. In Part III: Study Conduct, the overarching topic will be the implementation of the study protocol, starting with training and study preparations (Chap. 15).

References

Council for International Organizations of Medical Sciences (2009) International ethical guidelines for epidemiological studies. Council for International Organizations of Medical Sciences, Geneva, pp 1–128. ISBN 929036081X

Emanuel EJ, Wendler D, Grady C (2000) What makes clinical research ethical? J Am Med Assoc 283:2701–2711

The Belmont Report (1979) Federal register, vol 44, no 76, Wednesday, 18 Apr 1979, notices, pp 23192–23197

World Medical Association (2010) The declaration of Helsinki. http://www.wma.net/en/10home/index.html. Accessed Sept 2012

Part III

Study Conduct

Chapter 15: Training and Study Preparations
Chapter 16: Managing the Informed Consent Process
Chapter 17: Accrual, Retention, and Adherence
Chapter 18: Questionnaires
Chapter 19: Maintaining Data Integrity
Chapter 20: Data Cleaning
Chapter 21: Good Clinical Practice

Training and Study Preparations

15

Jan Van den Broeck, Shuaib Kauchali,
Jonathan R. Brestoff, and Meera Chhagan

It is impossible for a man to start learning what he thinks he already knows.

Epictetus

Abstract

Study implementation starts with the necessary training of personnel, planning of logistics, and establishment of infrastructure. At this stage of the study, management and training skills as well as practical experience become as important for the investigators as theoretical scientific skills. Literature on this important aspect of a study is sparse; therefore, in this chapter we present experience-based recommendations for preparing for the implementation of a study.

15.1 General Management of Training and of Study Preparations

Identifying members of the research team with strong management skills is one of the first tasks to complete when preparing for a study. The principal investigator is in a position of leadership and is a natural choice to serve as the primary manager, but

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

S. Kauchali, M.Phil., FCPaed • M. Chhagan, Ph.D., FCPaed
Department of Paediatrics, University of KwaZulu-Natal, Durban, South Africa
e-mail: kauchalis@ukzn.ac.za; Chhagan@ukzn.ac.za

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_15,

other members of the investigative team usually are responsible for managing some or most elements of the study. For example, it is common practice for investigators with many projects to delegate part of the management responsibilities to a study coordinator, a data manager, and other managers. Their level of involvement and responsibility may need to increase gradually because these persons may need some training themselves before becoming independent in the full range of their charged tasks. It is thus up to the investigators to make sure that the delegation of management duties is planned appropriately, taking into account that, by the time 'real' data are collected, all managers need to be fully immersed in the study and in constant communication with each other about harmonizing and complementing tasks. Panel 15.1 introduces selected terms relevant to this chapter

When hiring managers one can, among others, verify their familiarity with some of the management goals that need to be kept in mind during the study preparation phase. Goals for the study preparation phase are:

- To identify and document roles and responsibilities of each member of the team
- To train all study personnel to perform optimally before start of data collection; develop training material and plan sufficient time for practice in mock runs
- To get the full team operational before starting enrollments

Panel 15.1 Selected Terms and Concepts Relevant to Training and Study Preparations

Certificate An official document testifying of an individual's successful completion of a specific course or educational program, or of another significant educational or professional achievement

Fieldwork Study implementation activities carried out in places away from the study coordination center

Manager Member of study personnel responsible for the practical organization of a defined set of study procedures

Manual of operations *See:* Operations manual

Measurement standardization Making sure optimal measurement procedures are outlined, trained for and uniformly applied

Operations manual Manual describing in detail the procedures to be followed by study personnel during the execution of specific research activities

Protocol adherence Implementation of procedures prescribed by the official study protocol

Standard Operating Procedures (SOP) Detailed written instructions to achieve uniformity of the performance of a specific function

Study personnel All official employees and volunteers involved in research study operations

Supervisor Member of study personnel responsible for quality assurance and control regarding other personnel's performance of specific tasks

Training course A detailed plan to enhance learning about a specific topic or task

- To document all regulatory and ethical approvals
- To establish reporting and monitoring processes
- To inform all stakeholders before starting field operations
- To avoid or minimize delays in the start of operations
- To hold regular feedback meetings with staff to inform them of the progress of the study and discuss challenges
- To work with team members to develop acceptable solutions to recruitment challenges, data integrity problems, data flow issues, data management, and reporting

Many of the management goals in the study preparation phase are oriented towards achieving the target study start date or minimizing delays to starting the study. It is therefore helpful to know the most frequent bottlenecks for starting studies. In our experience, these include issues related to:

- Approvals from sponsors and ethics committees
- Ensuring support from communities
- Development of SOPs
- Money transfers
- Recruitment of personnel and training
- Logistics readiness (including laboratory preparedness)
- Printing of questionnaires or readiness of mobile devices for data collection

Even after addressing the issues listed above, it may be necessary to do additional groundwork before embarking on the main study's data collection. Examples of common preparatory activities include mock runs of all study procedures, ascertaining competency of each member of the team in assigned tasks, establishing routines, and ascertaining participant and community acceptance. For studies that involve biologic materials, compliance with safety precautions through procedural skills, use of personal protective gear, appropriate waste disposal and subsequent handling of biologic material need confirmation.

Hint

Before starting the data collection phase it is useful to have a pre-initiation checklist to verify one last time that all is in place, and to put copies of Standard Operating Procedures (SOP) at appropriate locations in case procedural questions arise.

15.2 Selection of Personnel for Enrollment and Data Collection

The success of data collection starts with well-considered selections of personnel that will be involved in enrollment and data collection. No gender, age, or race discrimination are acceptable in hiring study personnel; however, for reasons of study acceptability and validity, one may prefer hiring personnel with a particular profile. For example, female interviewers may be preferable in a community-based study on sensitive female reproductive health issues. But even in this particular study, there may be subjects who prefer to speak with a male interviewer. Personnel profiles are complex but can generally be adequately summarized by their professional background, social skills, motivations, and skills/talents.

15.2.1 Professional Background of Personnel

It is usually advantageous to hire persons with experience working on similar types of studies, though it is not feasible in most circumstances to assemble a team entirely composed of such individuals. In addition, it is preferable to avoid hiring staff that are overqualified for the job, as this scenario can decrease motivation, reduce willingness to learn from training, and increase the likelihood of quitting the job during the data collection period. Thus, one must strive to balance various professional backgrounds. Some additional points to consider include the following:

- Secondary school level education is usually sufficient for surveys, questionnaire administration, and simple measurements (e.g., anthropometry)
- The candidates should have no emigration plans for the study's data collection period
- In community-based studies, interviewers' knowledge of the study area and living in that area may be preferable to being a stranger in an unknown territory. On the other hand, participants may prefer to be interviewed by strangers if highly sensitive information is being collected

15.2.2 Desirable Social Skills of Personnel

Communication skills are crucial for recruitment and enrollment activities (such as conducting the informed consent process) and for collecting data. The capacities to empathize, to express an ethical attitude, and (if relevant) to be 'good with children' are among the determinants of a successfully conducted informed consent process. For data collection, good communication skills are always important but especially so for conversational interviewing (*See:* Sect. 18.4) and for follow-up studies where one needs to build good long-term relationships with subjects.

Data collectors must speak the local languages except small minority languages (though not every data collector must be able to speak every language spoken in the area). Non-verbal communication issues also have their importance. They can affect the participant-interviewer relationship, especially during face-to-face interviews or during any other measurements requiring close proximity. In such cases it is good to pay attention to:

- Culturally appropriate dress code
- Personal hygiene
- Body language; having a warm voice may help interviewers, particularly during telephone interviews
- A confident demeanor may inspire confidence in participants, but on overly confident demeanor can cause some subjects to feel uncomfortable

Individuals with good social skills are usually able to succeed in a team-based work environment. Indeed, good functioning of the team requires quality interactions among its members, but it also requires a clear distribution of tasks and responsibilities. Candidates must accept the idea of being supervised and not perceive supervision as an act of distrust. Data collectors sometimes have to work in small teams, e.g., in anthropometry or for fieldwork in 'difficult' locations (rough terrain or unsafe). Good personal relationships within such field teams are crucial.

Hint

Where field work involves household visits, adequate preparation is needed to ensure easy identification of field workers by members of the communities involved in the study; this facilitates acceptance and hence response rates.

15.2.3 Motivation of Prospective Personnel

Poor motivation can affect accuracy of recorded data, lead to various forms of disrespect towards participants, and reduce work efficiency. Each of these issues has the potential to considerably slow down the research process. Keeping motivation of personnel high is certainly a responsibility of the investigator, but baseline motivation should be present already at the stage of hiring, as shown by:

- Agreement with proposed remuneration scheme
- Interest in contributing to research
- Eagerness to learn about the research project
- Willingness to be trained

Ongoing motivation is enhanced by regular feedback to personnel on progress of operations; encouraging discussion on project goals and procedures; sharing project milestones with operational staff; and, for studies with a long duration, providing appropriate raises and promotions to those who earn them.

15.2.4 Measurement Skills or Talent

Last but not least, one should verify the relevant measurement skills of prospective data collection personnel, or at least the talent and aptitude for measuring.

- Basic writing skills and – for some tasks – mathematical skills are a pre-requisite
- Specialized types of measurement may require prior experience and qualifications such as for phlebotomy, ultrasound examinations, and use of GIS equipment
- Anthropometry requires good visual acuity (with or without corrective lenses) to read values from instrument displays and good fine motor skills
- Data collection personnel should be inclined to handle instruments and source documents carefully
- Candidates must have the ability and discipline to meticulously follow detailed standard operating procedures. They should not be easily distracted when focusing on a task and should not have a tendency for satisficing

15.3 Who Needs Training and Who Should Train?

15.3.1 Who Needs Training and Why?

In any study it is rare to have any persons who do not need any training or retraining at some point during the study.

15.3.1.1 Training for Investigators and Study Coordinators

The principal investigators and study coordinators are never too experienced, skilled, or intelligent to undergo some degree of training for a particular study. This training may be relatively informal and consist of an active effort to ensure that one's skillset is in accordance with the highly detailed study protocol. Indeed, in clinical research, the study protocol and operations manual may prescribe procedures that are more detailed than usual in a clinical care context. These procedures might even differ from one's traditional approach, thereby necessitating re-training for study purposes. This is especially crucial where new types of devices and tools will be used. Investigators and study coordinators need to be exceptionally well versed in the procedural aspects of a given study if they intend to participate in data collection or train other staff.

15.3.1.2 Training for Enrollment and Data Collection Personnel

For fieldworkers and other data collection personnel one purpose of training is to be well aware of the important aspects of the study protocol. They should also be prepared for answering questions from the part of the participants and be trained in research ethics. Data collectors further require training to use appropriate communication skills, to be sensitive when addressing concerns of participants, and to know when to consult the supervisor. Finally, they need training on the use of devices and other tools / instruments.

Hint

In studies with long data collection periods there can be considerable turn-over of personnel and new recruits tend to undergo less extensive training than the initial group. This pitfall may be avoided by training some **reservists** similarly in respect of methods, schedules, and certification criteria.

15.3.1.3 Data Handling Personnel and Technical Collaborators

Personnel involved in data handling, laboratory personnel and technical staff needs training to acquire a good understanding of data collection and data handling procedures. They should be especially aware of flow of procedures between different sections of the study and their role in interacting with other components. An example would be the response of the data team to missing data or to observed problems with legibility. For data handling personnel and technical staff it is very important to acquire or refresh skills in Good Clinical Practice or Good Clinical Laboratory Practice, as applicable.

15.3.2 Who Should Train?

Investigators, study coordinator, and managers will usually be the key trainers. Investigators and study coordinators may take on a variety of training roles, but managers should introduce the trainees to the specific domain of activities they manage. The head of the research institution or a person delegated by her / him may welcome the trainees and introduce them to the institution's organizational structure and infrastructure. Furthermore it is important to put trainees in contact with same-level personnel with prior experience but also with first-in-line supervisors. Generally, trainees are very curious and sometimes anxious to get to know their fellows, their supervisors, and the methods of supervision. External experts may be called in for specific training, for example for particular measurements or for patient counseling, e.g., HIV counseling and testing or psychosocial counseling. Research projects may have several capacity strengthening aims incorporated, and these require planning so that they happen at appropriate stages of the study and maintain staff motivation.

15.4 Training Modules

In our experience it is worthwhile training all study personnel on the following topics:

15.4.1 Introduction to Scientific Research

Research is a systematic activity aimed at achieving new knowledge, but not all research is scientific. Journalists, for instance, can do research but they do not (usually) do so in a scientific manner. By eliciting, during the training, some characteristics of the scientific methods used in epidemiology one easily arrives at principles and responsibilities that are relevant to all study personnel, such as the pursuit of validity, standardization of procedures, documentation of procedures, and quality control measures. It may also be useful to clarify basic concepts like precision and bias.

15.4.2 Elementary Training in Research Ethics and Good Clinical (Laboratory) Practice

Formal research ethics training with certification is a Good Clinical Practice requirement for study personnel of clinical trials. However, ethics training is recommended for all epidemiologic studies. For English-speaking trainees with access to the Internet, this can be done online, for example on the website of The United States National Institutes of Health (2011). Academic institutions may also offer online courses and certification. In low- and middle-income countries, it may be a better option to design a local research ethics training guide or adapt an existing one (particularly for data collectors in field studies).

A first notion to convey during this training is that for research to be ethical, it should be relevant, feasible, and valid. Further ethics training must focus on such basic ethical principles as autonomy, non-maleficence, beneficence, and justice/fairness as well as basic epidemiological principles (*See:* Chap. 1). The main

Textbox 15.1 The Importance of an Ethical Attitude

Knowledge of ethical principles and the following of guidelines that are derived from them are insufficient to guarantee the ethical conduct of research. The other pillar of ethical research is the expression of an **ethical attitude**.

The need for ethical attitude implies, first of all, that study personnel should be empowered to use internalized ethical principles as a guide to making decisions in difficult or **unforeseen situations** for which no clear guidelines exist. This should be accompanied by a keen sense of when it is necessary to consult with a study coordinator, physician, or principle investigator before taking action. Case scenarios can be useful for learning about management of dilemmas and other unforeseen crisis situations.

Secondly, an ethical attitude is often necessary for **good relationships** among members of the study team. Responsibilities to colleagues and collaborators, to the community at large, and to sponsors and other stakeholders are sub-domains of research ethics that have remained relatively underdeveloped in comparison with the sub-domain of research participant protection. It can be devastating for individual researchers, members of study teams, and/or entire studies when some persons are treated unethically. This goes for hierarchical relationships within the study as well as for how personnel operating within the same hierarchical level treat each other. Each member of study personnel should have regular chances to express concerns, make suggestions, and be taken seriously. An open discussion can be held with the trainees on how this will be done in the upcoming study. As far as relationships among collaborating scientists are concerned, respect for colleagues includes treating them as scientists, e.g., responding in a scholarly way to their scholarly concerns about a scientific issue.

concern should be for each trainee to understand how ethical principles translate into concrete day-to-day responsibilities, or, viewed from another angle, how there is an ethical dimension to all decisions and activities (*See:* Textbox 15.1). For discussions of these ethical and epidemiologic principles, we refer readers to Chaps. 1, 16, and 21.

15.4.3 Training on Study Protocol, Study Organization, and General Expectations

All members of the study team must know important aspects of study protocol, organization and context, so that all are capable of explaining the study to interested parties. To achieve this, the following aspects commonly need to be addressed during training:

- There may be a need to give an introduction about the health-related issue investigated in the study.

- Information should be given about the community and setting in which the research will take place. The study's organizational structure with the main lines of responsibility and reporting must be clear. It is useful to have an organizational chart clarifying this.
- It is necessary to outline field operations and quality assurance and control activities. The concept of measurement standardization must be conveyed to all. Sometimes studies employ treatment regimens that differ slightly from national or regional standards of care. For trials this is true almost by definition. It may then be difficult to convince data collection personnel, health workers and sometimes participants that a particular regimen should be adhered to. Field staff should be committed to the regimen or regimens used in the study and should be empowered to deal with questions and objections.
- One should outline to all personnel the points of view and expectations of the different stakeholders of the study regarding projected achievements of the project and general performance standards of each type of study personnel.

15.4.4 Job-Specific and Task-Specific Training Modules

Crucial for smooth study operations is that everybody should know exactly what is expected of them. This should be reflected in unambiguous job descriptions, task lists, and SOP. Tasks that require intensive technical training usually include tasks related to bio-measurements (*See:* Chap. 10), questionnaire administration (*See:* Chap. 18), and data handling (*See:* Chap. 12). Training should cover but not be restricted to technical acts and procedures; training should also deal with the principles behind the guidelines and all other aspects that normally form part of an operations manual, such as timing and scheduling, getting the equipment ready, informing and preparing participants, personal preparation for the task, recording of information, limits of responsibility, reporting of difficulties, risk management, communication with stakeholders, and general administration. During these training elements, objective job-specific performance standards must be made clear to all.

Hint

Consider training personnel for more skills than just for those that will be needed for their routine tasks. People generally like to broaden the range of their skill set, and any study can use reservists for unforeseen circumstances.

Staff who are in direct contact with study participants may occasionally encounter situations where there is a need to refer participants to public services. It is important to provide some training for this task. An example is participants who need referral for co-existing conditions to medical or social services. The identification, referral process and anticipated handling by the participant and services need to be clarified. Simulated case scenarios may be used to establish an acceptable referral procedure.

Panel 15.2 Considerations for the Format of Task-Specific Training Modules

- Use the operations manual as one of the training tools
- Discuss and practice the correct implementation of SOPs repeatedly
- Consider various training aids, e.g., a video showing a correct measurement act, online training modules, group discussions, etc.
- Always incorporate individually supervised training of technical acts
- Leave opportunity for discussion and for improved local adaptation of operational procedures
- Document each individual's level of skills before the start of data collection, preferably using intra-observer accuracy and reproducibility statistics (*See:* Chap. 11)
- Document team-level performance before start of data collection, preferably using group accuracy and reliability statistics (*See:* Chap. 11)
- Provide formal certification of the successful completion of training

15.4.5 Training Format

General training is best organized as a group event during which new personnel get the opportunity to talk with each other and with investigators, study coordinators, managers, supervisors, and perhaps stakeholders. Job-specific or task-specific training modules are usually best organized in stages. The first stage of a session can again be a group event, but at some point individual hands-on training may need to be provided, followed by a period of practice under supervision. Sometimes formal testing of skills with subsequent certifications of aptitude is appropriate (such certificates can be useful motivators during the training sessions). For the format of task-specific training, consider the recommendations listed in Panel 15.2.

15.5 Infrastructure and Logistics

When establishing the infrastructural and logistical environment of a study, the investigator may be confronted with some difficult choices and be forced to make some compromises. Consider, for instance, the issues of sharing, borrowing, and using materials originally intended for purposes other than research (e.g., personal vehicles, cellular phones, computers, desk space, etc.) On the one hand, use of these materials could lead to cost savings and streamlining of activities. On the other hand, the quality of already used materials might not be adequate, a scenario that could endanger study quality. The safer approach is to try and develop a study environment that is rather autonomous with strict regulations about uses of dedicated equipment and space.

15.5.1 Space Requirements and Office Equipment

'Borrowing' research space from service sites such as schools, hospitals, and work environments necessitates respect for scheduling systems and etiquette prevailing in that environment. This can be accommodated through formal agreements or memoranda of understanding, with the ultimate goal being the establishment of an environment that is conducive to the activities of all parties. In clinical care settings, overlap of research with routine clinical activities may become necessary, such as overlap in the use of routine waste disposal systems. In school settings, especially those with limited space available for researchers, special consideration needs to be given to the appropriateness of available space for the nature of research activities, scheduling around exams, sports events, vacations, and class schedules. Irrespective of whether the borrowed space is in a hospital, industrial setting, or school, the expectation generally is to limit disruption of routine non-research activities at the study site. These logistics can be tested in the study's preparatory phase.

Logistics of space allocation need careful consideration. A useful example is that of studies in which investigators are conducting interviews that result in sharing sensitive personal information. One should aim to achieve an operational set-up that not only allows the interview to take place in adequate privacy but also allows the participant some 'personal space' away from a busy waiting room. It is also important to prevent participants being identifiable by other participants or community members as having a stigmatizing condition. This can easily occur if field-workers known to be involved in a study that only enrolls HIV-infected participants are seen making regular visits to specific homes in the community. Figure 15.1 is a floor plan of a research site organized with the intention to avoid stigma.

In addition to the types of space shown in Fig. 15.1, a study may also need:

- Desk space for study personnel, with computers (if necessary) and office supplies
- Fridge and freezer space for samples (must be separate from fridge and freezer space for foods and beverages)
- Pharmacy space with secure drug cabinets and refrigerators with back-up power
- Source document archiving space with secure filing cabinets

During study preparations, the optimal functioning of each study space must be established. This optimization process usually begins with strategizing the use of available space so that the participants are ensured privacy during interviews, the data team has ready access to instruments, and the filing systems allow easy retrieval of source documents. If biologic samples, medications, or disposable test kits are stored, responsibilities for handling inventories and expiration dates are described in an SOP and assigned to a qualified staff member.

In Chap. 12 advice can be found on the choice of hardware and software. All hardware and software should be available and tested during the preparatory phase. Ongoing availability and licenses should be confirmed for the duration of the study. If researchers are using open-source software, then support systems should be identified prior to the initiation of use.

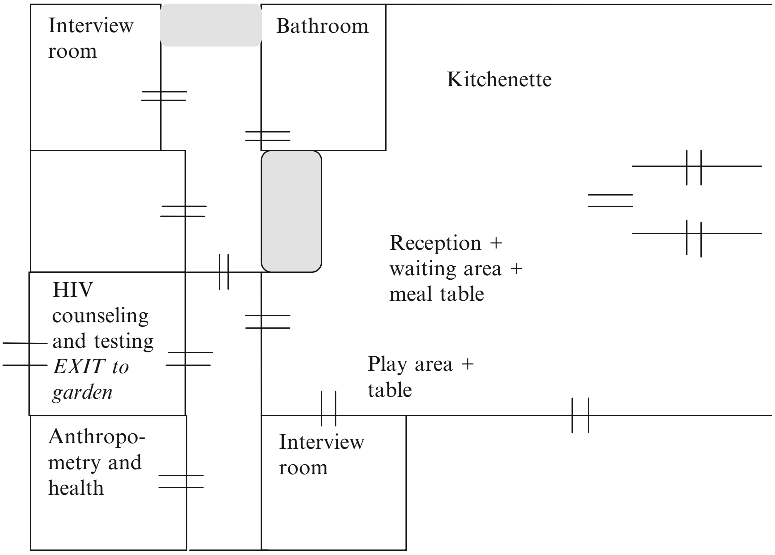


Fig. 15.1 Floor plan of research site. Private entrances are denoted by || and there is separate exit from the HIV testing room directly to the garden, bypassing the common waiting area. These design elements help to protect the privacy of subjects

15.5.2 Supply Systems, Local Transport, and Communication

Transport and communication solutions are especially important when recruitment, enrollment, interventions, or data collection happen in several geographically distant places. Persons and supplies must reach their destinations efficiently and in good condition. In turn, source documents and samples need return to the study coordinating center or laboratories promptly to ensure data integrity. This indicates a need for good coordination between field logistics and data management teams.

15.5.2.1 Supply Systems and Transportation

Supply systems tend to become more efficient when the number of intermittent steps is decreased and when systems for various items or activities are combined and harmonized, perhaps even with other studies or with existing health care supply systems. It is important to preserve and monitor the good condition of all the moving persons and objects, e.g., by ensuring an optimal temperature during transport. Likewise, storage of items must be under conditions that preserve their integrity and any undue exposures and expiration must be monitored. These guidelines may translate, for instance, into setting up fridge and freezer temperature tracking systems, establishing efficient stocks and flow management systems, and putting in place various safety regulations and contingency plans. One might also consider the purchase of a sufficiently powered electricity generator and a CO_2 backup system for freezers if biological samples need to be transported or stored at a certain temperature.

Depending on pre-existing means of transport and on considerations of cost and timing, budgeted solutions for transport may include subsidized personal transport; the use of public transport; and / or the purchase of vehicles, motorbikes, or bicycles for the project. Large research institutions may even establish their own car rental system and charge individual research projects for the use of vehicles, fuel, and maintenance services.

Hint

Study preparations should include administrative procedures for maintaining transport logbooks, procurement procedures, and a database of suppliers.

15.5.2.2 Communications

Difficulties during field operations require fast communication possibilities. Depending on the existence of wireless networks or other communication systems and considerations of cost and timing, budgeted solutions for fast communication may include subsidized air time for personal cellular telephones, use of public telephones, purchase of cellular phones and air time for the project, Internet-based communication, etc. Such communication systems are obviously in addition to the much-needed opportunities for face-to-face communication and regular meetings. It is important to realize that communication should not only be about what goes wrong but also about what goes well.

15.5.3 Printing

Printing of forms is done in at least two batches: one small batch for piloting, followed by one or more larger batches of documents that were amended based on the pilot. Expedited review by ethical oversight committees may be required for substantial changes made to printed documents in the piloting phase. Printing may be needed of:

- Questionnaire forms, including adverse events forms
- Standard operating procedures
- Manuals of operation
- Study information sheets
- Informed consent forms and participant statement forms
- (Barcode) labeling system of samples

Note that, even if data collection is designed to be paperless and based on mobile devices, a paper back-up system may be desirable, especially in resource-poor areas with unstable Internet connectivity or erratic electricity supply. Depending on the available printing capacity within the project, study coordinators may decide to outsource printing jobs or use in-house production. Especially with regular in-house printing of data collection forms, responsibility should be assigned to ensure that correct form versions are printed, and that formatting is maintained.

For data collection forms in clinical trials, the choice may go to a write-through type of form (e.g., carbon copies) that consists of duplicate or triplicate sheets. Commonly, one sheet is then needed for the patient file and one for data processing and archiving.

15.5.4 Laboratory Readiness and Technical Collaborations

In some clinical studies, consideration needs to go to the question if one should set up a special new research laboratory or use an existing one. If an existing laboratory needs to be contracted, the question frequently is: should it be a local laboratory that will perhaps need some capacity building to deal with the specific analyses for the study or, alternatively, one that has very high standards but is located farther away? There may also be a need for one or more small transportable laboratories to process samples before transport. In these matters local capacity building is often preferable but sometimes difficult. In resource-poor areas especially, capacities of local laboratories may be already stretched because of purely clinical work. Using an existing lab may still require extra training, quality control, and bio-safety precautions. Thus, there may be instances where local capacity building would become an unaffordable cost for a single research project. Laboratory capacity building can take considerable time and effort and is a responsibility that may transcend the individual investigator, as it also concerns institutions and health authorities. There are examples (Wertheim et al. 2010) of networks of hospitals and research institutions that have tried to enhance laboratory capacity in a region. Investigators and sponsors may wish to check the possibility of initiating or joining such initiatives.

15.6 Preparatory Information Gathering

In prospective studies, before starting enrollment and data collection, there is commonly a need to gather various types of information for the purpose of fine-tuning methods and procedures. The specific purposes of preparatory information collection may include those listed in Panel 15.3:

Panel 15.3 Possible Purposes of Preparatory Data Collection in Prospective Studies

- To construct an up-to-date sampling frame
- To learn more about the nature and optimal measurement of relevant attributes
- To fine-tune the sample size and power calculations and the analysis plan via estimation of frequency of modifiers, confounders, and population size
- To develop a new measurement tool or adapt a measurement scale to a local context
- To fine-tune questionnaires, user's manuals, and informed consent form
- To optimize efficiency in logistics and quality control
- To learn about the required level of liaison and awareness among stakeholders and how best to achieve this
- To get an idea about likely enrolment rates, potential for follow-up, likely frequency of particular reasons for refusal, and overall feasibility of a study

Methods employed for the creation of this preparatory information may include:

- Rapid assessment procedures
- A census
- A preparatory survey, perhaps combined with a census
- A measurement scale development exercise
- Methods-oriented studies
- Simple pilot test runs
- Focus group discussions and other qualitative research methods

Rapid assessments and preparatory surveys are especially relevant for large population-based studies. They can provide useful information, among others, for:

- Establishing a statistical sampling frame
- Documenting characteristics of the target population
- Describing the distribution matrix of exposures, effect modifiers, and confounders in view of refining object design and to inform study size planning
- Announcing the upcoming main study (a role in recruitment)
- Assessing perceptions around the upcoming main study

The aim of rapid assessments can also be the definition and description of the exposures if these are unclear from the outset. This is often the case in fact-finding investigations in occupational and environmental epidemiology. Thereby it is often useful to define exposure zones within the work or living environment and use these as strata within which study participants will be sampled. This exercise has been called ‘zoning’ (Corn and Esmen 1979; Corn 1985). The advantage of this approach is that one can maximize representations of persons from extreme exposure zones in the study sample. White et al. (2008) provided a good summary of the definition and use of exposure zones.

Sometimes rapid assessments aim at finding correction factors or equations for exposure measurement. For instance, when past occupational exposure to an agent must be estimated by current exposure, a small study may yield correction factors that take into account changes in production processes, physical characteristics of agents, protective devices, and output of the product of interest (Esmen 1979; Cherrie et al. 1987; White et al. 2008).

The preparation of an epidemiological study may require an extensive preparatory sub-study to develop or optimize a particular measurement tool. This will often be needed when one or more of the attributes in the occurrence relation cannot be measured directly but only indirectly via a series of questions in a questions-based measurement tool. This is particularly frequent with attributes that are mental-behavioral characteristics. The development and adaptation of questions-based measurement tools is discussed in Chap. 10 (The Measurement Plan).

15.7 Pilot Test Runs

Subsequent to all preparatory activities already described, and prior to embarking on formal data collection, a pilot test run is conducted. During this pilot all procedures are conducted to mimic the real study’s situation and context. After such a pilot run, all study teams and investigators meet to share feedback and strengthen communication. During such feedback, compliance with protocol and standard operating procedures are confirmed or adapted; completeness of data and range checks verified; and electronic data entry systems tested. The latter may include documenting the time required for and the accuracy of transcription. All steps, from data collection to data export, should be clearly tracked and sources of errors documented in a study diary or logbook.

Pilot testing the database is coupled with piloting the process of data entry. Associated with the latter are documenting the time required for and accuracy of transcription, as well as piloting the data verification procedures. The lag between data collection and electronic database update is established and taken into account when designing monitoring plans. Backup routines are tested during study preparation. This includes testing whether the back-up routine is successful with multiple data entry personnel working in parallel. Strict organization of current and archived data folders is needed to ensure data integrity. Data export routines should be assessed for syntax errors or 'bugs'. This is the best opportunity to amend these errors.

In this chapter we started our discussion of various aspects of study implementation. Here, we considered the appropriate selection and training of personnel and the establishment of adequate logistics and a functional infrastructure. When these elements have been achieved and stakeholders are informed about the study team's readiness to proceed, the actual enrollment process can go ahead. This enrollment process is dependent on an informed consent process, the topic of the next chapter.

References

Cherrie J et al (1987) An experimental simulation of an early rock wool/slag wool production process. Ann Occup Hyg 31:583–593

Corn M (1985) Strategies of air sampling. Scand J Work Environ Health 11:173–180

Corn M, Esmen NA (1979) Workplace exposure zones for classification of employee exposures to physical and chemical agents. Am Ind Hyg Assoc J 40:47–57

Esmen N (1979) Retrospective industrial hygiene surveys. Am Ind Hyg Assoc J 40:58–65

National Institutes of Health Office for Extramural Research (2011) Protecting human research participants. http://phrp.nihtraining.com/users/login.php. Accessed Sept 2012

Wertheim HFL et al (2010) Laboratory capacity building in Asia for infectious disease research: experiences from the South East Asia Infectious Disease Clinical Research Network (SEAICRN). PLoS Med 7(4):e1000231

White E, Armstrong BK, Saracci R (2008) Principles of exposure measurement in epidemiology. Collecting, evaluating, and improving measures of disease risk factors, 2nd edn. Oxford University Press, Oxford, pp 1–428. ISBN 9780198509851

Managing the Informed Consent Process 16

Douladel Willie, Jan Van den Broeck,
Jonathan R. Brestoff, and Ingvild Fossgard Sandøy

> *Yes and No are very short words to say, but we should think for some length of time before saying them.*
>
> Anonymous

Abstract

Informed consent is the process of fully informing potential study subjects about the study and obtaining their voluntary agreement to participate or (if already enrolled in the study) to continue their participation. Informed consent is an ongoing process and a key responsibility of researchers using information or biological samples provided by human subjects. Ethics committees play vital roles in ensuring that necessary steps are taken to fulfill the ethical obligations linked to the informed consent process, but even with the best intentions, the informed consent process can be mismanaged. The chapter outlines the principles and the stages of the informed consent process, and it highlights issues to consider when managing and executing the informed consent process.

D. Willie, M.Sc. (✉)
Epidemiology Research Unit, and Department of Child Health and Psychiatry,
University of the West Indies, Mona, Jamaica
e-mail: depatri28@yahoo.com

J. Van den Broeck, M.D., Ph.D. • I.F. Sandøy, M.D., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no; Ingvild.Sandoy@cih.uib.no

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_16,

16.1 Informed Consent as a Key Responsibility in Human Subjects Research

Whenever a member of an investigative team approaches a candidate research participant, there is informational asymmetry that puts the candidate in a vulnerable position. The staff member knows the reasons for the study, the procedures that will take place should the potential subject enroll, and the risks and benefits associated with those procedures. The potential subject, on the other hand, may be assumed to have no prior knowledge of the study or its basis. Consequently, that individual cannot make a rational, un-coerced decision to participate without being informed about the study and its potential risks and benefits. The informed consent process is specifically intended to ensure that potential subjects are fully informed about the study so that they can make an autonomous, voluntary decision about whether or not to participate, a decision that must be made without any coercion. Indeed, it is a key responsibility of research staff to obtain informed voluntary consent (also referred to as informed consent) from all subjects prior to enrollment and to continually confirm informed consent as a study progresses.

The principles of informed voluntary consent are contained within the very meanings of the three words informed, voluntary, and consent (IVC):

- *Informed:* Full disclosure of pertinent study information is an absolute requirement
- *Voluntary:* The decision to participate should be made freely and without coercion or undue influence
- *Consent:* Agreement to participate must be given prior to study participation and should be clearly indicated orally or in writing

Terms and concepts relevant to the informed consent process are provided in Panel 16.1.

Panel 16.1 Selected Terms and Concepts Relevant to the Informed Consent Process

Assent Documented agreement to participation after thoughtful consideration by a subject not in a position to give informed consent legally

Coercion Intentional threat of harm, explicit or implicit, to obtain agreement

Community consent General approval from a community through its leader(s) that an activity may proceed in that community. This does not replace individual consent but may be a necessary prelude to obtaining individual consent[#]

Enrolment (1) (- procedure) Interactive process composed of sampling, eligibility screening and informed voluntary consent, intended to lead to official study participation (2) (- act) Official inclusion as participant

Harms Foreseen or unforeseen effects of a research procedure or of research participation experienced by the participant as burdensome or negatively affecting her/his health, wellbeing or (perceived) position in her/his community or the wider society

(continued)

Panel 16.1 (continued)

ICF *See:* Informed consent form

Impartial witness An independent person who attends the informed consent process, reads the informed consent form to the potential study participant and may co-sign the participant's statement

Informed consent form Written information in lay terms, read to, discussed with and given to, potential study subjects to enable them to voluntarily decide about study participation

Informed consent process Process of fully informing potential study subjects about the study and of obtaining their voluntary agreement to participate or to continue participation

Participant's statement Statement, signed by study participant or legal representative and usually co-signed by enroller and witness, that all study information was given and understood and that participation is voluntary

Research ethics Discipline providing ethical principles and guidelines for the design, conduct, analysis and dissemination of research involving human subjects

Therapeutic misconception Tendency of prospective or actual research participants to assume that participation in a research study will provide them with more health benefits than non-participation

Undue influence An offer, explicit or implicit, of an excessive, unwarranted or improper reward in order to obtain agreement for study participation

Vulnerable subject Potential research subject whose willingness to participate may be unduly influenced by expectations of benefits, fear of retaliation, or lack of capacity to engage in a comprehensive informed consent process

Yes-doctor syndrome A pattern of behavior often observed in the study enrollment process where potential research participants tend to relinquish their autonomy before an authoritative figure, such as a doctor

[#] Definition contributed by Dr. M. Chhagan.

16.1.1 Historical Glance at Informed Voluntary Consent

The 2009 edition of the International Ethical Guidelines for Epidemiological Studies, prepared by the Council for International Organizations of Medical Sciences in collaboration with the World Health Organization, defines Informed Voluntary Consent as:

> …a decision to participate in research, taken by a competent individual who has received the necessary information; who has adequately understood the information, has arrived at a decision without having been subjected to coercion, undue influence or inducement, or intimidation.

Textbox 16.1 The Tuskegee Study of Untreated Syphilis

In 1932, the United States Public Health Service and the Tuskegee Institute recruited 399 African American men with untreated latent syphilis and 201 healthy controls to a study on the natural course of untreated syphilis. The study was conducted in Tuskegee, Alabama, and has become infamous due to a number of ethical violations. Among the most problematic was the **failure to obtain informed voluntary consent** from the participants. The men were not informed about the real aims of the study and were told that they would receive free medical examinations and treatment for "bad blood," a non-specific colloquial term for many disorders or diseases such as syphilis, anemia, and fatigue. Those with syphilis were not told of their diagnoses or offered proper treatment, even after penicillin was discovered in the mid 1940s to be effective in treating syphilis. Instead, men were given placebo injections and minerals, thereby allowing the disease to progress beyond latency to tertiary disease and death. In addition, participants were not given an option to quit the study, an element of the study that also constitutes an **informed consent failure**. The study lasted for 40 years, and was not stopped until a newspaper article about the study caused public protests.

There was a time when IVC was not among the primary considerations in research involving human subjects. In fact, it was not until 1948 that the first international document addressing voluntary participation and informed consent was produced, in the form of the Nuremberg Code. The code, which states that "the voluntary consent of the human subject is absolutely essential" in experiments, was established following the Doctors' Trial in Nuremberg, Germany, 1946–1947. In this trial, German doctors and administrators were charged for conducting medical experiments (which often resulted in death or disability) on concentration camp prisoners without their consent. However, even though the Nuremberg Code emphasized the principle of IVC worldwide, it took time for research practices to change. For example, the Tuskegee Study of Untreated Syphilis (Textbox 16.1) employed intentionally misleading practices to obtain and retain participants (White 2000).

Since then, IVC has assumed great importance in epidemiology. This is evidenced by its inclusion in all the major human subject research guidelines, including the original (1964) and successive versions of the *Declaration of Helsinki* (World Medical Association 2010) and the *Belmont Report* (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research). The latter guideline came about as a result of public outcry against the aforementioned Tuskegee Study, leading to the National Research Act of 1974 in the United States. The main thrust of the Belmont report is the identification of some key ethical principles that should govern the practice and regulation of research involving human participants. One such principle – *respect for persons* – captures the essence of IVC and calls for the recognition of persons as autonomous agents.

16.2 The Informed Consent Process

The informed consent process may vary among studies with respect to duration and exact procedures. Possible reasons for variation include:

- The nature of the study (e.g., questionnaire only versus taking biopsies)
- Candidate research participants (e.g., literacy levels and cultural norms)
- Enrollment setting (e.g., crowded clinic versus private home)

Despite variations in modalities, a well-managed informed consent process should, at a minimum, have the following stages (also *See:* Fig. 16.1):

Stage-1: Presentation of complete study information

This information should be adapted to the capacities of the candidate research participant. The candidate should be made aware of (*See also:* Sect.16.3):

- The purpose of the study
- The reason (s)he is being invited to participate
- The nature and frequency of the procedures that will be undertaken
- The known risks and benefits associated with her/his participation
- The alternatives to participation
- Any compensation
- Conditions that might accompany participation
- How the collected information will be stored and used, including who will have access to the information
- Individuals who can be contacted for further study information
- General research advice

This constitutes a lot of information, so there is a clear need for persons conducting the informed consent process (the enrollers) to have an excellent knowledge and understanding of all aspects of the study and to have good communication skills. Training of enrollers is highly recommended, and often required, to achieve this.

Stage-2: Discussion

There should be many opportunities for the candidate research participant to ask questions. As with the previous step, the discussion should be held at a level that is comprehendible to the potential participant, and the enroller should be knowledgeable about the study proposal. The process should be conducted in strict privacy.

Stage-3: Assessment of understanding

Assessing the candidate participant's understanding of the study details is often based on perceptions during the discussion but can also be based on explicit questions or more formal assessments when this is acceptable.

Stage-4: Time to reflect and discuss with significant others

After the enroller has ensured that all the relevant information is adequately understood, the candidate participants must be allowed to make their own balance of the *pros* and *cons* of participation. They should be clearly advised that they are free to choose whether or not to participate and that once they participate they are free to quit at any time without any repercussions. It is preferable that the informed consent process not be conducted by a doctor, direct care-giver, or any other person who is

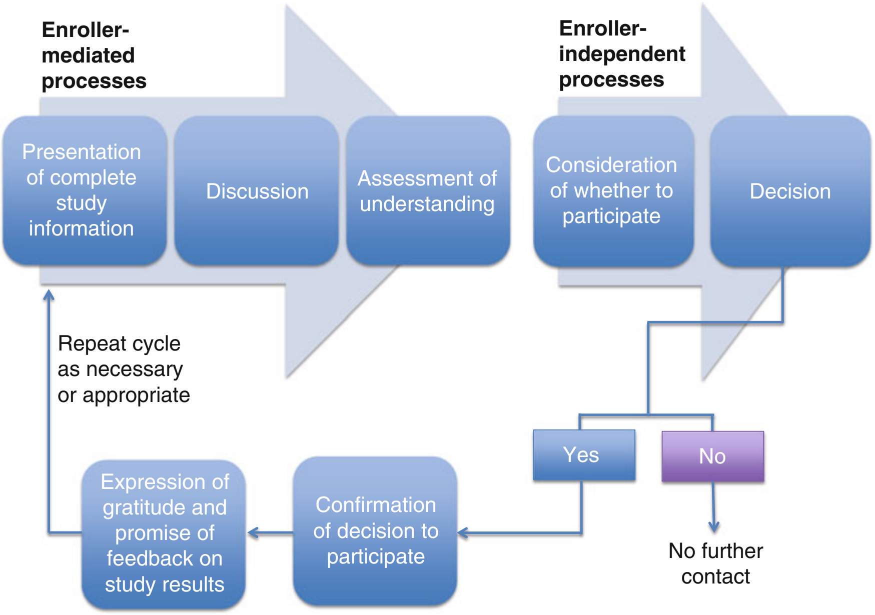

Enroller-mediated processes
Enroller-independent processes
Presentation of complete study information
Discussion
Assessment of understanding
Consideration of whether to participate
Decision
Repeat cycle as necessary or appropriate
Yes
No
No further contact
Expression of gratitude and promise of feedback on study results
Confirmation of decision to participate

Fig. 16.1 The steps of a typical informed consent process. The informed consent process usually begins with a series of enroller-mediated steps, the first of which is the presentation of complete study information. This explanation should be adapted to the capacities of candidate research participants and be appropriate for the study and setting. This presentation naturally leads to a discussion between the enroller and the candidate participant. The candidate participant should be given ample opportunities to ask questions and seek clarifications. During and after this discussion, the enroller will assess the candidate participant's understanding using multiple strategies (some validated, some creative). Based on the information conveyed to the candidate participant, that individual will contemplate whether or not to participate. This is the first enroller-independent step, and it usually involves self-reflection and discussions with friends and family. Ultimately, the participant makes a decision about whether or not to enroll in the study and communicates that decision to the enroller. If the decision is "No," then there should be no further research-related contact. If the decision is "Yes," then the enroller must confirm that decision by obtaining the necessary information and signatures on the informed consent form (or by other actions, e.g., by clicking "Agree to Participate" button in an online form). The enroller should express their gratitude and promise to provide feedback regarding study results. If necessary or appropriate, this entire process may need to be repeated on multiple occasions. Common reasons for repeating this process are follow-up studies, cases of suspected misunderstanding, changes in health status of the participant, and/or availability of new information

(or is perceived to be) at an unequal power level with the candidate. The doctor or investigator should, however, be available for clarification if necessary. Panel 16.2 lists factors that are commonly perceived as pros or cons of participation by candidate research participants.

Researchers are generally required to prepare an informed consent form containing all relevant study information (*See:* Sect. 16.3). It is important that the candidate participant reads this form carefully (with or without assistance, as the case warrants).

Panel 16.2 Factors Commonly Perceived as Pros or Cons of Participation by Candidate Research Participants

Pros:

- Making a contribution to society
- Liking of any foreseen benefits and incentives
- Social and entertainment aspects of participation
- Opportunity to please the doctor, superiors, or group

Cons:

- Costs of time, effort, mental-emotional commitment, money
- Possible harms to health, well-being, privacy, and confidentiality
- Fear that participation will bring displeasure to one's spouse, family, employer/colleagues, or friends

Enrollers should accept that the time required to do this may vary among individuals. As candidates read, discussions between the enroller and the candidate participant are encouraged and should be facilitated by the setting. After reading the informed consent form, ideally the candidate participant should be allowed as much time as they need to make a decision. Ideally, this can even include going home to discuss participation with family members and/or friends. In practice, however, the latter is not practiced in many studies, especially in surveys in which researchers do same-day data collection. Nevertheless, in clinical studies involving interventions, invasive biological sampling, or overnight hospital stays, allowing enough time for decision-making is of the utmost importance. Irrespective of the study details, candidate participants should never be made to feel as though they need to make a decision hurriedly.

Stage-5: Communication of the decision made

After all due considerations are made, the candidate participant will usually arrive at a decision and communicate it to the enroller. The enroller must accept this decision as final. In the case of a decision of non-participation, the enroller should not make further efforts to influence the decision, and all research-related contact should stop here. The remaining steps pertain only to a decision to participate.

Stage-6: Official confirmation of decision to participate

Confirmation of a decision to participate is normally accomplished by having the participant sign a statement. Confirmation of agreement to participate is considered 'giving consent.'

The signature of the individual or that of her/his legal guardian is generally required, though there are situations in which other indications of consent are acceptable. Thumbprints can be used in cases where persons are unable to sign. In some cases verbal consent is acceptable. In other cases completion of the questionnaire (for example a mailed questionnaire) is considered *implied consent* where a person's follow-up actions demonstrate agreement to participate in the absence of a signed document. However, questions are often raised regarding this approach since there is no easy way of assessing that such an implied consent is truly informed. In the case of minors, *assent* is also usually sought after receiving consent from the parent/guardian (*See:* Sect. 16.4.3).

If a study includes both interviews and collection of biological samples for tests that give results of a sensitive nature, such as testing for HIV and sexually transmitted infections, separate informed consent must be obtained from the participants for each test (*See:* UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance: Guidelines for measuring HIV Prevalence in population-based surveys 2005). If there is an intention to provide feedback on the results of such tests to each individual participant, it is also necessary to obtain permission for this action because some participants may not wish to receive such information. As part of the IVC process, it should also be made clear if some test results are mandatorily reported to government agencies.

Stage-7: Issuing renewed information and discussions during follow-up
The informed consent process extends beyond initial enrollment. This is usually apparent in studies that are ongoing for prolonged periods, involve follow up visits, and/or have lengthy data collection periods. In these cases new information sometimes becomes available, either directly from the particular study (through interim results) or from other non-study related activities in that particular area, which puts the study in a new light. It is important that this information is communicated to participants, as this is essential for ongoing consent to be truly informed. One practical way of providing new information is issuing timely study newsletters to participants and their families. Renewed consent may be required; consultation with the ethics advisory boards may help to determine whether renewed consent is required.

In the event that a participant decides to withdraw from a study, that decision should be fully respected. However, if the researcher is certain that the decision was made because of a misunderstanding it may be acceptable to attempt to clarify issues so that an informed decision is made.

In situations where the participant's health status changes over time, thus making the conditions and potential study effects different from those under which (s)he initially had agreed to participate, it may be necessary to provide such persons with additional information. Instances in which a person's understanding of study information fades or changes over time may also arise. It is generally expected that under these circumstances researchers should repeat the presentation of relevant information.

Stage-8: Expressions of thanks and feedback
This can be done after each study contact. Participants should be promised feedback on study results, and this promise should be honored.

16.3 The Informed Consent Form

As mentioned above, in addition to dialogue between the enroller and the candidate participant, information is usually presented in writing in the informed consent form.

16.3.1 Recommended Content of an Informed Consent Form

Panel 16.3 is a checklist of desirable elements of an informed consent form.

It is common practice to write the informed consent form in a conversational mode, as if the researcher were speaking directly to the candidate participant. Researchers should be careful to use terminology appropriate for the candidate participants, defaulting to simple language, avoiding technical terms, and including relevant explanations as far as possible and necessary. In studies with lengthy and complicated procedures, researchers often include a shortened version of the consent form. If this short form is used it is essential that it is presented along with, not instead of, the full version. The latter should be available for perusal.

Panel 16.3 Checklist for the Content of an Informed Consent Form

A well written informed consent form typically contains:

- An explanation that this is research; if necessary, explain what research is
- Aims of the research study
- Foreseen number of participants
- Reasons for inviting particular individuals to participate
- A description of the number of contacts, types of contacts, duration of the study, interventions, and procedures
- A description of what is experimental (if there is an experimental component)
- A description of foreseen benefits and risks to the participant and society, including minor discomforts
- A statement that there may be no direct health benefits; in trials this requires explaining that the test intervention may or may not turn out to be superior to the comparator intervention
- Reference to any compensation for travel and time costs and for possible injuries
- Description of how privacy and confidentiality will be maintained
- A statement indicating the voluntary aspect of participation and right to withdraw at any time
- Emphasis on the absence of any repercussions or changes in quality of health care in case of non-participation or withdrawal
- Information on who the sponsor and investigator of this study are
- Information on whom to contact with questions (e.g., sponsor, investigator, study physician, ethics committee)
- The participant's statement declaring that the presented information is understood and participation agreed
- Signatures of enroller, participant, (and sometimes impartial witness) and dates of all signatures

16.3.2 Consent for Storage and Future Use of Biological Specimens

The main thrust of bio-banking is to create a repository of biological specimens to be made available for research. Two types of bio-banks exist based on whether they are intended for research in a specific area of health or whether there is no specific area of future focus. The issue of consent for storage and future use of specimens has generated much debate within the scientific community and among other stakeholders (Van Diest and Savulescu 2002). This is because the norms of the IVC process are challenged and in many instances are impractical and even impossible to uphold. The informed consent process previously described in this chapter has full disclosure of study information and voluntariness as its tenets. The former is impossible if future uses are not defined when obtaining informed consent. Similarly, the ability to 'volunteer' or even opt out of specific sub-studies is also under threat.

Different positions have been arrived at through considerations of general ethical principles and empirical data on persons' preferences. One school of thought supports giving prospective consent, whereas another supports consent for each new study being done (Wendler 2008). In the former scenario, participants are asked either (1) to give blanket consent at the initial collection for their samples to be used for any study in *specific* area(s) of health/disease or (2) give blanket consent at the initial collection for their samples to be used for *any* area of health research. Variations of the above-described approaches also allow for decisions to be made regarding how the samples are stored, who uses them, whether data will be de-identified or anonymized, and what information (if any) will be provided to the donors for each of the studies for which their data are used.

The tremendous potential for research and development that lies in bio-banking cannot be ignored. Managing the informed consent process in this context requires selecting the position that ensures ethical treatment of research participants while serving research interests. Empirical data suggest that most persons are not averse to having their samples stored for future use (Wendler 2008). One assessment showed that the majority of persons are satisfied with giving initial consent for future use and then having an ethical review board make decisions regarding particular new studies (Clayton 2005). The distinction between research for academic versus commercial purposes, however, also needs to be considered as people have been shown to be slightly less likely to give samples for research for commercial gain. It may be best to present the potential research participant with several options regarding the use and storage of their samples and ways in which they can opt out of research after submitting samples. Researchers should consider empirical data available, cultural norms, and ethical principles as these options are being developed. Effective communication is a key factor in the actual process to ensure that prospective participants understand the options being offered and are aware of what they are getting involved in. This may even include awareness of the uncertainty related to what the samples may be used for in the future.

16.4 Challenges to Informed Voluntary Consent

16.4.1 The 'Yes-Doctor' Syndrome

The 'yes-doctor' or, more generally the 'yes-authority' syndrome, is a pattern of behavior often observed during a study where (potential) research participants tend to relinquish part of their autonomy before an authoritative figure, such as a doctor. Many people tend to decrease their role as an autonomous agent within an unequal relationship. Some examples of statements of participants revealing the 'yes-doctor' syndrome are:

- "If the doctor says it's okay for me to participate, then surely it's in my best interest to do so"
- "I am happy to let you, doctor, decide whether I should participate or not"
- "I don't want to disappoint the doctor because I have a good relationship with her/him"

What can be done to avoid the 'yes-doctor' syndrome?

- Enrollment, in our view, should be done by a person perceived to be the candidate participants' 'equal' whenever possible
- Never accept that people give up their autonomy if they have the capacity to decide for themselves
- Where appropriate, potential participants should be encouraged to carefully consider any burdens and risks as well as any other pros apart from the chance to please a doctor they like

16.4.2 Therapeutic Misconceptions

In clinical trials misunderstandings are common among (potential) participants and health care personnel about the purposes of the study and about the concepts of equipoise and randomization (Flory et al. 2008). Many fail to fully understand that none of the intervention arms may be superior to the others and that they might well be allocated to the comparison arm rather than the test intervention. There is a general trend towards optimism that participation will bring extra health benefits that are beyond what a non-participant could expect from routine treatment (Flory et al. 2008). Such expectations can certainly be caused by an inappropriate or unsuccessful informed consent process and can, in that case, be called a 'therapeutic misconception.' However, the expectation of extra health benefits is often present among persons who understand equipoise and randomization and who are fully informed about risks and potential benefits. In fact, expecting some extra health benefits is often not unrealistic as patients/clients in many trials often receive more frequent and higher quality medical attention than they would get in the routine health care system. This is often the case with trials in resource-constrained settings where issues concerning health care accessibility exist. In low- and middle-income countries especially, it is therefore not unusual that trial participation is (quite realistically) perceived as bringing extra health benefits and even as offering an increase in the quality of life. For many potential participants, these are all good reasons to have high expectations.

A brief digression on ‘post-trial obligations’ and therapeutic misconceptions is appropriate here. Some ethical guidelines prescribe that, after a trial, the best treatment should be given to all participants. This type of requirement remains somewhat controversial among epidemiologists. Results of a single trial can rarely be used in isolation to create firm knowledge or to establish a new care practice. In reality one cannot normally equate a positive observed effect in a single trial with general knowledge about safety, efficacy, and effectiveness of the treatment. More than one trial is normally required to provide a proper basis for establishing a new standard of care. Thus, promising continued treatment may be realistic, but promising the *best* treatment may be unrealistic and can induce a therapeutic misconception.

Based on the above considerations, a therapeutic misconception can be defined as an unrealistic level of optimism about health benefits of trial participation. Of special interest is the degree to which this level of optimism is induced by failures during the informed consent process prior to or after enrollment.

What can be done to avoid therapeutic misconceptions?

- The uncertainty related to potential *direct* health benefits of study participation should be emphasized in the informed consent process
- During the informed consent process the concepts of equipoise and randomization should be conveyed
- Concerning post-trial provisions, one should take care not to make promises that are unrealistic
- In resource-poor settings, some strengthening of the routine care system via the research project may be envisaged, discussed, and carried out when deemed acceptable by the local health authorities

16.4.3 Vulnerable Persons

In research, vulnerable persons are those who for some specific reason(s) have diminished capacity to protect their own interests. These reasons can include, but are not limited to, legal restrictions against decision making; diminished mental capacity due to age or impairment; and reduced social and/or financial standing. Special justification will be required to invite vulnerable persons to participate in research, and the *CIOMS Guidelines* require that safeguards are employed to protect their rights and welfare. These may involve, among others, further research regarding their condition, limits to the amount of risk they may be exposed to, employment of consent monitors, and proxy consent provided by caregivers using a ‘best interests standard’ (For more information on these approaches, *See:* CIOMS 2009, 2010).

Research among persons with vulnerability raises ethical concerns because of the existence of a real possibility for exploitation or unintentional maltreatment of these persons. Researchers have the responsibility to ensure that their actions are in the best interest of these persons and in accordance with existing ethical guidelines and relevant legal requirements.

16.4.3.1 Children

Children should not be enrolled as research participants unless it would otherwise be impossible to answer a research question relevant to their health and well-being. Moreover, The Declaration of Helsinki, Article 25, states: "When a subject deemed legally incompetent, such as a minor child, is able to give assent to decisions about participation in research, the investigator must obtain that assent in addition to the consent of the legally authorized representative." If the legally authorized representative consents to the child's participation, the child should also be informed about the study (describing the rationale and procedures to the extent that (s)he can understand) and asked to sign an assent form or to give verbal assent in the presence of the legal representative and a witness. It should be made clear to the child that (s)he should honestly indicate her/his willingness to participate and that (s)he is free to decline even if the parents or guardians consented to her/his participation. In such a case the researcher should yield to the wishes of the child. To determine if it is necessary or appropriate to receive assent, the researcher will need to be fully conversant with the legal requirements of the particular locale regarding the age of consent in research. The researcher will also need to assess objectively the child's ability to give assent.

16.4.3.2 Persons in Dependent Positions

Persons in dependent positions include college students, employees, and prisoners. When seeking to involve these persons in research, researchers need to take care that participation is informed and voluntary and not due to fear of adverse repercussions of non-participation, such as loss of a job and/or privileges. The converse is also important: participation should not be due to expectations of unwarranted benefits.

16.4.3.3 Patients with Cognitive Impairment

Similar procedures as those utilized in research involving children should be followed in research involving persons with cognitive impairment. It is also advised that research should not involve these persons unless it would otherwise be impossible to answer research questions that are essential to their health and well-being.

16.4.4 Cultural Challenges and Vulnerable Communities

General guidelines and best practices for the informed consent process have been described above. It is important to note that situations may arise in which these guidelines and practices have to be adjusted to suit the cultural context or setting in which the research is (to be) undertaken. This is a very important consideration that should be duly acknowledged, as study participants must deem research acceptable.

16.4.4.1 Community Consent Versus Individual Consent

The importance of treating the individual as an autonomous agent was emphasized earlier. Notwithstanding, researchers have come to recognize that in many settings the community (which may be delineated by geographical, social, cultural or religious boundaries) has pivotal roles in influencing and determining an individual's decision. These situations can be difficult to navigate, as the researcher is called upon to balance usual research standards and practices with community acceptability. The researcher will need to be creative and prudent in these negotiations but also needs to ensure that persons are participating because they have a personal desire to do so, not because community leaders have told them to. The specific procedures by which one should obtain community consent vary among different communities, and there may be important local variations within a country. It is therefore useful to gather information from candidate study communities regarding the procedures 'outsiders' are expected to abide by to gain the trust and consent of the community to recruit and enroll participants. Failure to abide by such required procedures may result in the whole community refusing to participate.

16.4.4.2 Internet-Based Surveys and Informed Consent

Internet-based surveys are simple, efficient instruments that are easy to create, distribute, and administer. Data management and analysis is also usually highly efficient in Internet-based surveys. It is not surprising therefore that many researchers are drawn to this method. Internet-based surveys often include information about the study along with an invitation to participate and a means of declaring agreement to participate. It is readily noted, however, that it is difficult or impossible to conduct many of the recommended activities for gaining IVC for such surveys. For example, researchers will not be easily able to assess whether the consent was truly informed since it is difficult to guarantee that the respondent read or understood the study information (though questions about the study itself can precede the survey questions as an additional safeguard). Many of these surveys rely on implied consent, and even if a participant statement is signed, there is no guarantee that the person who signs is the one who fills out the questionnaire, or that either of those persons is the intended respondent. There are also fewer assurances that can be given regarding confidentiality and privacy, as IP addresses can be traced (removing anonymity) and computers where data are stored can be hacked (though it can be argued that offices, filing cabinets, and offline databases can also be compromised). The benefits of internet-based surveys cannot be downplayed, so the guidelines for informed consent will need to 'catch up' to the technology and advise accordingly. Notwithstanding, the 'spirit' or essence of the IVC process is still applicable and the onus is on the researcher to ensure that research activities are still in keeping with these ideals.

16.4.5 Manipulative Temptations of Researchers Regarding Informed Consent

Investigators, enrollers, and data collectors all have direct responsibilities in the informed consent process. They also have professional and career interests and often have time constraints and undergo peer pressure. Under such circumstances they may succumb to a temptation to shorten the time taken to inform and discuss with candidate participants, or they may be brief in their answers to questions from participants, forget to mention some minor but relevant aspects of risks, delay a planned information session on study progress and interim findings, or postpone a meeting with a community leader. Subtle signs of the yes-doctor syndrome may be disregarded or even welcomed as a way to increase study efficiency and gain time for other activities. These examples concern relatively mild but regrettable deviations from an optimal informed consent process. More serious deviations also occur that clearly fall into the category of violation of duties and that create liability to prosecution in most countries. These include failure to provide relevant information, enrollment without consent, coercion, and failure to document consent, among others.

16.5 Informed Consent, Scientific Validity, and Study Efficiency

In this chapter the essence of the informed consent process has been described as an expert-guided information exchange and as a process of (re-)enabling participants to make a personal choice and follow it through. To enable free choice, the candidates are informed about study purposes, risks, and benefits and about the nature of the activities required. They need to be well aware that their contribution is significant to the study only if they are able to provide accurate information and adhere to scheduled study activities/interventions. The link between the informed consent process and scientific validity is therefore quite direct; consent to participate should be associated with a commitment to contribute to scientific validity.

There are also negative links between informed consent, scientific validity, and study efficiency. Misinformation or misunderstandings about study purposes and procedures may ultimately lead to disappointments about the actual burden of participation and be detrimental to the relationship with study personnel. Two types of reactions are common in such situations: manifest refusal or ‘hidden refusal’ with poor adherence and misreporting. Both may also be attributed to factors other than a poorly managed informed consent process. Frequent refusals, widespread inaccurate reporting, extensive loss to follow-up, and poor adherence are signs of the existence of strong factors that tend to swing people’s perception of the balance of pros and cons of study participation to the negative side.

In this chapter we discussed the informed consent process, which enables persons to make decisions about participation in the study. Accrual rates in a study strongly depend on both the content and the style of communication, and so do rates of retention and levels of adherence to prescribed study procedures. This brings us to the topic of Chap. 17: Accrual, Retention, and Adherence.

References

Clayton EW (2005) Informed consent and biobanks. J Law Med Ethics 33:15–21

Council for International Organizations of Medical Sciences (2009) International ethical guidelines for epidemiological studies. CIOMS, Geneva, pp 1–128. ISBN 929036081X

Council for International Organizations of Medical Sciences (2010) International ethical guidelines for biomedical research involving human subjects. CIOMS, Geneva. http://www.cioms.ch. Accessed Sept 2012

Flory JH, Wendler D, Emanuel EJ (2008) Empirical issues in informed consent for research. In: Emanuel EJ et al (eds) The Oxford textbook of clinical research ethics. Oxford University Press, Oxford, pp 645–660. ISBN 9780195168655

UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance (2005) Guidelines for measuring national HIV prevalence in population-based surveys. WHO, Geneva, pp 1–67. ISBN 9241593709

Van Diest PJ, Savulescu J (2002) No consent should be needed for using leftover body material for scientific purposes: for and against. BMJ 325:648–651

Wendler D (2008) Research with biological samples. In: Emanuel EJ et al (eds) The Oxford textbook of clinical research ethics. Oxford University Press, Oxford, pp 290–297. ISBN 9780195168655

White RM (2000) Unraveling the Tuskegee study of untreated syphilis. Arch Intern Med 160:585–598

World Medical Association (2010) The declaration of Helsinki. http://www.wma.net/en/10home/index.html. Accessed Sept 2012

Accrual, Retention, and Adherence 17

Jan Van den Broeck and Thorkild Tylleskär

Every man is guilty of all the good he did not do.

Voltaire

Abstract

This chapter offers practical advice on how to monitor and optimize enrollment rates, retention rates, and adherence levels in a study. The advice is partly based on a framework of thinking about dynamics of research participation. When problems arise with levels of research participation, two dynamics are worth considering to gain insight into causes of the problems and to start designing responses. The first is that potential and enrolled research participants continuously entertain fundamental concerns about personal safety, personal gain, social acceptability, and personal competing interests. The second is that these concerns are continuously influenced by changing internal and external factors, the latter of which includes practical circumstances and opinions of various stakeholders or opponents of participation with the research study.

17.1 Dynamics of Research Participation

Relatively few studies have formally compared strategies for optimizing recruitment, enrollment, retention, or adherence. However, a wealth of useful information has come from experiential knowledge shared by epidemiologists and from individual, mostly non-methods-oriented studies (e.g., Lovato et al. 1997; Swanson and Ward 1995; Ross et al. 1999; White et al. 2008). Indeed, the anecdotal evidence has

J. Van den Broeck, M.D., Ph.D. (✉) • T. Tylleskär, M.D., Ph.D., MA
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_17,

Panel 17.1 Selected Terms and Concepts Around Accrual, Retention, and Adherence

Accrual The gradual completion of the foreseen number of enrolled participants

Adherence to treatment The correct following by the participant of all instructions concerning the allocated treatment regimen

Dropout rate Rate of early unplanned cessation of individual follow-up among study participants

Enrollment rate Rate at which new participants are officially included in the study

Loss to follow-up Dropout from study participation not due to death

Participation rate Rate of participation among a group of persons one would have liked to participate completely

Refusal rate Rate of non-participation among eligible observation units invited to participate

Retention Avoidance of loss to follow-up

been the most informative on this topic, as many authors have reported their experiences dealing with recruitment, enrollment, and retention problems and their successes with changes in strategy (for terminology *See:* Panel 17.1). From these sources some basic insights into the dynamics of research participation can be gained. Strategies for optimizing accrual, retention, and adherence need to be based on a minimum understanding of these dynamics of research participation, of which two are particularly helpful to understand well.

1. It is important to keep in mind that potential and enrolled participants have variable levels of motivation to participate. They can perceive benefits of participation of a financial, social, health-related, or other nature. However, before or after enrollment, there are many factors that can offset the perceived benefits at any time. These so-called ‘negative factors’ include misinformation, fears, dislikes of particular study-related persons/institutions, social desirability concerns, practical feasibility concerns, and real burdens. External factors (e.g., unforeseen financial hardship) and internal factors (e.g., health-related experiences that may occur during participation) tend to influence the person’s balancing of the pros and cons of participation. When participation levels are a problem, the challenge for the researchers is therefore to (a) find out whether there is a problem with any of the aforementioned negative factors, and (b) to find appropriate ways to resolve those problems.
2. Problems with research participation and strategies to resolve them are highly situational. However, in general, it is evident that problems with enrollment rates, retention rates, and adherence levels (and solutions to those problems) are associated with one or more broad levels of influence. Table 17.1 lists several levels of influence and illustrates how each can support or impair enrollment,

Table 17.1 Levels of influence on individual research participation

Level of influence	Examples of influence on		
	Enrollment	Continued participation	Adherence
Family, friends	Negative advice from family member	A sick family member requires increased care effort	Husband helps with taking study medication correctly
Local community	Positive advice from community leaders	A negative rumor about the study is circulating	Adherence is promoted during community meetings
Co-participants	Enthusiasm shown by those already enrolled	Within 2 weeks two study children died	One participant convinces others that there is no strict need to adhere
Circumstances	Intention to emigrate during study period	The participant found a job and is now extremely busy	Bus fares have gone up recently, making it more difficult to attend study visits
Study personnel	Enrollers show empathy and respect	The newly hired interviewer makes many mistakes, which is noted by participants	Time is taken to repeat explanations of procedures at each study visit
Investigators and research institutions	Good reputation, perceived competence	Institution received positive media attention	Due attention is paid by the investigator to monitoring adherence
Study characteristics	Adequate provisions are in place to preserve confidentiality	A study protocol amendment increases the burden for participants	The schedule for taking study medication is complex
Personal health	The person easily gets tired when traveling to the health center	Death	Participant tends to become nauseous when taking study medication

retention, and adherence. In examining Table 17.1, it should be kept in mind that the various levels of influence are linked in complex ways. For example, study personnel can be members of the local community, so their practices and attitudes are influenced by prevailing perceptions in the community, local opinion leaders, and local authorities. Generally speaking, what happens on one level of influence also affects what happens on other levels; therefore, each influence has both direct and indirect effects on an individual's practices and attitudes with respect to study participation.

17.2 Monitoring and Optimizing Enrollment Rates

Once enrollment has started, accrual needs close monitoring to be successful (Lovato et al. 1997). The accrual trend can be shown as monthly or weekly numbers with a smoothed trend line. Ideally, accrual rate should be about constant to minimize uneven or excessive workloads (Schoenberger 1987; Lovato et al. 1997).

Based on extrapolation of the trend line, one can forecast the number and proportion enrolled by the end of the official enrollment period, perhaps with allowance for some expected future changes in enrollment capacity. The forecast can serve as a basis to accelerate or decelerate enrollment. The forecast may also be used for reports to the Data Monitoring Committee (DMC), also known as Data and Safety Monitoring Board (DSMB), in clinical trials. In prospective follow-up studies it may be used in the context of periodic renewal of ethics approval.

In addition to total enrollment numbers and forecasts, the quantification of response rates and their relation to exposure categories is a common concern. The monitoring process may reveal a worrying differential rate among important exposure categories. The concepts of 'response rate' and 'participation rate' have been unclear and variably defined in the epidemiologic literature because the nature of the denominator population varies among studies who report these rates (Galea and Tracy 2007). Investigators variably use as the denominator: persons sampled; sampled and approached; sampled, approached, and screened; or, sampled, approached, screened, and eligible. When reporting a 'response rate' or 'participation rate,' the exact nature of the denominator should be described and justified.

17.2.1 Reasons for Poor Enrollment Rates

Enrollment rates are often slower than expected (Lovato et al. 1997) or they can initially be as expected only to taper gradually later in the study. When that happens, a careful situation assessment is warranted to identify possible reasons and to develop for a suitable remedy. First, one examines at what stage of the process the problems appear. This point is referred to as an *enrollment bottleneck*. Examples of common reasons for enrollment bottlenecks are listed in Panel 17.2. For each bottleneck, one may want to verify if any of the types of sources of failure listed previously in Table 17.1 could be in play.

Panel 17.2 Examples of Enrollment Bottle Necks (Listed in Succession)

- The pool of eligible subjects is smaller than expected
- The recruitment strategy fails to reach a sufficient proportion of eligible subjects for the first contact
- The recruitment strategy fails to convince contacted subjects to agree with eligibility screening
- The recruitment strategy fails to initiate informed consent among a proportion of subjects screened eligible
- The informed consent process fails to convince potential participants
- Some people agree to participate but then disappear or are found not to be eligible after all ('early exclusions')

17.2.1.1 Main Motivations for Refusal by Subjects

Reasons for refusal may vary from study to study and from individual to individual. Reported reasons may be difficult to interpret when they are obtained by interviewing (especially standardized interviewing) as subjects may not wish to reveal their true motivation or the complexity of it. A focus group discussion in the preparatory phase of a study may give insight in prevailing perceptions and likely frequency of refusals.

Among reported reasons for refusal, one often finds one or several of the following:

- Fear for physical side effects, especially if reported by persons already enrolled
- Time and money concerns
- Planned out-migration
- Fear of stigma; fear of being asked sensitive questions
- Fear of legal consequences
- Fear of disapproval by friends or family members
- Negative rumors about the study
- Unfavorable opinions expressed by opinion leaders or local authorities; lack of desired community consent
- Lack of perceived potential benefit
- Dislike of personality of enroller or dislike of investigator or research institution
- Being unhappy with clarifications given by enroller

We mentioned in Chap. 9 that cover letters, information sheets, informed consent forms, personal introductions by enrollers, and other recruitment strategies need to be culturally adapted and must show respect, politeness, and goodwill. They must also give reassurance about participants' concerns (where possible to do so) and show professional seriousness. If they do not, enrollment rates are bound to be lower.

Motivations and associated refusal rates may differ among important study subgroups. It is well known that in case–control studies, cases tend to be more motivated to participate than controls (White et al. 2008). In studies of the effect of a potentially harmful exposure, refusal rates tend to be lower among those who are aware of their own suspicious exposure level. Studies in Anglo-Saxon high-income countries have revealed that, in such settings, many other personal characteristics can influence response rates and be a source of selection bias (White et al. 2008).

It is useful to make a distinction between *soft and hard refusals*. One may assume that many candidate participants are unclear about where their own balance of pros and cons actually is. Many are hesitant to make a decision and they will try to defer it when they can. Thus, non-response to a mailed questionnaire or to a reminder does not always imply hard refusal to participate. When forced to decide quickly, a 'soft' but explicit refusal is what will usually follow, which could be seen as a failure of the informed consent process.

17.2.2 Strategies to Improve Enrollment Rates

After locating the enrollment bottlenecks (Panel 17.2) and assessing main reasons for refusal (previous section), the following response strategies can be weighed and considered according to scientific, ethical, cultural, and efficiency criteria (Panel 17.3).

Panel 17.3 Possible Strategies to Deal with Low Enrollment Rates

- Prolong the official recruitment and enrollment period
- Decrease sample size; settle for less precise estimates or less statistical power
- Relax eligibility criteria (this is only valid if done after the slow enrollment rate was discovered during a pilot study)
- Increase stakeholder support and improve image; advertise and inform more frequently; solicit support from opinion leaders and community leaders; combat misunderstandings and unfounded negative rumors about the study
- Increase the study recruitment area; go multi-center; go international
- Increase recruitment capacity by increasing the number of recruiters and enrollers; increase budget for recruitment; provide better training for recruiters and enrollers; reduce number of relatively unsuccessful enrollers
- Make recruitment and enrollment procedurally more efficient, better culturally adapted, more respectful and polite, more informative on sensitive issues
- Improve the quality of the informed consent process; give undecided candidates more time; approach them again later
- Make study procedures less burdensome if the perceived burden seems a major reason for refusal
- Offer re-imbursement of travel and time costs; provide for meals and drinks while waiting; provide small tokens of goodwill
- Provisionally recruit people who are not yet eligible but are expected to become eligible before the end of the recruitment and enrollment period
- Repeat contact with more experienced recruiter some time after a soft refusal to a less experienced recruiter

Whatever the strategy used to boost enrollment rates, time is of the essence. Deployment of the new strategy needs to be well considered and well planned but fast. Lovato et al. (1997) pointed out that the inability to alter existing plans rapidly and to implement other strategies is a common reason for enrollment rate and study size problems. They also suggested that clear lines of authority and clarity of responsibilities might help to avoid such delays. Finally, one must realize that the implementation of these new strategies involves obtaining ethics approval for a protocol amendment and approval of the various stakeholders as relevant to the particular study. These approvals can also be time-consuming.

Caldwell et al. (2010) reviewed studies that compared strategies for increasing recruitment into trials. Formal comparisons of survey methods also exist (e.g., Edwards et al. 2009) but most of those are studies of middle-class U.S., hardly relevant to other populations.

17.2.3 Faster than Expected Enrollment Rates

Enrollment rates may turn out to be faster than expected. This may seem advantageous in terms of time investment, cost reduction, avoidance of stakeholder relationship problems, or avoidance of newly emerging confounders or study fatigue. And often it is. There are some caveats, however. First, increased work volumes per unit of time can saturate a system and ultimately decrease its quality. This point is relevant to all study types, but it is especially critical for follow-up studies because a shorter enrollment period implies that individual follow-ups will also be concentrated in a shorter total calendar period. Consequently, high work volumes and potential decreases in quality can be recapitulated at each follow-up. Moreover, contractual obligations to staff and for some equipment usually cannot be amended simply because the study enrolled its subjects faster than expected. Finally, faster than expected enrollment rates can be problematic if the study aims for equal rates of enrollment over a year to avoid seasonality effects.

17.2.4 Access to Secondary Data

In studies that use secondary data there may or may not be a need to ask each individual for consent to use the data. For example, the ethics committee may decide that there is no need for new consent to use historical hospital data and patient files. When there is a need, then some of the recruitment and enrollment strategies described above may be useful. These strategies may be greatly supported if the persons or institutions previously involved in the primary data collection have a role in the current recruitment effort. In any case a proposed useful sample size or a foreseen rate of extraction of data may turn out to be unachievable. Occasionally extra data collection is then envisaged.

17.3 Monitoring and Optimizing Retention Rates

Poor retention can be a source of bias and imprecision of study findings (*See:* Chap. 1). Good retention is an important challenge in all studies with new data collection, not only in prospective follow-up studies. In cross-sectional studies, for example, participants may decide to discontinue an ongoing measurement session. Monitoring rates of loss to follow-up (i.e., lack of retention) in an ongoing study is not always a straightforward exercise. For example, when multiple study contacts are scheduled, a participant may not attend on several successive occasions and seem to have been lost to follow-up, only to unexpectedly reappear at a later scheduled contact. A similar situation arises when out-migrations are followed by a return to the study area.

17.3.1 Reasons for Poor Retention

Potential factors contributing to retention are listed in Table 17.1. When investigating reasons for poor retention it is often helpful to keep in mind that participants

Panel 17.4 Questions Study Participants Continuously Ask Themselves During a Study

- Are these researchers still as trustworthy and competent as I thought they would be?
- Is the issue that they are doing research about and the research itself still relevant to me and my community?
- Are the research staff members still communicating with me in a respectful manner?
- Does what I am getting out of this study correspond to promises made or established expectations? Is it as exciting as I thought it would be?
- Can I deal with the burdens of participating? Is it still easy enough for me to continue participation?
- Is it still safe to participate? Is it as safe as I was told?
- Do other people still think this project is okay?
- How many people have already participated? Are others also continuing their participation?

frequently express their concerns. In Panel 9.3 we listed frequent concerns that people have about participation in a research study and ways of dealing with these concerns from the earliest recruitment phase of a study. These are the same questions that participants will continue to seek answers for during measurement sessions and during their individual follow-up period (and beyond). The reason is that these questions are rooted in permanent fundamental concerns about personal safety, personal gain, social acceptability, and personal competing interests. Translated into the follow-up situation, these recurrent questions become those listed in Panel 17.4.

17.3.2 Strategies to Improve Retention

When considering methods of improving retention it is helpful to take the above-mentioned list of concerns into account. Some of the questions may have straightforward answers, but most do not. Providing participants with more updates on study progress, study safety aspects, and on what is known about the topic in general are obvious choices to improve retention, as is keeping the study in the news. Some other methods of maximizing retention in prospective longitudinal studies have been reviewed by Hunt and White (1998) and discussed in White et al. (2008). In such studies the aim should be to create a group sense of identification with the study. Bonding strategies should be employed and may consist of:

- Frequent contact, preferably with the same person if there is a good relationship, but not so frequent as to annoy people

- Empathy, enthusiasm, and commitment from study personnel
- Respect for personal needs and preferences of participants
- Newsletters
- Social gatherings of participants
- Small gifts and/or cards

In prospective longitudinal studies, one should be alert to signs of study fatigue and respond to misunderstandings. Telling signs frequently precede loss-to-follow-up. Bad rumors about the study can circulate among participants or in the wider community and this may announce imminent retention problems. Urgent action is then required. A re-activation of the informed consent process may help, but one should always keep in mind that participants are free to discontinue participation at any time. Sometimes a loss can be prevented by switching to another data collection mode, one that is better perceived by the participant. Sometimes a loss can be prevented by linking a new data collector to the subject or by a chat with a study supervisor, coordinator, or investigator.

Contact details should be available on all participants and one close contact person if that second person agrees. These contact details should be updated regularly, and if this fails one should trace losses to follow-up using a variety of inventive strategies within legal boundaries. In the course of a longitudinal study, the data collector gets to know the participants better and may even acquire a picture of their social network, e.g., memberships or other participants who have become friends. This may help to trace some of those lost to follow-up.

Hint

Out-migration from the study area does not always need to be a reason for dropout. It may be possible to maintain follow-up of some variables (e.g., vital status) over a distance via mailings, phone calls, and the Internet. Some resourceful studies have even resorted to long-distance traveling in order to do follow-up measurements on emigrated subjects.

17.4 Monitoring and Optimizing Participants' Adherence Levels

The measurement and monitoring of individual participants' adherence to interventions has been discussed in Chap. 10 (The Measurement Plan). In trials it is useful to monitor the distribution and trends of levels of adherence in the study sample and its relevant subgroups. Such monitoring may be a requirement of the DMC/DSMB of a trial. Similar to what we discussed for accrual and retention rates, here too, a worsening trend in adherence levels may trigger the need to assess reasons and implement remedial strategies.

17.4.1 Reasons for Poor Adherence by Participants

In every study all possible sources of lack of adherence need to be foreseen and prevented as much as possible. Failure to do so can on and by itself be a cause of suboptimal adherence levels in the study. There are many reasons for poor adherence, including the following:

- The informed consent process was inadequate
- Bad study management
- Poor forecasting of the time and volume of new orders of products
- Poor communication with pharmacists and suppliers
- The intervention regimen is complicated and contains several substances
- Adherence to the intervention regimen is very strict, e.g., in some HIV drug trials, the drug regimen does not allow omissions.
- Calamities
- The study intervention has undesirable side-effects; for pediatric trials, this also includes poor palatability
- The participants may forget details of the instructions given to them
- The participants forgot appointments made far in advance, which can lead to interruption of the intervention
- The participant intentionally manipulates the intervention to make sure they (or their children) receive the test intervention, e.g., a mother might mix the tablets received for their children in order to make sure that every child got at least part of the test drug (this was documented for an early AZT antiretroviral treatment trial)
- The participant lacks a high level of discipline
- The participant lacks adequate social support
- A participant has not disclosed the study intervention to people with whom (s)he is living
- The participant is travelling away from home

17.4.2 Strategies to Improve Participants' Adherence Levels

Remedial strategies may be individually tailored to the particular situation of each participant, or they can have a more general approach. The latter is needed when structural and logistical problems are found to be a reason for generally suboptimal adherence rates. Reminders shortly before an appointment can help to avoid missed contacts for handing over new drug supplies and discussing and treating side effects. It is important that, at each contact, adherence is discussed seriously and that the participant is well aware that her or his adherence levels are monitored.

A common strategy in HIV treatment trials is that the participant identifies a 'buddy,' a friend or a family member who encourages the participant. The buddy may also receive some information and training from the study team in order to carry out the support function properly. Discussing in advance the situations that may be extra challenging and have the participants think through such situations can improve adherence rates.

In this chapter we discussed dynamics of study participation and ways in which the investigator can monitor and manage study participation issues. The success of this aspect of study implementation is crucial for obtaining high quality data. Another crucial aspect in this regard is the way in which questionnaires are designed and administered. That is the topic of the next chapter.

References

Caldwell PHY et al (2010) Strategies for increasing recruitment to randomized controlled trials: systematic review. PLoS Med 7(11):e1000368

Edwards PJ et al (2009) Methods to increase response rates to postal questionnaires. Cochrane Database Syst Rev 2:MR000008

Galea S, Tracy M (2007) Participation rates in epidemiologic studies. Ann Epidemiol 17:643–653

Hunt JR, White E (1998) Retaining and tracking cohort study members. Epidemiol Rev 20:57–70

Lovato LC et al (1997) Recruitment for controlled clinical trials: literature summary and annotated bibliography. Contr Clin Trials 18:328–357

Ross S et al (1999) Barriers to participation in randomized controlled trials: a systematic review. J Clin Epidemiol 52:1143–1156

Schoenberger JA (1987) Recruitment experience in the aspirin myocardial infarction study (AMIS). Contr Clin Trials 8:74S–78S

Swanson GM, Ward AJ (1995) Recruiting minorities into clinical trials: toward a participant-friendly system. J Natl Cancer Inst 87:1747–1759

White E, Armstrong BK, Saracci R (2008) Principles of exposure measurement in epidemiology: collecting, evaluating, and improving measures of disease risk factors, 2nd edn. Oxford University Press, Oxford, pp 1–428. ISBN 9780198509851

Questionnaires 18

Jan Van den Broeck, Meera Chhagan, and Shuaib Kauchali

Many recall questions would never be asked if researchers first tried to answer them themselves.

N.Schwartz and D.Oyserman

Abstract

A questionnaire is a measurement tool consisting of a list of questions accompanied with instructions, response options, and answering spaces. It guides the respondent and sometimes also an interviewer in finding and recording measurement information. As such, the questionnaire is a source document: it is very close to the source of data, the respondent. Errors at this point tend to considerably and sometimes irreversibly affect the validity of the evidence generated in the study. This chapter first deals with the response process, as a good understanding of the psychological stages of the response process can help questionnaire designers and interviewers to avoid recall error and misreporting. Second, this chapter provides practical recommendations for questionnaire design and administration. Study objectives, types of measurements planned, error-avoidance concerns (including prevention of errors in later data processing and analysis), and ethical concerns guide questionnaire design. Panel 18.1 introduces terminology used in this chapter.

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

M. Chhagan, Ph.D., FCPaed • S. Kauchali, M.Phil., FCPaed
Department of Paediatrics, University of KwaZulu-Natal, Durban, South Africa
e-mail: Chhagan@ukzn.ac.za; kauchalis@ukzn.ac.za

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_18,

18.1 The Response Process

There are several psychological stages of the response process during questionnaire administration (Tourangeau et al. 2000). A common view distinguishes five stages (Schwarz and Oyserman 2001; Streiner and Norman 2008), as shown in Fig. 18.1.

Knowledge of these stages is helpful in evaluating the usefulness of potential questions and in minimizing recall errors and misreporting. We discuss each stage in some detail and highlight sources of bias constituted by deficiencies at each stage.

18.1.1 Response Process, Stage-1: Understanding the Question

The process starts with the respondent reading or hearing the question and attempting to understand what information is being requested. Culture, language, and individual interpretations influence this understanding. Understanding of the question may also be influenced by 'context effects,' i.e., by information that appears on the questionnaire (e.g. previous questions) or by any suggestion that the researcher or the research is interested in particular types of behaviors or other characteristics. The way the question is formulated is crucial, but, in addition to the question itself, it is often the list of response options that clarifies to a respondent what the question actually means or leads them to assume a certain meaning of it (Schwarz and Oyserman 2001). Errors arising at this stage are called 'comprehension errors,' meaning that the respondent does not understand the question or understands it in a way unintended by the researcher. An example is an item on a questionnaire that was designed outside Africa and used in an African country. That item aimed to capture the occurrence of severe respiratory and circulatory compromise in the newborn period by asking the mother, "What was the color of the baby at birth?" The response options were 'normal,' 'blue,' or 'white.' The item had been imported into an African setting in which approximately 90 % of deliveries occur at a health facility were newborns are rushed off for urgent attention, without a mother observing the baby's condition. Moreover, the only time a respondent would see a 'white'

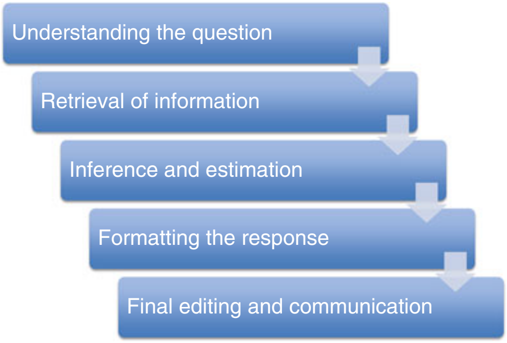


Fig. 18.1 Psychological stages of responding to questions

baby is if the parents had white skin. Understandably, the question was mostly ill understood and created a lot of confusion.

The result at the end of the response process may be that the respondent does not answer the question or that an inaccurate answer is given, and a lack of comprehension may or may not become clear to the interviewer. Not all respondents will ask for clarification when they are aware of their lack of comprehension. When it *is* clear to the interviewer, clarification of the question may not always succeed, as there may be cultural and language barriers. Moreover, people's personal interpretation frameworks are not always easy to change. The implications for questionnaire design and administration, discussed later in the chapter, are multiple and include the need to phrase questions in culturally appropriate terms and in the language of the respondent. Comprehension errors may be related to personal characteristics such as education level, alertness, socioeconomic status, etc. Any comprehension errors can be sources of considerable information bias, missing data, and hence imprecision. If questions are used to assess eligibility criteria, comprehension errors can result in selection bias.

Panel 18.1 Selected Terms and Concepts Related to Questionnaire Design and Administration

Conversational interviewing Style of interviewing in which interviewers interact freely with respondents after the question is read

Interview Method of data collection based on asking questions orally (face-to-face or over some communication medium) to persons and recording the elicited responses or their inferred meaning

Interviewee The person invited to answer the questions during an interview

Interviewer The person who asks the questions during an interview

Interviewers' guide/manual Document containing detailed step-by-step descriptions of prescribed procedures for preparing and carrying out an interview

Leading question A question that, by its phrasing, by the tone in which it is asked, by its positioning among other questions or by the ordering of its possible response categories suggests that a certain answer is or is not expected or socially acceptable

Options list A list of possible solutions to a problem statement

Question Written or spoken sentence, addressed to a research participant, aimed at eliciting a response or action that will assist with the measurement of an attribute/experience of the participant

Questionnaire Measurement tool composed of written questions, clarifications, answering spaces, and instructions on how to answer and how to proceed to other questions

(continued)

Panel 18.1 (continued)

Respondent The person who answers a question or questions

Satisficing Tendency for questionnaire respondents to settle for an approximate but less than optimal accuracy in their response

Self-administered questionnaire Mode of questionnaire administration whereby the respondent reads the questions and instructions (or hears a recorded version) and records answers

Standardized interviewing Mode of administering a questionnaire based on detailed instructions on how exactly the interviewer is to ask the questions and interact with the respondent

18.1.2 Response Process, Stage-2: Retrieval of Information

Given the respondent's understanding of the question, (s)he will now try to retrieve the information considered necessary. Information retrieval refers to facts retrieved from memory or from external sources, such as family members' memories, co-workers' memories, databases, diaries, or household files. For an event or experience to be remembered or retrieved, a record of it must be available, either under the form of physical data or a stored memory. Respondents cannot be expected to retrieve facts that have never been or are no longer encoded in memory or saved as an accessible physical or electronic record. Errors arising from a deficiency at this stage of the measurement process are termed 'encoding errors' in psychology. When the deficiency is one in retrieving from memory, they are called 'recall errors.' These may again take the form, at the end of the response process, of non-response or of misreporting. They can be related to participant attributes and lead to biased estimates and decreased precision.

Forgetting is the major process leading to recall errors and hence to recall bias (*See:* Chap. 2). In general, experiences must be very stressful or otherwise highly impactful and infrequent to be remembered for a long time (say more than a year). Questionnaire designers must keep in mind that asking respondents to count and report a frequency of a *common behavior* in some defined calendar period in the past is among the most difficult tasks one can ask of a respondent. For example, the question "How often have you eaten chicken in the last 12 months?" is a cognitively extremely demanding question (Jobe et al. 1990). One difficulty with it is that people's memories tend to relate to typical episodes in their personal history ('the time I lived in village *x*,' 'the time I worked for employer *y*), rather than to the defined calendar time episodes the researcher would like to know about (Schwarz and Oyserman 2001). This inherent memory structuring helps to explain why the construction of personal history calendars as an initial part of an interview process can often enhance recall accuracy of behavioral information, especially the accuracy of event dates.

Another problem with event dating is *telescoping*. Forward telescoping may be the most common problem and often concerns stressful events that are remembered as more recent than they actually were. Backward telescoping happens when recent events are remembered as more distant than they actually were.

As to short- and medium-term memory, recall accuracy is commonly an object of methods-oriented epidemiological investigation, and designers of questionnaires should thus verify the available evidence in the literature about what is a reasonable recall period for the specific type of event of interest. For example, a period of 2 weeks is generally considered the maximum recall period for questions to mothers about diarrhea in their children (Martorell et al. 1976). Schwarz and Oyserman (2001) suggested that, for events that are highly memorable, recall accuracy tends to increase by decomposing the recall period in sub-periods about which separate questions are asked. In taking this approach, one should work back from more recent periods to earlier periods rather than the other way around. Recall accuracy also tends to increase when the participant is given more time to think. The accuracy of retrieved information depends on how much effort the respondent is able and willing to make to remember and/or lookup information. 'Satisficing' can occur at this stage, meaning that the respondent settles for making little mental effort in tracing the information. Researchers should be aware that recalling relevant behaviors from memory can be time-consuming and that satisficing may be induced by any form of pressure to speed up the response process.

18.1.3 Response Process, Stage-3: Inference and Estimation

Additional mental effort is often required to further use the remembered events for counting or estimating total numbers of events; estimating average ('usual') frequencies or intensities; comparing various events to decide about the most intense or the least intense; and calculating durations (e.g., elapsed times) or other abstractions. For these tasks, too, the respondent decides what amount of motivation and time (s)he will spend and what level of accuracy (s)he will aim for. Satisficing occurs when the task seems too daunting (*hint:* terminal digit preference in the reporting of numerical values can be a manifestation of satisficing).

When questions are asked about prolonged periods, such as 'in the last year,' one naturally remembers best the last few weeks or months. Respondents may therefore be tempted to extrapolate a current or recent pattern to a longer time span.

When questions are asked about average intensity or usual intensity of a fluctuating or recurrent subjective experience (pain, anxiety, etc.), the answer may be positively biased because respondents' memories tend to be heavily influenced by the worst episode or the peak in experience as well as by the most recent episode (Streiner and Norman 2008).

Context effects may also influence reported past behavior. For example, in the evaluation of behavioral interventions, reported pre-intervention behavior tends to be worse when it is asked about after the intervention than when it is asked about before the intervention (Ross and Conway 1986). Another example of a context effect on inference and estimation is that respondents tend to report higher frequencies and severities of common mental-behavioral characteristics when the response options list contains mostly higher frequency/intensity options than when the list contains mostly lower frequency/intensity options (Schwarz and Oyserman 2001).

18.1.4 Response Process, Stage-4: Formatting the Response

The next mental process for the respondent is to prepare a response to the question in the format expected by the researcher. The major types of formats are open answer versus lists of response categories. As a concrete example of the latter, the respondent may have estimated a usual frequency of nine alcoholic drinks per day but may need to choose from a list of response options (e.g., '0–3', '4–6' and '7 or more'). Preceding options lists, there may be instructions about:

- How to choose (e.g., 'tick on option' or 'tick all applicable options')
- The measurement units to use (e.g., the form asks for stature in centimeters)
- The measurement scale to use (e.g., the form asks for the number of alcohol servings rather than the number of drinks, as 'drink' could be interpreted to mean 'glass,' each of which might contain more or less than one serving of alcohol)

Satisficing can also occur at this stage, especially if the list of response options is long or difficult to read. The length of the response options is therefore important. Five to seven options are often seen as a maximum. Options in the beginning of the list tend to be chosen more often in self-administered questionnaires whereas options at the end of the list tend to be chosen more often during telephone or face-to-face interviews (Schwarz and Oyserman 2001). This implies that, except for short options lists, response options should rather be presented as separate questions.

18.1.5 Response Process, Stage-5: Final Editing and Communication

In the final stage of the response process, the prepared response (chosen category, value, or reply) is briefly reflected upon and then communicated to the interviewer or written (ticked, circled, etc.) on the questionnaire. The respondent may, however, decide to edit the answer before communicating it, bringing in considerations other than accuracy. These considerations may concern social desirability or fear of disclosure. For example, the respondent may think that ticking the box '7 or more' alcoholic drinks per day will be seen by the researcher as abnormal and decide at the last moment, for the sake of her/his own reputation, to tick the box '4–6' instead.

Social desirability motives may be pursued consciously or unconsciously. They can show as a tendency to present oneself as healthier, more adherent to treatment, more 'normal,' and wiser than one actually is. Reported financial income is also prone to these effects, and within a single survey, different groups of participants may edit their responses for differing reasons: lower income groups may under-report income because of anticipated financial assistance or over-report to avoid stigma, whereas wealthier participants may under-report income to avoid social or tax repercussions. Sometimes a phenomenon opposite to social desirability occurs, if a direct benefit of ‘faking bad/unhealthy/deviant’ is expected. When social desirability motives affect the measurement of an attribute, the possible consequences in epidemiological studies include social desirability bias through: (1) under-estimation of the frequency and/or magnitude of socially undesirable attributes; (2) over-estimation of the frequency and/or magnitude of socially desirable attributes; and (3) biased estimates of the strength of association with other attributes.

The so-called *hello-goodbye effect* (Streiner and Norman 2008) means that before an intervention some people have a tendency to exaggerate their condition in the hope of getting the best possible care, whereas after an intervention, they may tend to present themselves as healthier than they are as a form of gratitude to the health workers. The consequence for interview-based research is obviously the danger of a falsely strong observed effect of the intervention on self-perceived health or on outcomes that rely on questions about symptoms.

18.1.6 Personal Characteristics of Respondents Affecting Responses

18.1.6.1 Personal Reference Points for Judgments

Another important lesson that epidemiologists have learned from cognitive psychology and from methods-oriented research about health surveys concerns the way people rate their preferences and intensities of experiences. When asked for such information, persons may take various reference points as a basis for making their judgment (Fienberg et al. 1985). The importance of this phenomenon for research was well illustrated by Groves (1991). He asked two questions about general health [reformulated]:

1. Would you say that your own health in general is excellent, good, fair, or poor?
2. When you answered question-1 about your health, what were you thinking about?
   - Others of the same age?
   - Myself at a younger age?
   - Myself now as compared to 1 year ago?
   - Other

The frequencies of the answers to the second question were highly revealing about the general and important issue of personal points of reference for judgments.

This implies that the researcher designing a question must try to know about (or at least anticipate) possible variations in such reference points and, if necessary, to learn about them in a pilot exercise. When the variation in reference points is important, one should provide the respondent with one clear reference point, or, split the question into several questions each with a specific reference point. For example, when asking a question about self-perceived general health, as above, one could ask “When you compare your health now with your health 1 year ago, would you say that your health now is good, fair, or poor?” Yet this approach would still be less than ideal because many people do not have an accurate recall of their health status 1 year ago. Indeed, personal reference points for judgments may shift considerably over time. This has important consequences for the validity of assessing changes in subjective attributes, which as a rule should be viewed with considerable skepticism, especially when efficacy of an intervention on a subjective attribute is evaluated.

18.1.6.2 Personal Characteristics Affecting Response Accuracy

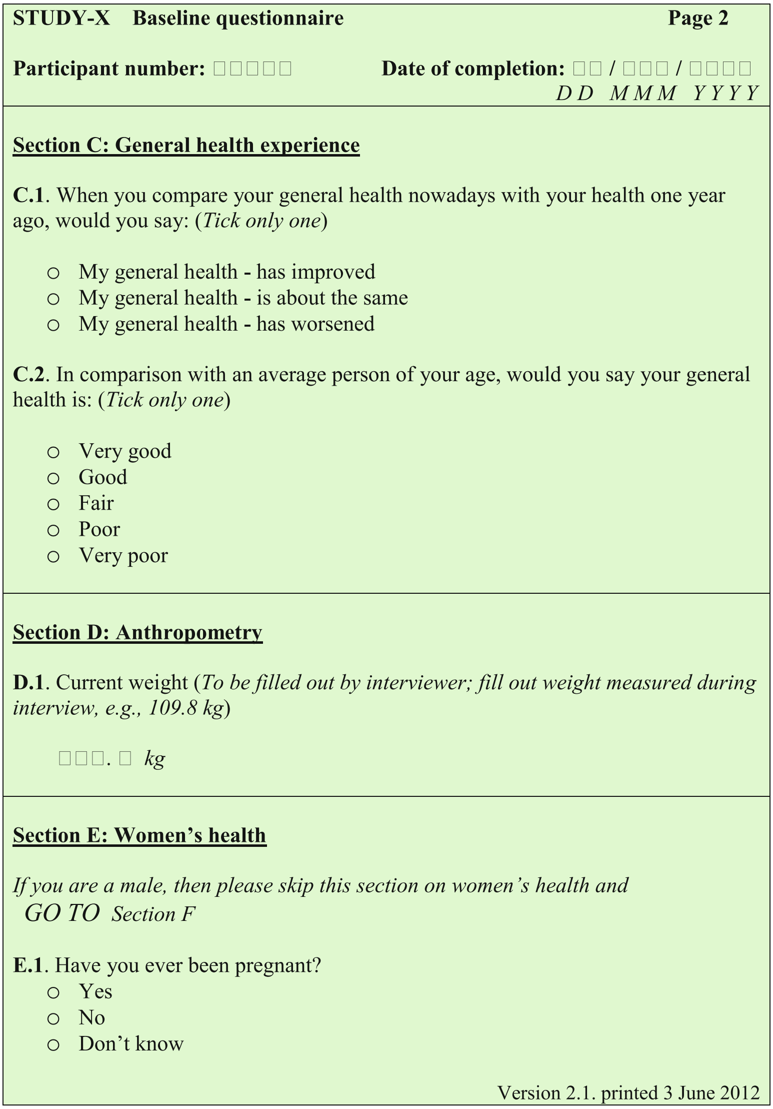
Inclination to satisficing or optimizing may vary individually, and so may the susceptibility to be influenced by social desirability motives or fears of disclosure. ‘Yeah-saying’ and ‘nay-saying’ mean a preference for ‘yes’ and ‘true’ answers or ‘no’ and ‘false’ answers, respectively. Many people do have a slight tendency, and some have a strong tendency for one of them. A way to minimize the effects of this is to make sure that questions are formulated such that, for the average respondent, one expects that about half of the answers will be ‘yes’/’true’ and half of the answers will be ‘no’/’false’ (Streiner and Norman 2008). ‘End aversion’ is a reluctance of many people to use the extreme options in an options list of answers. The consequence is an under-estimation of frequencies of extreme categories. A possible solution, if one wants to minimize the effects of this phenomenon is to broaden the extreme categories (Streiner and Norman 2008). For example one could use ‘always or nearly always’ instead of ‘always’ and ‘almost never or never’ instead of ‘never’. Alternatively one may conceal the true extreme categories by adding extremes of a nearly impossible magnitude that nobody is expected to choose. Finally, epidemiologists should remember that age, illness, sickness, and treatments can affect all stages of the response process.

18.2 Questionnaire Design

18.2.1 Standard Components of a Questionnaire

The main building blocks of a questionnaire are ‘items,’ which are units composed of a question with instructions, response options, and answering spaces. Items about a common theme are arranged in clearly delineated sections and linked through alphanumerical sequencing, combined with skip instructions when appropriate. In addition to the items, there may be spaces on the questionnaire that serve administrative or quality control purposes. Most questionnaires will have several onscreen or printed pages. Printed questionnaires may have one or several write-through pages attached to each numbered page (e.g., one for data entry and one for archiving). Studies may use several questionnaires administered in the same session or over multiple sessions.

Figure 18.2 shows the classical components of a questionnaire. Each single page of a questionnaire has a header section that identifies, as a minimum, the study, the questionnaire within the study (if several exist), the page number, the participant identification number, and the date of completion. Participant numbers and dates of completion may be pre-printed. Note that all instructions are traditionally given in *italics*. A small footer indicates the version of the questionnaire and the printing date.



STUDY-X Baseline questionnaire **Page 2**

Participant number: □□□□□ **Date of completion:** □□ / □□□ / □□□□
DD MMM YYYY

Section C: General health experience

C.1. When you compare your general health nowadays with your health one year ago, would you say: (*Tick only one*)

○ My general health - has improved
○ My general health - is about the same
○ My general health - has worsened

C.2. In comparison with an average person of your age, would you say your general health is: (*Tick only one*)

○ Very good
○ Good
○ Fair
○ Poor
○ Very poor

Section D: Anthropometry

D.1. Current weight (*To be filled out by interviewer; fill out weight measured during interview, e.g., 109.8 kg*)

□□□. □ *kg*

Section E: Women's health

If you are a male, then please skip this section on women's health and GO TO Section F

E.1. Have you ever been pregnant?
○ Yes
○ No
○ Don't know

Version 2.1. printed 3 June 2012

Fig. 18.2 Excerpt of a questionnaire form with the classical components of header section, items organized into sections, questions, answering spaces, options lists, and instructions

18.2.2 General Approach to Questionnaire Development

The first element in the general strategy to developing a questionnaire is to avoid anything that could confuse, bore, embarrass, or otherwise burden either the interviewer or respondent. This element encompasses (1) making the questionnaire as clear, short, simple, friendly, and attractive as possible, and (2) making all possible efforts to keep motivation high.

The second element is to account for what is known about psychological response stages and influences of personal characteristics as discussed above.

The third element is to draw from what is known already about the validity of specific questions. It is unwise to produce a questionnaire item de novo if a suitable version of the item is known to exist, has been used in other studies, and has produced reliable and accurate information, except when there are reasons to believe that a translation, update, or cultural adaptation is necessary. Questionnaire developers' websites or organizational repositories may provide access to adapted and/or translated versions that are suited to a particular research site. For example, the developers of the 'Strengths and Difficulties' questionnaire hosts a website that provides details about the questionnaire and a repository of versions translated into various languages (http://www.sdqinfo.com/). Another example is the World Health Organization research tools for substance abuse (http://www.who.int/substance_abuse/research_tools/en/). More examples are given in Table 10.4. That being said, one should not assume that an item is acceptable for use and has been validated merely because it has been used in other studies.

The fourth element is to make maximal use of possibilities to promote data integrity after questionnaire filling (details discussed below).

18.2.3 Practical Recommendations for Questionnaire Design

Panels 18.2, 18.3 and 18.4 are checklists for the content and format of questionnaire items and for the formatting of the entire questionnaire.

Panel 18.2 Checklist for the Content of a Questionnaire

- Do not collect personal identifying information unless necessary
- Avoid culturally sensitive questions or ask them sensitive questions only after an extensive and appropriate introduction with an explanation; place sensitive questions them towards the end of the questionnaire or relevant section
- Avoid leading questions or leading sets of response options. The phrasing of the question and the wording and sequence of the response options can be suggestive of a socially desirable or a typical 'normal' answer
- Avoid confusing and unclear questions or response options
- Avoid the use of specialized terms and medical jargon and the use of abbreviations and acronyms
- Avoid vague references to the past, e.g., "Compared to baseline…" or "Since last visit…" Respondents may not understand what exactly is meant by that

(continued)

Panel 18.2 (continued)

- Provide a common point of reference when asking about a current situation
- Avoid questions that refer to periods too far in the past or that would otherwise be challenging to answer based a high likelihood of forgetting
- Try to minimize the telescoping (i.e., the event is remembered but the date is inaccurate). Most emotional events tend to be remembered as more recent than they really were
- Choose an appropriate recall period
- Avoid open-ended questions as much as possible. If open-ended questions are necessary, provide sufficient space. Give instructions on desirable elements and degree of specification of the answers to be recorded in the open-answer field
- Design items only for data that will be analyzed
- Avoid duplication of items. Validity of responses can be checked by asking several questions related to the same topic to see if responses are consistent, but these questions should be phrased in different ways

Panel 18.3 Checklist for the Format of Questionnaire Items

- Plan and sketch the design of the items before committing to paper
- Formulate the items in the language of the respondent
- Make options lists that are non-overlapping and as exhaustive as possible
- Add an option for ‘other’ or ‘don’t know’ whenever relevant
- Avoid complete non-response by adding options such as ‘Prefer not to respond’ or ‘Unsure of how to respond’
- Clearly indicate whether only one option should be chosen (using *circles* as tick boxes) or whether multiple options can be chosen (using *squares* as tick boxes)
- If more than one response option can be selected, it is usually preferable to list all with yes/no/don’t know options for each rather than providing an overarching instruction to ‘select one or more’ from the list.
- Use answering spaces of appropriate length
- Clearly indicate the unit of measurement, e.g., cm, kg (SI Units)
- Use consistent units of measurement for similar items in the questionnaire
- Adapt questions to mode of administration. In telephone-administered interviews options lists cannot be as long as in self-administered questionnaires

Panel 18.4 Checklist for Questionnaire Formatting

- Include all main components of a questionnaire (Illustrated in Fig. 18.1)
- Format all items according to Panel 18.3
- Format the printed questionnaire pages so that they resemble the electronic version used for data entry
- When producing another language version, check translation accuracy by comparing an independent back-translation with the original. Guidelines are available at: http://www.who.int/substance_abuse/research_tools/translation/en/
- Provide detailed instructions on questionnaire administration in a user's manual; make sure each interviewer is trained accordingly and has the instructions available during each interview
- In self-administered questionnaires, visual attractiveness of the forms and large enough font sizes are of extra importance
- Surround sets of related items with a box
- Avoid crowding the questions on a page; separate questions clearly
- Use no more than two columns per page
- Separate columns with clearly visible lines
- Use consistent page designs
- Avoid splitting an item across pages; especially avoid having the response options cross pages
- Be consistent with codes and options list throughout the questionnaire, e.g., not: 'yes, no, don't know' on one page and 'no, yes, don't know' on another page
- Make sparse use of skip and stop instructions; limit skips by optimal placement of answers; tell where to go next, not what to skip
- For printed versions, use thick paper that can withstand repeated handling; if finances allow use transparent plastic folders for each form
- Use consistent date formats throughout the questionnaire; the least confusing date format is DD-MMM-YYYY (e.g., 08-FEB-2012)
- Avoid questionnaires that are very long
- Finalize the questionnaire after several practice runs
- Avoid loss of printed pages by properly attaching all the pages

18.2.4 Questionnaire Design Decisions to Facilitate Data Entry and Analysis

To facilitate data entry, one can consider the following options when developing a questionnaire:

- Design the electronic data entry form to resemble the paper form as much as possible
- Provide code lists on the form as much as possible (perhaps in italics and with a smaller font size)

- Be consistent with codes and options lists throughout the questionnaire. Try not to use separate, different code lists for data entry: data entry persons should ideally be able to type directly what they see
- For closed answers, use boxes with a space for each character. Mind the appropriate number of characters and the number of decimal places
- Design the questionnaire in such a way that a data entry screen can be easily made with a similar design
- Ask feedback from data entry persons before finalizing a questionnaire
- Mimic interview skip patterns in data entry forms

For making the questionnaire analysis-oriented consider the following options:

- Envisage the analysis when designing questionnaire items
- Only collect data that will be used in planned analyses of the primary and secondary outcomes
- If the analysis uses derived variables (computed from raw data) make sure all necessary elements for the computation are collected on the forms, e.g., data elements for socio-economic status, dates for length of follow-up, etc.

18.3 Types of Items in Questionnaires

18.3.1 Structured, Semi-structured, and Open-Ended Items

All types of items include a worded question but they differ in the way responses are recorded. A *fully structured item* provides a clear measurement scale on which one or more specific values can be placed. For instance, it may provide a list of response options from which one or more need to be chosen. Another example is an item that depicts a visual analog scale (Fig. 18.3), on which a single value needs to be indicated. Yet another example is an item with clearly indicated spaces to record measured height.

A *semi-structured item* equally represents a clear range of options, but one or more of the options trigger a sub-question, the response to which is to be recorded as free text. The item is thus only structured to a certain level. This type of item is useful when an explanation or specification is desired of a chosen option. For example "If 'other,' please specify: ___________" or "If yes, please explain reasons: ______________."

A *fully open-ended item* simply provides a dedicated open space where the respondent or interviewer can freely write a textual answer to the question. Though

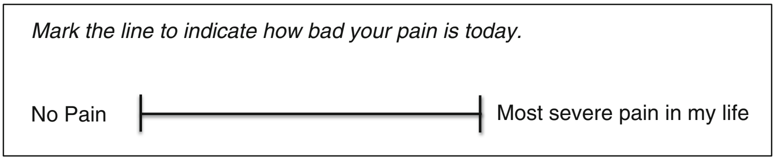

Mark the line to indicate how bad your pain is today.
No Pain
Most severe pain in my life

Fig. 18.3 A visual analog scale – VAS

this text is free in principle, the open-ended item can include instructions (e.g., algorithms) to help focus the respondent on particular aspects of content or instructions to request certain restrictions in the format (e.g., length) of the response.

Questionnaires that are mostly composed of structured and semi-structured items are called *structured questionnaires*, and those mostly containing open-ended items are called *open-ended questionnaires*.

18.3.2 Items for Counts and Continuous Attributes

For the redaction of items for continuous attributes, most of the guidelines in Panel 18.2 are relevant. Here we will discuss some particular issues and some typical forms of items.

As to the precision and units of measurement, it would be unfair to ask respondents to report a quantity in units they are unfamiliar with, or to ask to report it with a precision that is unlikely to be remembered or traced. Thus, the item needs adaptation to locally used units of measurement and locally used precision. Several options may need to be offered if there is heterogeneity in this local tradition.

18.3.2.1 Respondent-Reported Measurement Values

The concept of *respondent-reported* includes both ‘self-reported’ and ‘reported for a child or other person.’ It is usually understood that this relates to information retrieved from memory. For example, self-reported weight and height are commonly understood to be the most recent weight and height measurement values the respondent remembers. Obviously some measurements must have been done at some point in the past, but when exactly this measurement was done, how accurate the measurement value was, and how well it is remembered and reported are unknown and highly variable. In addition, remembered values may be outdated, e.g., the respondent may have gained or lost a lot of weight since the measurement (s)he remembers.

Numerous studies have indeed shown the lack of reliability of self-reported weight and height values. In general, respondent-reported numerical values based uniquely on memory need to be avoided as much as possible. In mailed survey questionnaires or in other situations where direct measurement by an observer is impossible, it can be useful to request that respondents use additional sources other than memory. For example, the item in the mailed questionnaire could include an instruction for the respondent to trace or verify the numerical value, time or date, with the help of a diary or by looking up other written information. It may also contain a request to perform the measurement again before answering, e.g., using an available scale to measure one’s weight. Whenever different sources are possible it becomes important to include a sub-question to record the sources used (e.g., memory, documentation, new measurement, or combinations).

18.3.2.2 The Item for Age Determination

Age is frequently used as an eligibility criterion and also as a study variable. Errors in age determination can thus potentially lead to selection bias, information bias, and confounding. Age is a continuous attribute commonly defined as ‘time elapsed since birth.' Age is often measured by calculating the time interval between two dates: the date of birth and the date of filling the questionnaire. Both these dates are normally recorded in epidemiological studies and can often be accurately provided by respondents themselves. Alternatively, but somewhat less reliably, one can ask the respondent directly for an age or an age at last birthday. This alternative approach is based on the assumption that respondents know their birthday, remember their age at their last birthday, and sometimes that they can calculate months elapsed since their last birthday. Some participants may have difficulty remembering or counting months. Also, not everybody is familiar with the months of the Gregorian calendar: in some societies, one rather calculates in moon cycles than in months.

Consequently, the birth date-based method is generally preferable over asking for age. It makes sense to include into the item an instruction asking respondents to verify any document that may contain the birth date, preferably the birth certificate or an identification card. The same is true if the questionnaire is to be interviewer-administered, as the interviewer can then verify the documents. In areas where such documents are not systematically available, asking for a document-endorsed birth date or an age may be helpful for some but not for all. The measurement of age in such areas should then, for some of the participants, involve an interview during which approximate birth dates are derived with the help of a local events calendar or via reference to people of the same age (e.g., former class mates) who *do* know their birth date or age exactly. A sub-question is then useful to distinguish participants for whom this method was applied.

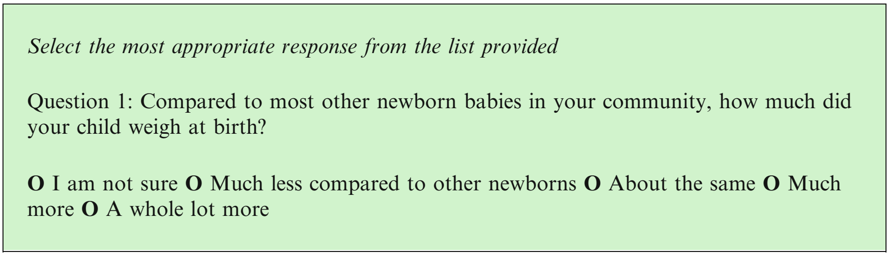
18.3.2.3 Visual Analog Scales: VAS

A VAS consists of a line and two described endpoints representing the least possible and the most possible amount of an attribute (Fig. 18.3). There are strengths and weaknesses of this method (Streiner and Norman 2008). A VAS is generally appealing although some respondents may not find it easy to understand. The optimal wording to describe the endpoints can be a problem and a source of variation. For example, an endpoint described as 'the worst possible anger' may mean totally different things to different respondents depending on their experiences and imagination.

18.3.2.4 Ordinalized Scales

Sometimes the measured attribute is continuous but the scale for measurement is ordinalized (presented as a sequence of ordinal levels). The optimal number of levels is usually in the range of 5–7. Ordinalized scales include the following types:

- Horizontal options lists with circles (Fig. 18.4)
- Likert scales (Fig. 18.5): These are often used to measure subjective levels of agreement, acceptance, or perceived likelihood. They are characterized by the fact that there are levels of opinion in either direction away from a neutral opinion. The neutral opinion itself may or may not be mentioned as a separate level, but it usually is
- Juster scales (Fig. 18.6) are used mostly for subjectively estimating the probability of an event. The ordinal levels are described by a numerical probability combined with a worded interpretation of that same probability

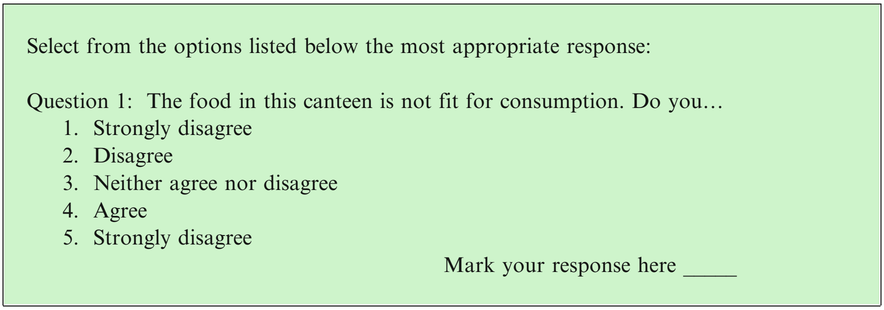


Select the most appropriate response from the list provided

Question 1: Compared to most other newborn babies in your community, how much did your child weigh at birth?

O I am not sure **O** Much less compared to other newborns **O** About the same **O** Much more **O** A whole lot more

Fig. 18.4 A horizontal options list with circles showing incremental values



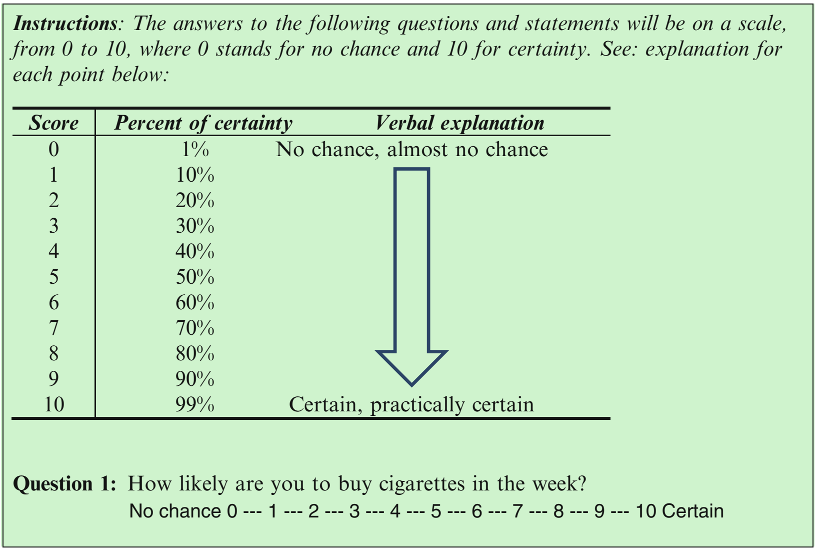
Select from the options listed below the most appropriate response:

Question 1: The food in this canteen is not fit for consumption. Do you…

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly disagree

Mark your response here _____

Fig. 18.5 A Likert scale



Instructions*: The answers to the following questions and statements will be on a scale, from 0 to 10, where 0 stands for no chance and 10 for certainty. See: explanation for each point below:*

Score	***Percent of certainty***	***Verbal explanation***
0	1%	No chance, almost no chance
1	10%	
2	20%	
3	30%	
4	40%	
5	50%	
6	60%	
7	70%	
8	80%	
9	90%	
10	99%	Certain, practically certain

Question 1: How likely are you to buy cigarettes in the week?
No chance 0 --- 1 --- 2 --- 3 --- 4 --- 5 --- 6 --- 7 --- 8 --- 9 --- 10 Certain

Fig. 18.6 A Juster scale

Instructions: Ask the child to show you on the face scale how much the foot hurts today

Question 1: Foot pain level indicated by the child

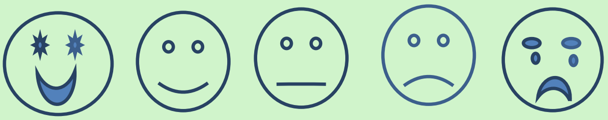


Fig. 18.7 A face scale

- Face scales (Fig. 18.7): The ordinal levels are represented by faces expressing a range of moods or of pain (Stinson et al. 2006). This makes it more feasible for children and for those with reading difficulties. Face scales can be seen as a special form of Likert scale

18.3.3 Items for Non-continuous Attributes

For the redaction of items for non-continuous attributes, most of the guidelines of Panel 18.3 are relevant. We briefly discuss some particular types of items of interest.

18.3.3.1 The Item for Sex Determination

*S*ex is one of most frequently used variables in health research. There exist distinctions between chromosomal/biological sex, genital sex, other phenotypic sexual characteristics, sexual orientation, gender, and gender-related behavior patterns. Based on this, a small rate of mismatch is expected between respondent-reported sex and interviewer-reported sex. Respondent-reported sex and gender are expected to be more strongly correlated with each other, whereas interviewer-reported sex may be more influenced by phenotypic sexual characteristics and dress code followed. In practice, however, unless the research directly concerns issues around biological sex or gender, the mismatch will be negligible. Thus, a simple question with two response options (male/female or boy/girl) will usually be appropriate in all types of questionnaires and for all modes of administration.

Hint

Biological sex and gender are often used interchangeably, but they are in fact very different concepts. Biological sex refers principally to chromosomal patterning, where males are defined by the presence of a Y chromosome (i.e., XY, though XXY and XXYY are rare variants) and females are defined by the absence of a Y chromosome (i.e., XX, though XO and XXX are rare variants). Gender, on the other hand, is a social construct defined by behavior, actions, roles in society, and sexual orientation. Gender identification refers to a self-selected gender.

18.3.3.2 Items for Measuring Dichotomous Phenomena

The following types of phenomena are commonly measured:

- Whether or not a past event, experience, or activity has occurred, e.g., by the question 'Have you ever taken oral contraceptive pills?'
- Whether or not a state is present or absent, e.g., by the question 'Are you currently married?' Note that attributes can be nested and hierarchical and that, for this reason, a particular level of one attribute may be seen as a dichotomous attribute on its own. For example, age is a continuous attribute but being an adult can be considered to be an attribute on its own
- Opinions about whether a particular statement is true or false

The items for these types of attributes often contain short questions with 'Yes-No-Don't Know' or 'True-False' response options. Multiple dichotomous characteristics can be measured in a single item starting with a general question such as 'Have you ever taken any of the following medicines?' or 'Have you ever had one of the following illnesses?' or 'Are the following statements true or false?' Such items assessing several dichotomous attributes may do so with the aim of measuring a higher-level latent attribute. For instance, a list of questions about the use of particular medications may aim at measuring whether treatment for a particular illness was given. Or, an item containing a list of statements with 'True-False' options may aim at measuring a level of knowledge or a psychological-behavioral characteristic. These examples are illustrations of the fact that attributes can be multi-dimensional and nested.

18.4 Questionnaire Administration

For questionnaire administration it is important to keep in mind that anything that can confuse, distract, bore, embarrass, or otherwise burden the respondent or the interviewer will tend to adversely affect accuracy and completeness of the recorded responses. In this section we will discuss administration styles, specific training, user's manuals and ethical issues of questionnaire administration with a special concern for maximizing accuracy and completeness. As a reminder, in Chap. 10 we discussed *modes* of administration in the context of designing a measurement plan. The important choices to make included:

- Self-administered vs. interviewer-administered
- Face-to-face vs. internet vs. telephone vs. mixed administration
- Administration at home vs. clinical care settings vs. other
- Proxy-respondents vs. interviewing enrolled study subjects

One should make sure to always record the type of respondent used, for example self-about-self, mother-about-child, other-caregiver-about-child, etc. When an adult is reporting about a child, especially in environments with extended care-giving practices, it may be necessary to define the relationship of the adult to ensure validity of responses. Generally speaking, proxy-respondents must be avoided as much as possible if the enrolled subject is capable of providing accurate answers.

18.4.1 Styles of Interviewing

The style of interviewing tends to have an influence on the accuracy of the responses. Panel 18.5 lists the main styles and the expected effects on responses.

18.4.2 Training of Questionnaire Administration

Panel 18.6 shows a checklist of selected training topics around questionnaire administration.

Panel 18.5 Main Styles of Interviewing

Standardized interviewing

- All interactions with respondent are prescribed and written in the interviewer's guide as a step-by-step process. This style rules out most interviewer influences on responses.

Conversational interviewing

- This mode allows interviewers to interact freely with respondents, which minimizes errors due to poor understanding of the question by the respondent but also introduces some interviewer variance. The interviewer's guide in this case may contain information on common misunderstandings and (perhaps several) possible ways of responding to them.

Conversationally flexible interviewing

- This mode of administration combines both previous styles: a standardized part and a free part to each question. Alternatively, there can be a standardized approach for one question and a conversational one for another question (Biemer and Lyberg 2003). Conversationally flexible interviewing leads to the same accuracy as standardized interviewing when the question is easy to answer, and it has been found to allow for better accuracy than standardized interviewing when the question is difficult.

Panel 18.6 Checklist for Training on Questionnaire Administration

- Provide detailed instructions in a user's manual; make sure each interviewer is trained accordingly and has the manual available during each interview; this includes moving through the questionnaire at an appropriate pace, writing legibly, using permanent ink, etc.

(continued)

Panel 18.6 (continued)

- Sufficient training and supervision during the preparatory phase should ensure that the interviewer establishes rapport with the participant during face-to-face interviews even when the user's manual is constantly referred to
- Special training on the use of code lists
- Special training on skip patterns
- Special training on uniform date recording
- Special training on items that require complex probing, e.g., age or date assessments based on a calendar of local events
- Special training on the specificity of terminology, length of text, etc. for items involving free-text
- If optical scanning and recognition is used for data entry, organize special training to avoid common types of computer-misreads

18.4.2.1 Source Document Standards

ICH Good Clinical Practice guidelines state that 'Source data is all information in original records and certified copies of original records of clinical findings, observations, or other activities in a clinical trial necessary for the reconstruction, evaluation and validation of the trial. Minimum standards as to the quality of source data are currently prescribed for clinical trials only. However, many of the specific guidelines are potentially useful for other types of studies as part of a strategy of maximizing data quality. Selected examples of this are listed below.

- No changes to original data can be made without signed justification
- No personal identifiers on questionnaires except with special permission
- All questionnaires and any copies must be signed, credentialed, and dated. Copies must be certified to be an exact reflection of the original
- Any questionnaire as well as any written communication about the participant (e.g., lab report) must mention subject study number
- A master-list must be kept linking study number to personal information, only accessible by the investigator (not data management personnel)
- Every protocol deviation (e.g., missed visit) should be documented with reasons for the deviation stated
- Never obliterate entries that require correction (no barring, no use of white-out)
- Never destroy original documents if they require error correction
- Follow-up questionnaires must be kept in chronological order
- Enrollment forms must document compliance with each single eligibility criterion
- All source documents must be kept either in a same place or in a way that a monitor can easily access them during a monitoring visit

18.4.3 The Questionnaire User's Manual

Also known as the Interviewer's Guide or Instruction Sheet, the questionnaire User Manual contains detailed instructions on the use of the questionnaire form. User's manuals usually have a section with general guidelines as well as question-specific sections. The content is influenced largely by the chosen style of interviewing. One can also consider providing a library of pre-coded answer sets in the user's manual, e.g., occupational categories. One should make sure that each interviewer is trained extensively on how and when to use the instruction sheets. It should be a formal obligation for the interviewers to have the instruction sheets available for consultation during each interview. It is habitual to prepare a Standard Operating Procedure based on the User Manual and field logistics; this will prevent deviation from the study protocol.

18.4.4 Ethical Considerations Around Questionnaire Administration

It is good to ensure privacy during questionnaire administration and to avoid non-intended disclosures. These measures optimize accuracy and limit item non-response rates. When the subject matter is anticipated to reveal emotionally sensitive issues, such as partner violence or mental distress, protocols should include details on emergency counseling and professional services. Periodic counseling of interviewers is also advised in such research, though data collectors should not do this counseling. Finally, adherence to source document standards, as described above, is another ethical imperative.

This chapter discussed questionnaire design and administration. Every time a direct measurement value or response is recorded or a biological sample is taken, a further challenge lays ahead, namely to preserve the integrity of these data and samples while they are processed. The maintenance of data and sample integrity is therefore the topic of the next chapter.

References

Biemer PP, Lyberg LE (2003) Introduction to survey quality. Wiley, Hoboken, pp 1–402. ISBN 0471193755

Fienberg SE, Loftus EF, Tanur JM (1985) Cognitive aspects of health survey methodology: an overview. Milbank Mem Fund Q 63:547–564

Groves RM et al (1991) Direct questioning about comprehension in a survey setting. In: Tanur JM (ed) Questions about questions. The Russell Sage Foundation, New York, pp 1–308. ISBN 9780871548429

Jobe JB et al (1990) Recall strategies and memory for health-care visits. Milbank Q 68:171–189

Martorell R et al (1976) Underreporting in fortnightly recall morbidity surveys. J Trop Pediatr 22:129–134

Ross M, Conway M (1986) Remembering one's own past: the construction of personal histories. In: Sorrentino RM, Higgins ET (eds) Handbook of motivation and cognition, vol 1. Guilford, New York, pp 122–144

Schwarz N, Oyserman D (2001) Asking questions about behavior: cognition, communication, and questionnaire construction. Am J Eval 22:127–160

SDQ (2012) Strengths and difficulties questionnaire. www.sdqinfo.com. Accessed Sept 2012

Stinson JN et al (2006) Systematic review of the psychometric properties, clinical utility and feasibility of self-report pain measures for use in clinical trials in children and adolescents. Pain 125:143–157

Streiner DL, Norman GR (2008) Health measurement scales. A practical guide to their development and use, 4th edn. Oxford University Press, Oxford, pp 1–431. ISBN 9780199231881

Tourangeau R, Rips LJ, Rasinski K (2000) The psychology of survey response. Cambridge University Press, Cambridge, pp 1–401. ISBN 9780521576291

World Health Organization. Process of translation and adaptation of instruments, YouthinMind. http://www.who.int/substance_abuse/research_tools/translation/en/. Accessed Sept 2012

Maintaining Data Integrity

19

Jan Van den Broeck, Jonathan R. Brestoff, and Meera Chhagan

Unfortunately part of the questionnaire forms have been eaten by termites.

Anonymous

Abstract

Data have integrity when they are free of data abnormalities and data manipulations. Maintaining data integrity is a responsibility of all those involved in research, not only data managers. The costs of data integrity problems and of responding to them when they are discovered can be high; therefore, prevention of data integrity problems is far better than correcting them after they have been made. However, even when good strategies are employed to prevent data integrity problems (a topic discussed previously), it is inevitable that some data integrity problems will occur. The specific foci of this chapter are thus on (1) operational problems occurring in spite of detailed quality assurance and data management plans and (2) adaptive responses. Some data integrity challenges and possible solutions in resource-limited settings are also highlighted.

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J.R.Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

M. Chhagan, Ph.D., FCPaed
Department of Paediatrics, University of KwaZulu-Natal,
Durban, South Africa
e-mail: Chhagan@ukzn.ac.za

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_19,

19.1 Concepts of Data Integrity

As we have discussed in previous chapters, data management is a complex and laborious process. It begins during the study planning stage, far in advance of recording the first data points. Along the way, the investigative team aims to limit the number of data errors and manipulations. Failure to do so adequately results in data *without* integrity and, consequently, study findings that are biased. Such unfortunate developments also result in substantial losses of time, effort, and resources and waste human subject data. On the other hand, data *with* integrity are free of errors and manipulations. Thus, maintaining data integrity is a key ethical responsibility of all epidemiologists and their investigative teams. Panel 19.1 highlights some of the key terms and concepts relevant to data integrity. These terms will recur repeatedly here and in Chap. 20.

19.2 Threats to Data Integrity After Initial Recording or Sampling

Threats to data integrity fall into three categories: (1) organizational weaknesses affecting individual performance levels during activities relevant to maintaining data integrity; (2) failures of equipment and infrastructure needed for maintaining

Panel 19.1 Terminology Related to Data Integrity

Analysis dataset Selection of fields and records extracted from a database, used for a particular statistical analysis

Data Recorded information

Data abnormality Data which are deficient, excessive, outlying or inconsistent in comparison with prior expectations

Data cleaning Process of detecting, diagnosing, and editing data abnormalities

Data flow Passage of recorded information through successive information carriers

Data integrity Freedom of data abnormalities and data manipulations

Data management The organization of data processing

Data manipulations Changes deliberately made to data values for reasons other than data cleaning

Data processing Recording, storing and extracting data, and cleaning and preparing data for analysis (*q.v.* data handling)

Database Organized set of data kept as a source for extracting analysis datasets

Erroneous data Missing or inaccurate (biased) data

Quality control of data Data management activity aiming at checking data integrity

Quality assurance of data Activities aiming at optimizing and maintaining data integrity

Table 19.1 Types of threats to data integrity

Players	Role in maintaining data integrity	Examples of threats to data integrity		
		Organizational weaknesses	Equipment and infrastructure problems	Deliberate human action
Communities and authorities	Support operations related to maintaining data integrity	No timely approvals for sample shipments	Irregular electricity supply for cold storage	Politically motivated or misperception-based boycotts
Sponsors	Provide resources for maintaining data integrity	Lack of interest in interim data reports or lack of capacity to issue queries	Insufficient budgets for QA/QC-related equipment	Retaliation against individuals who report data problems
	External monitoring of data operations			
Institutions	Provide resources for maintaining data integrity	Delayed hiring of QA/QC and data handling personnel	Providing insufficient space for QA/QC, data handling, sample storage, archiving	Misuse of research funds designated for data management
	Ethical oversight of data problems			
Scientific team	Supervise integrity of data processing	Not organizing restrictions on database access, no audit trails or performance metrics	Not insisting with sponsor and institution on more resources when needed	Hiding data problems; falsification or fabrication of data, analysis errors
Study personnel	Supervise data handling and chain of custody in data and sample flow	Loss of completed questionnaires due to weather or extreme circumstances; excess data entry errors	Use of equipment for purposes not described in protocol	Deliberate destruction of part of the data; neglecting back-up and archiving procedures
	Maintain database (backups and archiving)			
	Extract analysis datasets			
Study participants	Provide source data/samples	Not motivated to communicate well with study personnel	Incorrectly follow instructions on handling forms and samples	Deliberately reporting erroneous data

data integrity, and (3) deliberate human action. Many players in research studies have some capacity to support or undercut data integrity. Table 19.1 shows the main players involved in data-related study operations, lists their functional roles, and outlines types of threats to those roles.

Textbox 19.1 A Case Scenario Illustrating Data Integrity Problems with Samples and Laboratory-Based Assessments

Crucitti et al. (2010) reported issues of laboratory data validity that arose during a large international multi-center microbicide trial. The trial involved laboratory capacity building in peripheral trial sites, under the guidance of a central reference laboratory. In spite of careful logistical preparations, training, and intensive quality assurance and control, some serious problems arose in the course of the study. These included several false positive HIV test results from a peripheral lab site, DNA contamination of amplification-based tests, instances of sample mislabeling, and clerical errors in sample shipment lists. Interestingly, the false positive HIV tests were mostly due to fading skills gained during training. The DNA contamination was due to decreased attention to protocols for cleaning the laboratory and performing the necessary sample preparations. Upon discovery of these problems, corrective actions were taken, but there is no way to reverse the unnecessary anxieties caused by false-positive HIV test results or to recoup the costs associated with investigating sample contaminations.

In a rare case study of data integrity Crucitti et al. (2010) have shown that in well-designed and large studies with extensive quality assurance serious unexpected problems can occur. This case study (Textbox 19.1) illustrates well that most data integrity problems with samples and laboratory analyses arise from either deficient standard operating procedures or practical deviation from them related to human errors.

The case study of Van den Broeck et al. (2007) gives an account of biological sampling problems arising in an ongoing micronutrient trial. Proper quality control of samples started too late in the trial, mainly because the analyses of the samples only concerned a secondary outcome in the trial and the focus of quality assurance efforts had gone mostly to the primary outcome. This points to another commonly observed pattern of data integrity problems in general, namely that they tend to arise more commonly and/or tend to be more severe for data about secondary outcomes than for primary outcomes. In this trial, too, forgetfulness and simple human mistakes were found to be common causes of data integrity problems. Another interesting cause was programming errors in software for barcode printing that had remained undetected during piloting and validation.

Not all data integrity problems, however, arise from biological sampling, laboratory processes, or equipment. In fact, any process involving samples or information provided by subjects is liable to data integrity problems. For example, in our experience, there are quite commonly problems with questionnaire-based data collection, although their discovery always comes as an unfortunate surprise. These problems tend to relate to the following:

- Loss of, disappearance of, or damage to questionnaires
- Decreasing focus on supervision and quality control of filled-out questionnaires

- Discovery of cases of questionnaires that are inconsistent internally, inconsistent with data from previous questionnaires, or illegibly filled out by interviewers
- Temporary or permanent delays in data entry and data cleaning

In studies with high data quality expectations, such as clinical trials, the threshold for calling something a data integrity problem may be very low. What would be considered a minor problem in another type of study might be seen as a major problem by trial monitors. Trials and other longitudinal studies are more likely to face study fatigue among study personnel and among participants and to endure challenges associated with changing views of sponsors and local authorities. Indeed, these stakeholders and others (e.g., communities) are important in facilitating studies, but they can also cause unexpected dysfunctions or changes in study procedures. Such changes can be major threats to data integrity.

19.3 Adaptive Responses

A fast response is needed when serious data integrity problems surface spontaneously or when they are revealed by performance metrics, but not without a thorough examination of the nature of the problem. This examination must initially focus on people, not structures. The first matters are identifying the players involved, the quality of their interactions, their understanding of their own and others' responsibilities, and other possible reasons underlying poor behavior or performance. During this process structural weaknesses and possibilities for operational adaptations tend to emerge. After the sources of data integrity problems are identified, they should not be left to linger. A fast solution is critical. Fortunately, most problems have simple, easy-to-implement solutions. All that may be needed is a good conversation, an efficient meeting, or a simple change in timetable. In other instances, however, the solution may require complex and cumbersome procedures, and sometimes the problem and solution are clear enough but the resources needed to implement the change are insufficient. When complex problems and solutions are discovered, study investigators, coordinators, or data managers must not accept the fatalistic attitude of 'things are never perfect' and move on without an adequate solution. This attitude is analogous to satisficing, but the consequences of satisficing in dealing with data integrity problems are far more serious than those caused by a subject's satisficing while answering questions.

Data integrity problems and solutions tend to be particular in nature (i.e. they are specific to the study and problem at hand), but the following general questions deserve some attention in responding to any data integrity problem:

- Have the right people been contacted? Have the individuals been contacted who have the responsibility and power to do something about the problem?
- How can the decision mechanisms for facilitating the response be improved?
- Are there certain types of decisions that can be automated by a feedback link to some objective metric?
- What are the costs of various responses?
- What is the cost of not responding?

19.3.1 Structural Improvements of the Flow of Data and Samples

Quality control activities and data cleaning (*See:* Chap. 20) often allow the researchers to gain insight into the nature and severity of error-generating processes that depend on structural problems. Perhaps the most common structural problems encountered regarding data integrity relate to programming of data collection, data entry, post-entry data cleaning, data transformations, and data extractions. If these problems are identified early enough, solutions may be easier to develop and deploy. In this case, the researcher can often simply give feedback to operational staff to improve study validity and precision of outcomes. But late-discovered errors may be more difficult to address and may temporarily overburden data entry and QA/QC personnel.

Hint

When serious data integrity problems arise it may be necessary to amend the study protocol regarding design, timing, observer training, data collection, quality control procedures, and/or analysis strategy. In rare instances, it may even be necessary to restart the study. In these instances, it may be necessary to proceed with the advice and guidance of an ethics board.

19.3.2 Adjustments of Laboratory Practices in an Ongoing Study

When a laboratory testing problem surfaces, testing is usually put on hold and one must organize some fact-finding investigation to identify causes. Intensive communication is the rule. One must give immediate feedback, provide guidance, and retrain laboratory staff (for an example, *See:* Textbox 19.1). Guidance and re-training will normally consist of an on-site re-enforcement of Good Clinical Laboratory Practice guidelines (Ezzelle et al. 2008). This may involve bolstering QA/QC procedures by making them more intensive or detailed, and it might involve initiating new procedures previously considered unimportant, unaffordable, and/or cost-ineffective for the specific study context. Re-enforcements may concern the specific study only or relate to a sustainable upgrade of the accreditation level of the entire laboratory (e.g., ISO accreditation). Such quality upgrades, in turn, may be supported by and may fit in the context of strengthening wider networks, such as (inter)national laboratory systems (Wertheim et al. 2010; Nkengasong et al. 2009, 2010; Olmsted et al. 2010).

19.3.3 Budget Extensions and Supplements

Various problems may bring up a need to expand or supplement a study budget. The problem may be that the enrollment period needs to be extended, that costs were underestimated, that costs changed unexpectedly, or that received money was devalued. When dealing with these issues, it is good to keep in mind that:

- Funders often dislike budget extensions beyond what was initially granted
- It may be acceptable to the funder to transfer money from one line item to another so that any savings on one front can benefit the problematic budget items

- Any savings should not have more than a minimal impact on data quality
- The extra funds may have to come from another funder, perhaps from the research institution itself, or, in the extreme case, from the investigator
- Many funders do not object to other funders being involved
- Obtaining extra funds is very time consuming and painstaking

19.3.4 Monitoring the Success of Adaptive Responses

The chosen adaptive response needs to be monitored. In support of this monitoring, refined or intensified performance metrics may be needed to strengthen the QA/QC system. Overall, the success of maintaining data integrity can be assessed quantitatively by a database-to-source document comparison of values of key variables (Van den Broeck et al. 2007). For questionnaire-based data, an error rate of less than 1 % is sometimes set as a criterion for successful data handling in population-based epidemiological studies. The rate can be compared between a period before and after the adaptive response. For laboratory-based results, possible recurrence of false test results or other metrics of test performance will need to be monitored intensively.

19.4 Maintaining Data Integrity in Resource-Limited Settings

Obtaining and maintaining data integrity in rural areas and developing countries is feasible and can result in acceptable data quality.

19.4.1 Challenges to Maintaining Data Integrity in Resource-Limited Settings

There are some major challenges to maintaining data integrity in resource-limited settings. For example, there may be local unavailability of experienced programmers, data managers, IT personnel, laboratory technicians, or other important study personnel involved in maintaining data integrity (Van den Broeck et al. 2007). One is often forced to engage with less experienced and less skilled personnel who need more training. Once appropriate personnel is found and trained, the actual setting up of a local IT and data system may prove to be more time consuming and expensive than anticipated. However, when there is scarcity of experienced personnel or high turnover of personnel, sometimes one has no other option than to assign certain tasks to relatively inexperienced persons who need on-the-job training to gain experience. The lack of experience might reduce the quality of the work or even inspire lack of confidence and trust with research participants, which can contribute to enrollment bias and various other types of bias.

Moreover, in remote rural areas, it may be very difficult to conduct fieldwork supervision, field staff re-training, and query handling, and these processes may depend on streamlined transport over large distances. For example, study staff may occasionally forget or loose questionnaire forms or hand in forms with considerable delays. When these situations occur in remote areas with poor transport and communication infrastructure, it is more difficult to redo the examinations and interviews. Large distances and difficult travel and working conditions tend to mean fluctuating but generally reduced efficiency of supervision, individual field staff re-training, and query handling.

Directly linked to the practical problems of doing fieldwork in poor or rural areas are potential space constraints. Space for data collection and pressure for sharing scarce equipment resources may prove to be an unexpectedly big problem in rural clinics of developing countries. Facilities may become over-crowded and over-burdened during epidemics. Available spaces can experience interruptions of electricity, Internet connectivity, water supply, and other essential utilities to make working conditions difficult.

One might also need to address unexpected cultural differences between participants and bio-medically trained staff as well as any other tensions arising between communities and research team. For example, in some cultures, the mothers' time allocation is crucial for the survival of the household. Spending a lot of time with a mother every week is only sustainable if interviewers show real empathy and respect the mother's time schedule, which usually means visiting very early in the morning before the mother leaves the homestead. Field staff may eventually have serious difficulties with the long and odd working hours, traveling, and other burdens of working in difficult circumstances, and this can contribute to increased staff turnover and sometimes budget supplements and structural re-adjustments during the implementation phase.

19.4.1.1 Challenges to Maintaining Laboratory Capacity in Resource-Limited Settings

For a variety of reasons, high turnover of personnel is a typical problem likely to affect data integrity in ongoing studies in rural areas. In addition, laboratories in resource-constrained settings also tend to lack logistic and technical support from manufacturers and providers of laboratory equipment (Crucitti et al. 2010). For example, there is often less easy access to maintenance contracts, less easy communication about technical problems, and more difficult access to training symposia. Thus, laboratory activities may need longer interruptions when problems are detected.

An additional problem may arise when small peripheral laboratories are routinely involved in testing for clinical care purposes in addition to being involved in the research study. Such a setup may well be considered the most cost-efficient, especially when laboratory analyses only concern secondary outcomes in a study. This may mean that study-related activities are only part of the activities of the laboratory personnel. The problem is that due focus on the special SOPs for the research study may dissipate more easily with this setup, especially in periods of high workload for clinical care.

19.4.2 Potential Solutions to Maintaining Data Integrity in Resource-Limited Settings

The challenges described above should not be taken to imply that data integrity cannot be maintained in resource-constrained settings. There are examples of excellent research successfully completed in such circumstances, but there are also examples of failures (Doumbo 2005). It is crucial to identify general and contextual factors that threaten data integrity in resource-constrained circumstances and to take steps towards developing practice standards specifically tailored for these circumstances. Potential solutions are obviously a matter of national and international health policy related to building research capacity in general. However, insight into the described challenges should allow individual sponsors and investigators to take a number of special measures aimed at preserving data integrity in particular studies in remote, rural, or poor settings. Panel 19.2 lists some special precautions that can be taken to protect data integrity in remote settings.

Panel 19.2 Special Precautions to Protect Data Integrity in Remote or Poor Settings

- A pool of external reserve personnel should be trained and maintained (with regular retraining as required) as a security against high turnover of personnel
- Internal personnel should be trained for tasks they will not routinely perform but for which they will serve as a reservist
- Issues of flexible and odd working hours should be carefully discussed with local personnel, and salaries should be adapted accordingly
- Transport and communication infrastructures should be adapted to the local circumstances and resistant to potential extreme situations
- For some problems, such as space constraints, structural solutions may prove difficult to implement, and flexibility may be required to find solutions on a case-by-case basis. In general, however, sponsors should be aware of the possible need to assist with infrastructural problems
- Equipment maintenance contracts should be carefully negotiated with manufacturers and providers
- Standard operating procedures should include guidelines for how to deal with problems related to unexpected resource limitations
- There should be flexibility in making sudden budget adjustments, but these adjustments need to be well documented
- Good international and national networking can aid with (re-)training, supporting research capacity, and implementing elements of a study. In international multi-center research, this can take the form of one high-quality research center serving as the coordinating center that guides and monitors other centers

(continued)

Panel 19.2 (continued)

- External monitoring plays an important role in ensuring data integrity and validity. It can provide for expertise that is not available locally to look critically at all aspects of data collection and handling
- Intensive community liaison activities and a local Community Advisory Board can be very helpful to introduce the study to local communities and authorities and to boost local acceptability of the study throughout its implementation phase
- Efficient management is especially important in resource-constrained settings, as the number and severity of crisis situations tends to be high. Fast reactions to new problems are often needed

In this chapter we discussed how to maintain the integrity of the collected data. In spite of efforts to achieve high quality data, errors do occur. Sometimes they can be corrected when discovered. The detection, evaluation, and editing of data errors is a task known as 'data cleaning,' and this is the topic of the next chapter.

References

Crucitti T et al (2010) Obtaining valid laboratory data in clinical trials conducted in resource diverse settings: lessons learned from a microbicide phase III clinical trial. PLoS One 5(19):e13592

Doumbo O (2005) It takes a village: medical research and ethics in Mali. Science 307:679–681

Ezzelle J et al (2008) Guidelines on good clinical laboratory practice: bridging operations between research and clinical research laboratories. J Pharm Biomed Anal 46:18–29

Nkengasong JN et al (2009) Critical role of developing national strategic plans as a guide to strengthen laboratory health systems in resource-poor settings. Am J Clin Pathol 131:852–857

Nkengasong JN et al (2010) Laboratory systems and services are critical in global health. Am J Clin Pathol 134:368–373

Olmsted SS et al (2010) Strengthening laboratory systems in resource-limited settings. Am J Clin Pathol 134:374–380

Van den Broeck J et al (2007) Maintaining data integrity in a rural clinical trial. Clin Trials 4:572–582

Wertheim HFL et al (2010) Laboratory capacity building in Asia for infectious disease research: experiences from the South East Asia Infectious Disease Clinical Research Network (SEAICRN). PLoS Med 7(4):e1000231

Data Cleaning 20

Jan Van den Broeck and Lars Thore Fadnes

Our greatest glory is not in never falling, but in getting up every time we do.

Confucius

Abstract

This chapter offers practical advice for investigators on how to deal with errors in collected data. In epidemiological research, as in all research, errors do occur in spite of careful study design, well-conducted groundwork, and error prevention strategies. Data cleaning is a process in which one identifies and corrects these errors, or at least minimizes their effects on study results. The present chapter describes a conceptual framework for how to set up and carry out data cleaning efforts. The framework is built on the notions that waves of data cleaning should occur at various stages of data flow (from data entry to dataset construction) and that each wave involves a screening step, a diagnostic step, and a data editing step. We then discuss study-specific aspects of data cleaning and provide advice on how to document and report on data cleaning.

20.1 Data Cleaning as a Three-Step Process

Good error prevention strategies can reduce many data problems. However, data errors can rarely be eliminated. All studies, no matter how well designed and implemented, have to deal with errors of various sources and their effects on study results. This applies to experimental research as much as it does to observational

J. Van den Broeck, M.D., Ph.D (✉) • L.T. Fadnes, M.D., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_20,

research (Ki et al. 1995; Horn et al. 2001). Data cleaning is the process one takes to deal with data problems that arise.

Little guidance is currently available in the peer-reviewed literature on how to set up and carry out data cleaning efforts. In many epidemiology circles, study validity has been discussed predominantly as an issue of study design and measurement standardization. Aspects of data handling, such as data cleaning, have received much less attention but have an equal potential to affect the quality of study results. Only certain aspects of data cleaning, such as detection of statistical outliers, have received adequate attention (*See:* Snedecor and Cochran 1980; Hoaglin et al. 1981). The *data cleaning process*, with all its conceptual and practical aspects, has not been described comprehensively. In this chapter we briefly summarize the scattered literature on this subject and integrate what is known into a conceptual framework aimed at assisting investigators with planning and implementing data cleaning (Panel 20.1; *See also*: Van den Broeck et al. 2005). This framework is built on the notions that waves of data cleaning should occur at various stages of data flow (from data entry to dataset construction) and that each wave consists of three steps: a screening step, a diagnostic step, and a data editing step.

20.1.1 The Three Steps of Data Cleaning

We propose a process of data cleaning involving three steps (Van den Broeck et al. 2005):

1. Screening of the data
2. Diagnosing likely errors
3. Editing data abnormalities

Many data errors are detected incidentally during study activities other than data cleaning. Yet, it is more efficient to detect errors by actively searching for them in a planned way. This also saves time in the analysis and writing phase, as re-analysis after correction of every data error takes substantial amounts of time.

Panel 20.1 Selected Terms and Concepts Relevant to Data Cleaning

Data abnormality Data which are deficient, excessive, outlying or inconsistent in comparison with prior expectations

Data cleaning Process of detecting, diagnosing, and editing data abnormalities

Data editing Changing the value of data shown to be incorrect or missing

Data flow Passage of recorded information through successive information carriers

Data manipulations Changes deliberately made to data values for reasons other than data editing

Inliers Data points falling within the expected range

Outliers Data points falling outside the expected range

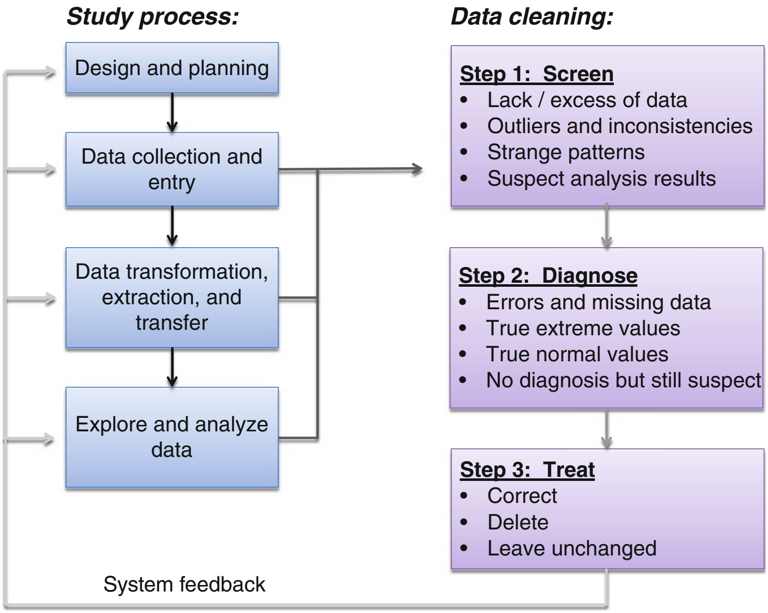


Fig. 20.1 A three-step data cleaning framework

It is not always immediately clear whether a data point is erroneous. In many cases there are no clear-cut differences between errors and true values. Usually, what is detected is a suspected data point or pattern that needs careful examination and must be assessed for the likelihood of being a true value. Similarly, missing values require further examination. They may be due to interruptions of the data flow or to the unavailability of the information. Pre-defined rules for dealing with errors and 'true' missing and extreme values are part of good practice. For these reasons, one should consider data cleaning as a systematic process of screening, diagnosing, and treating/editing data abnormalities. Figure 20.1 shows these steps, which can be initiated at three different stages of a study.

20.1.2 Sources of Errors Throughout the Data Flow

The concept of *data flow* encompasses all data management activities after a measurement is made. It begins with recording data on source forms and entry of that data into a database, and it ends at the time a cleaned dataset is analyzed. Hence, data flow involves repeated steps of data entry, transfers, extractions, selecting, editing, transformations, summarizations, and even presentations. It is important to realize that errors can and do occur at any stage of the data flow, including during data cleaning itself. Indeed, most of these problems are due to human error. One can screen for suspect features in questionnaires (Table 20.1), computer databases (Table 20.2), or analysis datasets (Table 20.3).

Table 20.1 Sources of data abnormalities in questionnaires

Lack or excess of data	Outliers and inconsistencies
Form missing	Correct value filled out in wrong box
Form collected repeatedly	Not readable
Answering box or options left blank	Writing error
More than one option selected when not allowed	Answer given is out of expected range
Errors in skip rules	Misunderstanding of question

Table 20.2 Sources of data abnormalities in the database

Lack or excess of data	Outliers and inconsistencies
Lack or excess of data from questionnaire	Outliers and inconsistencies from the questionnaire
Form or field not entered	Value incorrectly entered
Data erroneously entered twice or more	Value incorrectly changed during previous data cleaning
Value in wrong field	Transformation error
Inadvertent deletion or duplication during database handling	Use of wrong and not updated database file

Table 20.3 Sources of data abnormalities in the analysis dataset

Lack or excess of data	Outliers and inconsistencies
Lack or excess of data from database	Outliers and inconsistencies from the database
Data extraction or transfer error	Data extraction or transfer error
Deletions or duplications by analyst	Sorting errors (spreadsheets)
Use of wrong data format	Data cleaning errors
Inclusion of variable with extensive missing information in regression analysis	

Not all errors are of interest to data cleaning. Inaccuracy of a single measurement and data point might be entirely acceptable, depending on how much the degree of bias relates to the inherent technical error of the measurement instrument and compares to the range and distribution of values in the study population. Data cleaning focuses on errors that are beyond small technical variations and that constitute a major shift within or beyond the population distribution. This, in turn, points to the necessity for data cleaning to be based on some knowledge of technical errors and expected ranges of values.

In most epidemiological studies, errors that need to be cleaned at all costs include errors in outcome variables, sex, birth dates, and examination dates as well as duplications or merging of records and biological impossibilities. Sex and date errors are particularly important because they 'contaminate' derived variables. Prioritization can be of huge importance if the study is under time pressures or if resources for data cleaning are limited.

20.2 The Screening Phase of Data Cleaning

20.2.1 Types of Data Abnormalities

When screening data it is convenient to distinguish some basic types of abnormalities that one may encounter (some of which may or may not turn out to be due to errors after further examination):

- Lack or excess of data
- Outliers
- Inconsistencies
- Impossibilities
- Strange patterns
- Unexpected data formats
- Unexpected analysis results
- Other odd-looking types of inferences and abstractions

The screening method need not only be statistical. In reality, many outliers are detected by perceived non-conformity with prior expectations based on the investigator's experience, pilot studies, evidence in the literature, or just common sense. This may even happen during article review or after publication.

What can be done to make this screening objective and systematic? To allow the researcher to understand the data better, the data should first be examined with simple descriptive tools. Standard statistical packages or even spreadsheets make this easy to do. For identifying suspect data, one first pre-defines expectations about normal ranges, distribution shapes, and strengths of relationships (Bauer and Johnson 2000). Second, one applies these *pre-defined criteria* during or shortly after data collection, during data entry, and regularly thereafter. Third, one compares the data and screening criteria to *flag* dubious data, patterns, or results.

A special problem is that of *erroneous inliers*, i.e., data points generated by error but falling within the expected range. They will often escape detection. Major errors may result into values that are still within the expected data range. For example, if the true value is −1.5 Z-score (for example for growth data) and the error resulted in a value of +1.5 Z-score then the error was of the considerable magnitude of 3 Z-scores. Yet, the erroneous value is still within normal range and undetectable by a simple range check. Sometimes, inliers are discovered to be suspect if viewed in relation to other variables using scatter-plots, regression analyses, or consistency checks (Winkler 1998). One can also identify some erroneous inliers by examining the history of each data point or by re-measurement, but the latter option will rarely be feasible. Instead, one can examine and/or re-measure a sample of inliers to estimate an error rate (West and Winkler 1991).

In studies with follow-up data, it may be possible to use a more sophisticated approach than the one mentioned above. For example, one can determine changes in parameters among visits and compare those changes against a *pre-defined maximum plausible change* during that interval. If the change between the two visits exceeds the plausible change, this usually indicates that at least one of the values is erroneous. If possible, it is useful to assess both values with the ones preceding and following, to find out which of the two is most likely to be incorrect. This approach can often be programmed into a syntax file.

20.2.2 Useful Screening Methods

Useful screening methods include:

- Checking questionnaires using fixed algorithms
- Validated data entry and double data entry
- Browsing data tables after sorting
- Printouts of variables not passing range checks and of records not passing consistency checks
- Graphical explorations of distributions (e.g., box plots, histograms, scatter plots)
- Plots of repeated measurements on the same individual (e.g., growth curves)
- Frequency distributions and cross-tabulations
- Summary statistics
- Statistical outlier detection

The screening should be done after data are recorded, e.g., during supervisor checks of questionnaires, at data entry, during post-entry data cleaning, and during exploratory analyses.

20.3 The Diagnostic Phase of Data Cleaning

In this data cleaning step, the purpose is to clarify the true nature of the suspicious or implausible data points, patterns, and statistics and to identify more accurate data values whenever possible. Possible diagnoses for each suspicious data point are:

- Erroneous
- True extreme
- True normal, i.e., the prior expectation about the normal range was incorrect
- Undetermined, i.e., no explanation found but still suspect

Some data points are clearly logically or biologically impossible. Hence, one may pre-define not only screening cut-offs, as described above (*soft cut-offs*), but also cut-offs for immediate diagnosis of error (*hard cut-offs*) (Altman 1991). Figure 20.2 illustrates this method. Sometimes suspected errors will fall between soft and hard cut-offs, and diagnosis will be less straightforward. In these cases, it is necessary to apply a combination of diagnostic procedures.

One procedure is to check with the original data sources (previous stages of the data flow) to see whether a value is consistently the same. This requires accessibility of well-archived and documented data with justifications for any changes made at any stage (usually found in the data audit trail).

A second procedure is to look for information that could confirm the 'true extreme' status of an outlying data point. For example, a very low score for weight-for-age (e.g., –6 Z-scores) might be due to errors in the measurement of age or weight (which would be considered an erroneous value), or the subject may be

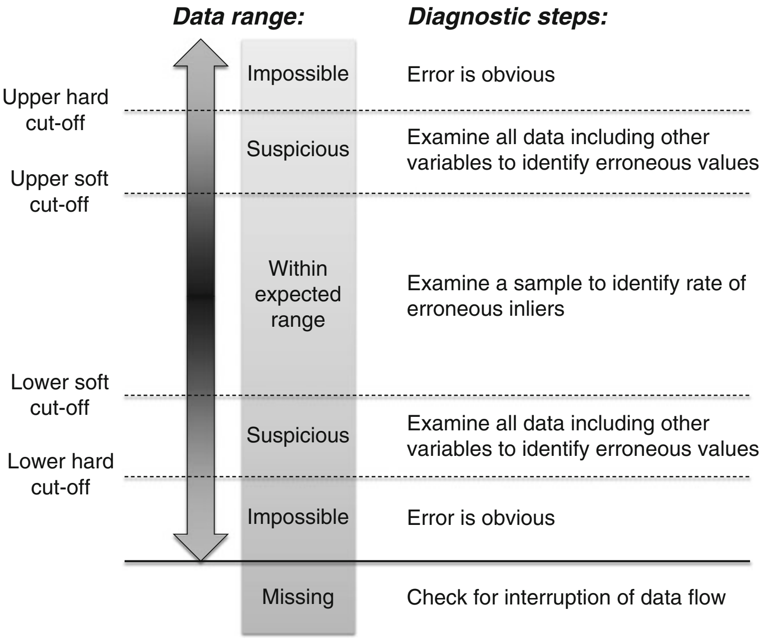

Data range:
Diagnostic steps:
Upper hard cut-off
Upper soft cut-off
Lower soft cut-off
Lower hard cut-off
Impossible
Error is obvious
Suspicious
Examine all data including other variables to identify erroneous values
Within expected range
Examine a sample to identify rate of erroneous inliers
Suspicious
Examine all data including other variables to identify erroneous values
Impossible
Error is obvious
Missing
Check for interruption of data flow

Fig. 20.2 Areas within the range of a continuous variable defined by hard- and soft cut-offs for error screening and –diagnosis, with recommended diagnostic steps for data points falling in each area

extremely malnourished, in which case other nutritional status variables should also have extreme values (supporting the notion that the extreme value is true). Individual patients' reports and aggregated information on related issues are helpful for this purpose. This type of procedure requires insight into the coherence of variables in a biological or statistical sense.

A third procedure is to collect *additional information* that supports making a decision about the diagnosis of a putative data error. One should be prepared to discuss with the interviewer/measurer what may have happened and, if possible, to repeat the measurement. This is a strong argument in favor of starting data cleaning as early as possible after data collection. Sometimes, re-measuring is only valuable very shortly after the initial measurement. In longitudinal studies, variables are often measured at specific ages or follow-up times. With such designs, the possibility of re-measuring or obtaining measurement values for missing data will often be limited to predefined allowable intervals around the target times. Such intervals can be set wider if the analysis foresees using age or follow-up time as a continuous variable.

Finding an acceptable value does not always depend on re-measurements, though. For some input errors, the correct value is immediately obvious. Examples include:

- Missing values in the database could be due to a data entry omission and the correct value would then be readily available on the source document

- Swapped values (e.g.,'infant length' was mistakenly noted under 'infant weight' and vice versa). In this case, a correction can only be made if the investigator's knowledge of the subject matter is adequate; therefore, this type of data error is usually only caught by highly trained members of the research team.

During the diagnostic phase, one may have to reconsider prior expectations and/or review QA/QC procedures. This phase is labor intensive and the budgetary, logistical, and personnel requirements are typically underestimated or even neglected at the study planning stage. How much effort must be spent? There is no good answer to this question yet, and the field is in desperate need of cost-effectiveness studies to answer this question. Experience tells us, however, that costs are likely to be lower if the data cleaning process is planned and starts early in data collection. Supporting tools during the diagnostic phase – including automated query generation and automated comparisons of successive datasets – can also help to reduce costs of data cleaning.

20.4 Data Point Editing and System Feedback

After identifying errors, missing values, and true (extreme or normal) values, the researcher must decide what to do with problematic data. Editing options are limited to:

- Correction of the data
- Deletion of the data (setting the value to 'missing')
- Leaving the data unchanged

There are some general rules in selecting which editing option is most appropriate:

1. Impossible values are never left unchanged. They should be corrected if a correct value can be found; else they should be deleted (usually set to missing).
2. For biological continuous variables, some within-subject variation and small measurement error affects every measurement. If a re-measurement is done very rapidly after the initial one and the two values are close enough to be explained by these small variations alone, accuracy may be enhanced by taking the average of both as the final value.
3. What to do with true extreme values, and with values that are still suspect after the diagnostic phase? The investigator may wish to further examine the influence of such data points individually and as a group on analysis results before deciding whether or not to leave the data unchanged. Statistical methods exist to help evaluate the influence of such data points on regression parameters. If extreme values are left unchanged, one can consider applying 'robust estimation' (e.g., robust regression) procedures in the statistical analysis, in order to minimize the effect of the remaining outliers on study findings
4. Some authors have recommended that true extreme values should never be deleted (Gardner and Altman 1994). In practice, exceptions are frequently made to that rule. The investigator may not want to consider the effect of true extreme values if they result from an unanticipated extraneous process. This becomes an *a posteriori* exclusion criterion and the data points should then be reported as 'excluded from the analysis.'

5. For values that need to remain missing or need to be set to 'missing,' there may be an option of imputation during analysis that allows for correction.

Data cleaning often leads to insight into the nature and severity of error-generating processes. The researcher can then give methodological feedback to improve study validity and precision of outcomes (*See:* Chap. 19). It may be necessary to amend the study protocol regarding design, tim ing, observer training, data collection, and quality control procedures. In extreme cases, it may be necessary to restart the study, including planning and data collection. Programming of data capture, data transformations, and data extractions may need to be revised and the analysis strategy adapted to include robust estimation or separate analyses with or without remaining outliers or imputation.

20.5 Study-Specific Aspects of Data Cleaning

The study objectives co-determine the required precision of the outcome measures, the error rate that is acceptable, and therefore the necessary investment in data cleaning. The sensitivity of the chosen statistical analysis method to outlying and missing values can also have consequences in terms of the amount of effort the investigator wants to invest in data cleaning. As an example, a study using median values when reporting their outcomes are less prone to extreme values than a similar study where the mean is used.

In clinical trials, there may be concerns about investigator bias resulting from the close data inspections that occur during data cleaning. In these studies, examination by an independent expert may be preferable. In intervention studies with interim evaluations of safety and/or efficacy, it is of particular importance to have reliable data available before the evaluations take place. This is another strong argument for the need to initiate and maintain an effective data cleaning process from the early start of a study.

Longitudinal studies differ in that checking the temporal consistency of data is essential. Plots of serial individual data such as growth data or repeated measurements of categorical variables often show a recognizable pattern from which a discordant data point clearly stands out.

In small studies, a single outlier will have a greater distorting effect on the results. Some screening methods such as eyeballing of data tables will be more effective, whereas others, such as statistical outlier detection, may become less valid with smaller samples. The volume of data will be smaller, hence the diagnostic phase can be cheaper and the whole procedure more complete. Smaller studies usually involve less people and the steps in the data flow may be fewer and more straightforward, allowing fewer opportunities for errors. These considerations should help in choosing the appropriate data cleaning tools listed above.

20.6 Documentation and Reporting of Data Cleaning

Statistical societies recommend that data cleaning be reported in the statistical methods as a standard in research articles (American Statistical Association 1999). What exactly to report under the various circumstances has remains mostly unanswered.

It is still rare to find any statements about data cleaning methods or error rates in study protocols or medical publications.

We recommend including a data-cleaning plan in study protocols. This should include budget and personnel requirements, prior expectations used to screen for suspect data, screening tools, diagnostic procedures used to discern errors from true values, and the decision rules to apply in the editing phase. Proper documentation should exist for each data point (e.g., a syntax file), including differential flagging of types of suspected features, 'diagnostic' information, and information on type of editing, dates, and personnel involved.

In large studies, data monitoring and safety committees should receive detailed reports on data cleaning, and procedural feedbacks on study design and conduct should be submitted to study steering and ethics committees. We recommend that medical scientific reports include a description of the data cleaning in the methods section. This should include error types and rates at least for the primary outcome variables, with the associated deletion and correction rates, justification for imputations when done, and differences in outcome with and without remaining outliers.

So far in Part III we have looked at all major phases and aspects of gathering and processing empirical data, from study preparations to data cleaning. The outlook has been to maximize the quality of data available for analysis and to ensure that the rights and safety of participants are preserved. For clinical trials, formal guidelines and regulations exist that aim exactly for the same goals of data quality and ethical research conduct. These guidelines are called 'Good Clinical Practice' guidelines and they are briefly introduced in the next chapter.

References

Altman DG (1991) Practical statistics in medical research. Chapman and Hall, London, pp 1–611. ISBN 0412276305

American Statistical Association (1999) Ethical guidelines for statistical practice. http://www.amstat.org/about/ethicalguidelines.cfm. Accessed Sept 2012

Bauer UE, Johnson TM (2000) What difference do consistency checks make? Am J Epidemiol 151:921–925

Gardner MJ, Altman DG (1994) Outliers. In: Statistics with confidence. BMJ Books, London, pp 1–90. ISBN 0727902229

Hoaglin DC, Iglewicz B, Tukey JW (1981) Small-sample performance of a resistant rule for outlier detection. In: Proceedings of the statistical computing section. American Statistical Association, Washington, DC, pp 144–52

Horn PS et al (2001) Effect of outliers and non-healthy individuals on reference interval estimation. Clin Chem 47:2137–2145

Ki FY et al (1995) The impact of outlying subjects on decision of bio-equivalence. J Biopharm Stat 5:71–94

Snedecor GW, Cochran WG (1980) Statistical methods, 7th edn. The Iowa State University Press, Ames, pp 1–507. ISBN 0813815606

Van den Broeck J et al (2005) Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med 2(10):e267

West M, Winkler RL (1991) Data base error trapping and prediction. J Am Stat Assoc 86(416):987–996

Winkler WE (1998) Problems with inliers. US Census Bureau, Research Reports Series RR98/05

Good Clinical Practice

21

Jan Van den Broeck, Vundli Ramokolo, and Jutta Dierkes

> *Good Clinical Practice is a standard for the design, conduct, performance, monitoring, auditing, recording, analyses, and reporting of clinical trials that provides assurance that the data and reported results are credible and accurate, and that the rights, integrity, and confidentiality of trial subjects are protected.*
>
> EMA, 2002

Abstract

Good Clinical Practice (GCP) is a set of guidelines for trial research, not for the practice of clinical care, as the name might suggest. This chapter aims to introduce the large topic of GCP, to orient those researchers who are unfamiliar with trial research to the essence and scope of GCP guidelines, and to discuss some practical GCP-related tasks. First, the concept of GCP as a standard rooted in general ethical principles and as a new paradigm in experimental research involving human subjects is explained. Next we review the wide scope of GCP-related responsibilities of investigators and discuss the resources required to establish minimum GCP capacity. This leads us to the topic of the relevance of GCP for observational research and implementation difficulties in

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

V. Ramokolo, M.Sc.
South African Medical Research Council, Cape Town, South Africa

J. Dierkes, Ph.D.
Faculty of Medicine and Dentistry, Department of Clinical Medicine,
University of Bergen, Bergen, Norway

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_21,

resource-poor areas. Finally, we introduce in some more detail three selected GCP-related activities that are of particular practical importance during execution of a trial: the maintenance of a regulatory file, adverse events reporting, and site monitoring visits. Basic terminology is listed in Panel 21.1

21.1 GCP as an Ethical Standard and a New Paradigm in Research

Good Clinical Practice (GCP) guidelines define responsibilities of trial investigators in the preparation of a trial protocol, a document that is to be reviewed and approved by the regulatory authorities before the commencement of the trial. The guidelines also prescribe minimum standards for procedures used by the investigator during the implementation of the trial and procedures after trial completion. Thus, the principal investigator is responsible for the design and conduct of a clinical trial and for the reporting of the findings. This wide scope of responsibilities is described in more detail in the next section. International GCP guidelines exist, but some countries, sponsors, and institutions use their own adaptations. Currently, the most commonly followed international GCP guidelines are the ICH-6 Guidelines (International Conference on Harmonization 6, www.ich.org) and the World Health Organization GCP Guidelines. In fact, GCP compliance is legally regulated in many countries. The European GCP guidelines can be found at the webpage of the European Medicines Agency (EMA, www.ema.europe.eu, Directive2005/28/EC).

Panel 21.1 Selected Terms and Concepts Relating to Compliance with Good Clinical Practice Guidelines

Adverse event Untoward health-related event in a trial participant (which does not necessarily have a causal relationship with the test product or intervention)

Adverse events report Investigator report on adverse events and serious adverse events given to sponsor, ethics committee and/or regulatory authorities

Clinical monitor Person designated by the trial sponsor to check if the actual trial procedures conform with study protocol, standard operating procedures and GCP guidelines

Clinical report form (CRF) A document designed to record all of the protocol required information to be reported to the sponsor on each trial participant

Clinical trial An intervention study of the pharmacological effects or dynamics and/or the efficacy and/or the safety of a test product or treatment

(continued)

Panel 21.1 (continued)

Compliance Implementation according to protocol or guideline

Good Clinical Practice guidelines A standard for all stages of conduct of clinical trials aimed at (1) optimizing validity and credibility of data and results, and (2) ensuring that the rights, integrity, and confidentiality of data of trial subjects are protected

Monitoring The act of overseeing the progress of a clinical trial, and of ensuring that it is conducted, recorded and reported in accordance with the protocol, SOPs, GCP guidelines and regulatory requirements.

Participant Individual who has officially consented to participate in a study and has not withdrawn from participation

Safety monitoring System or practice of detecting, characterizing following-up foreseen and unforeseen problems with participant safety during a research study

Serious adverse event Adverse event that leads to death, (prolongation of) hospitalization, disability or birth defect, or any condition that is immediately life-threatening

Sponsor An individual, company, institution, or organization that takes responsibility for the initiation, management, and/or financing of a study

Standard Operating Procedures (**SOP**) Detailed written instructions to achieve uniformity of the performance of a specific function

21.1.1 GCP as an Ethical Standard

GCP guidelines are rooted in ethical principles. For example, 13 ICH-6 GCP principles have been formulated. The first asserts that ‘Clinical trials should be conducted in accordance with the ethical principles that have their origin in the Declaration of Helsinki (World Medical Association), and that are consistent with GCP and the applicable regulatory requirement(s).’ This refers to the four *prima facie* principles of biomedical ethics, which are respect for autonomy, justice, beneficence, and non-maleficence and also underlie the principles of epidemiology proposed in Chap. 1. More information on these four basic ethical principles can be found in Beauchamps and Childress (2001). In ICH-6 the remaining twelve GCP principles are also derived from these *prima facie* principles. Three examples of how these ethical principles serve as the basis of ICH-6 GCP principles are illustrated in Table 21.1.

21.1.2 GCP as a Paradigm Shift in Experimental Research

The advent of GCP guidelines and compliance requirements has been the result of an important shift in thinking about study validity and other ethical aspects of clinical trial research. This shift is illustrated in Table 21.2. Essentially, the shift is

Table 21.1 Illustration of the ethical basis of selected ICH-6 GCP principles

Ethical principle	ICH-6 GCP principle
Respect for autonomy	'Freely given informed consent should be obtained from every participant prior to clinical trial participation'
Beneficence	'Before a trial is initiated, foreseeable risk and inconveniences should be weighed against the anticipated benefit for the individual trial participant and society. A trial should be initiated and continued if the anticipated benefits justify the risk'
Non-maleficence	'The rights, safety and well being of the trial participants are the most important considerations and should prevail over interest of science and society'

Table 21.2 Good clinical practice as a paradigm shift

Old paradigm	New paradigm
Validity and ethical value of clinical trial research mainly depend on:	Validity and ethical value of clinical trial research mainly depend on:
1. Study design/protocol	1. Study design/protocol
2. An honest and dedicated investigator *implementing the protocol* **to her/his best abilities**	2. Honest and dedicated investigator and sponsor **complying with GCP guidelines**
3. Peer review	3. Peer review

to be seen as a necessary step in the greater involvement of the scientific community and of society at large as stakeholders in research. This shift fits into the wider shift in thinking that has occurred after World War II and that has led to the universal requirement for ethical oversight and informed consent: It is a partial move away from a 'blind' trust in individual researchers' abilities to design and carry out high quality research towards greater oversight from other stakeholders. This greater oversight has also led to regulations and laws about GCP compliance. GCP compliance has become a major responsibility of sponsors as well.

During trials, site visits are made by the sponsor to assess study status and verify compliance with the protocol and GCP guidelines. Sponsors often delegate these visits to institutions specialized in clinical research monitoring. One or more (clinical) monitors, also known as Clinical Research Associates, visit the site(s). Monitors review the regulatory file and sometimes also the infrastructure and logistics with the assistance of the study coordinator and investigator(s).

21.2 The Scope of GCP-Related Responsibilities of Investigators

GCP guidelines concern any study-related activity during which accuracy of evidence or respect for participants could be compromised. The guidelines therefore define responsibilities concerning:

- Study design
- Practical implementation

Table 21.3 Scope of GCP-related responsibilities of investigators

	Prescribed practice	Reinforcement	Measurement of actual practice	Reporting of actual practice
Securing resources	Sufficient budget	Budget justification, grant approval	Accountability	Budget status reports
Establishing framework	Legal requirements, acceptability by stakeholders	Legal authority's and stakeholders' approval	Check new legal requirements, check acceptability repeatedly	Feedback to stakeholders
Study design and enrolment	Study protocol incl. Informed consent form and -procedures	Ethics committee approvals study steering committee	Track violations, keep enrolment and follow-up statistics, protocol amendments	Protocol history & violation reports, cohort status reports
Data collection	Fieldwork SOP[a]	(Re-)training, supervision, operations management	Quality control forms, audits	Quality control reports
Data handling	Data handling protocol and SOPs	Data management, IT management and data cleaning	Audit trails, database backups	Data management reports
Intervention delivery	Investigator brochure, Intervention delivery SOP, drug management plan	User info and demos, pharmacist or drug manager	Drug accountability; pharmacy temperature monitoring	Drug accountability reports
Intervention monitoring	Ethics plan, data monitoring plan	DSMB, ethics committee, (S)AE[a] monitoring	Interim analyses, SAE forms, AE forms	DSMB[a] reports, (S)AE reports
Analysis	Analysis plan, publication policy	Analysis/writing committee	Analysis syntax, analysis datasets	Scientific articles

[a] *SOP* standard operating procedures, *(S)AE* (serious) adverse events, *DSMB* data and safety monitoring board

- Measurement of actual practice
- Reporting of actual practice

The very wide scope of GCP-related responsibilities, as viewed from this angle, is illustrated in Table 21.3.

21.2.1 Scope of Training Needs and Regulatory Requirements

Compliance with GCP guidelines requires the necessary resources and training of study personnel. From the side of investigators and study managers, it requires a good knowledge of the specific GCP guidance document(s) and regulatory framework that will be adhered to in the trial. As indicated earlier, this chapter merely provides a brief orientation and some discussion on selected aspects. It is the GCP guidelines themselves that provide the most important resource for learning. A disclaimer is in place here to state that none of the content discussed in the chapter should be taken to represent the actual regulatory requirements relevant to any particular trial.

The sponsor of the trial will normally have an important say about which specific documents need to be followed. A research sponsor is an individual, company, institution, or organization taking responsibility for the initiation, management and/or financing of a research study. Situations can arise where several GCP guidelines need to be followed. For example, a clinical trial in South Africa, if sponsored by the United States National Institutes of Health (NIH), would need to comply with the South African GCP guidelines (Guidelines for Good Practice in the Conduct of Clinical Trials with Human Participants in South Africa, Department of Health 2006) as well as with those prescribed by NIH.

21.2.2 Relevance of GCP for Non-trials

By definition (EMA, 2002; *See:* opening quote of this chapter), GCP provides practice standards aiming to assure that reported results are credible and accurate, and that the rights, integrity, and confidentiality of trial subjects are protected. These concerns about validity and ethical value are universal concerns about all research with humans, not only clinical trials. They relate to all intervention research as well as to observational research. Whereas the concerns are universal, part of the GCP standards are clearly only relevant to trials, such as those parts that relate to test products, treatments, and their side effects. Many other guidelines, however, are directly applicable to all types of studies. The impetus to employ strict GCP-level oversight is commonly felt to be less strong in the case of observational studies because these studies often carry less risk for serious negative outcomes. Yet, GCP-level oversight is increasing in observational studies mainly because of a felt need to maximize the validity of observational research so that evidence from this type of studies gains greater credibility. In practice, observational research investigators do actually implement some of the GCP guidelines. Many are using GCP guidelines as a resource to improve study design and quality assurance strategy. Viewed from another angle, the wide scope of GCP guidelines for trials importantly overlaps with the practical guidelines in this book (*See:* Chap. 1) and with the guidelines for epidemiological studies from the Council for International Organizations of Medical Sciences (CIOMS 2009). The general principles of epidemiology presented here are rooted, like the GCP principles for trials, in *prima facie* ethical principles.

21.3 GCP Capacity

21.3.1 Developing Local GCP Capacity

Researchers and institutions embarking on trial research are faced with the challenge of developing GCP capacity. To develop this capacity, it is advisable that all researchers and managers who will be involved in the trials are trained and certified in human subject protection. Accredited online GCP training courses are available, such as the free online course provided by the National Institutes of Health in 2012 (NIH 2012). Some academic institutions also run GCP training programs. For example, the South African Tuberculosis Vaccine Initiative of the University of Cape Town offered introductory and refresher GCP training courses in 2012 (SATVI 2012). Training should cover adverse events, informed consent procedures, and source document standards for all field staff, including study physicians. Capacity building must, however, also involve trial coordinators, laboratory technicians, data managers, statisticians, and monitors.

21.3.1.1 Difficulties with GCP Compliance in Resource-Poor Settings

GCP capacity development is mainly an institutional responsibility and is difficult but feasible in resource-poor settings**.** GCP-related study activities need to be properly budgeted by the investigator and sponsor, carefully considering the special requirements in resource-poor settings (Acosta et al. 2007; Osrin et al. 2009). Difficulties to overcome may include:

- Poor accessibility of study participants for obtaining information on serious adverse events
- Lack of detailed medical information on adverse events; poor or absent hospital records
- Sub-optimal infrastructure and logistics for providing medical care; divergent opinions of various oversight committees on what exactly the level of care should be for the participants during and/or after the trial
- Validity of traditional serious adverse event parameters (hospitalization, death) may be reduced by problems of health care accessibility
- Difficulty finding experienced personnel with or without GCP training; increased training needs
- Difficulty building oversight committees using local resources; need for understanding the local situation when hiring people located far from the study site
- Communication problems with oversight committees may occur in case of prolonged loss of Internet connectivity
- Language barriers

21.3.2 GCP Compliance Budgeting

Costs related to GCP compliance are often underestimated at the design and budgeting stage of a clinical trial. Costs are higher if infrastructure is lacking or if specialized manpower is difficult to recruit or train, which is more often the case in resource-poor settings. Budget items should cover at minimum the following:

- Costs of ethical review
- Costs of data and safety monitoring board functioning
- Costs of data handling protocol implementation
- Costs of standard operating procedures development and training
- Costs of quality assurance and control procedures

- Costs of adverse events training and reporting
- Costs of site-monitoring/auditing
- Costs of infrastructure and manpower

21.4 The Regulatory File

Essential documents of a trial need to be compiled in a so-called regulatory file. According to ICH-6, Section 8, "Essential documents are those documents that individually and collectively permit evaluation of the conduct of a trial and the quality of the data produced". The regulatory file must be established at the beginning of each study and updated throughout the life of the study. It must be made available for inspection during site monitoring visits. After the study it must still be kept for some time. For example, in the United States of America, the regulatory file for drug approval trials needs to be kept for at least 2 years after Food and Drug Administration (FDA) approval or 2 years after the stop of drug development. Panel 21.2 contains a minimum list of essential documents based on the second edition South African Good Clinical Practice guidelines (Department of Health 2006). The list is divided into three sections in accordance with the stage in which the documents are usually generated in the trial. This list aims to illustrate that the regulatory file is comprehensive and detailed. This, in turn, should again illustrate the wide scope of GCP-related responsibilities of investigators and draw attention to the considerable resource requirements (including time investment) of GCP compliance.

Panel 21.2 Minimum Content of a Regulatory File; Example Based on the South African GCP Guidelines, with Minor Adaptations (Abbreviations Explained in Footnote)

Section A:
Documents generated before the formal commencement of the clinical trial:

1. Investigators' brochure
2. Signed protocol and amendments (if any) and sample CRF
3. Information given to participant (including ICF in all relevant languages)
4. Documents detailing financial aspects of trial
5. Insurance statement (if applicable)
6. Signed statement between parties involved in trial
7. Documented approval by IRB/IEC of the following: protocol and amendments, CRF (if applicable), ICFs, any written information given to participant (e.g. information sheets) and participant compensation (if applicable)

(continued)

Panel 21.2 (continued)

8. Outline of IRB/IEC composition
9. Regulatory authority’s approval of protocol (if required)
10. Documents with investigator and sub-investigator qualifications (including Curriculum Vitae)
11. Document with normal values and/or ranges for technical-medical procedures
12. Medical technical procedure certification or accreditation or QA/QC assessment methods
13. Sample of labeling on test product
14. Handling instructions for test product and other trial materials (if omitted in protocol or investigators brochure)
15. Shipping records for test product and other trial materials
16. Certificate(s) of analysis for shipped test products
17. Decoding procedures for blinded trials
18. Master randomization list
19. Pre-trial monitoring report
20. Trial initialization monitoring report

Section B:

Documents to be added to the regulatory file during the trial:

1. Updates on investigator’s brochure
2. Any revisions to protocol/amendment(s), CRF, ICF or any other information that is provided to participants
3. Documented IRB/IEC approval of protocol amendment(s) and revision(s) of the following: ICFs, approved documents and any other written information given to participants. If applicable, regulatory authority approvals for protocol amendment(s) and other documents
4. Investigator and/or sub-investigator qualifications
5. Updates to normal values and/or ranges for technical procedures
6. Updates to technical procedures certification or accreditation or QA/QC assessment methods
7. Documentation of trial article and shipment of trial-related materials
8. Certificate(s) of analysis for new batches of trial article
9. Site visit monitoring reports
10. Records of communication other than site visits (e.g., meeting minutes, emails)
11. Signed ICFs
12. Source documents
13. Signed, dated and completed CRFs
14. Documentation of CRF amendments
15. Investigator serious adverse event notifications (and related reports) to sponsor

(continued)

Panel 21.2 (continued)

16. Sponsor and/or investigator unexpected serious adverse drug reaction notifications to regulatory authorities and IRB/IEC
17. Safety information provided to the investigator by sponsor
18. Interim or annual reports provided to IRB/IEC or regulatory authorities
19. Participant screening log
20. Participant identification code list
21. Participant enrolment log
22. Investigational products accountability at study site
23. Signature sheet
24. Record of retained body fluid or tissue samples (if applicable)

Section C:
Documents to be added to the regulatory file after completion or termination of the trial:

1. Test product(s) accountability at site
2. Documentation of test product disposition
3. Completed participant identification code list
4. Audit certificate (if applicable)
5. Final trial close-out monitoring report
6. Treatment allocation and decoding documentation
7. Final report by investigator to IRB/IEC, where applicable, to regulatory authorities
8. Clinical study report

Abbreviations: CRF = Case Record Form; ICF = Informed Consent Form; IRB/IEC = Institutional Review Board/Independent Ethics Committee; QA/QC = Quality Assurance/Quality Control

21.5 Adverse Events Reporting

GCP guidelines around participant safety comprise, among others, the requirement for a qualified physician (or dentist, as appropriate) to give medical care to participants, and, as a main responsibility of this study physician, the timely detection, assessment, and reporting of adverse events in trial participants. Typically, adverse event reports are to be sent to the sponsor, ethics committee(s), and data and safety monitoring board within 24 h of awareness of the event. The purposes of adverse events reporting are:

- To maximize individual participation safety
- To allow methodical evaluation of clinical safety data for study participants both individually and as a group
- To help in developing accurate drug toxicity profiles

Panel 21.3 Adverse Events and Serious Adverse Events: An Example of How the Distinction Can Be Made. Exact Distinctions to Be Made During a Trial Depend on Sponsor or Oversight Body

Any adverse event – An adverse event is any untoward medical occurrence in a study participant who has received a test article or intervention that may or may not have a causal relationship with this test-treatment. It can be any unfavorable or unintended sign, including:

- Abnormal laboratory findings
- Symptoms or diseases
- Worsening of a baseline condition
- Protocol-defined events

Serious adverse event – This is any adverse event occurring at any intervention level that results in any of the following outcomes:

- Death
- Immediately life threatening event
- Persistent or significant disability or incapacity
- Hospitalization or prolongation of hospitalization
- Congenital anomaly or birth defect
- Conditions considered so serious that they required medical or surgical intervention to prevent one of the above outcomes

21.5.1 Adverse Events and Serious Adverse Events

Currently used adverse events reporting systems typically make a distinction between adverse events and *serious* adverse events. Definitions vary slightly according to sponsor or oversight body, but a typical distinction could be as described in Panel 21.3.

21.5.2 Content of an Adverse Events Report

Administratively, the following could be a list of officially required content of an adverse event report in a clinical trial:

1. Name of the event.
   An adverse event can be named as a medical diagnosis, or, if a diagnosis is not available, as signs and symptoms. The name of a procedure such as a surgical operation is not eligible as a name of an adverse event.
2. Start date/stop date/time.
3. Treatment given.

4. Intensity.
   Sponsors and oversight bodies can provide scales for grading adverse events. For example:
   Grade 1 = Mild
   Grade 2 = Moderate
   Grade 3 = Severe
   Grade 4 = Life threatening
   Grade 5 = Death
5. Relationship to a study drug and action taken with a study drug.
   Categories describing the likelihood of a relationship with a study drug are assigned by the study physician in agreement with the sponsor and ethics committee. An example of an officially required categorization is:
   Impossible
   Unlikely
   Possible
   Certain
6. Outcome, resolution.
   All serious adverse events should be followed intensively. At the time of reporting the following categories might be considered applicable and in need of inclusion in the report:
   Event has stabilized
   Condition returned to normal
   Condition resolved
   Condition no longer meets serious adverse event criteria
7. Was the event expected?

21.6 Site Monitoring Visits

The monitoring process is defined as oversight of the progress of a clinical trial, and of ensuring that it is conducted, recorded, and reported in accordance with the protocol, SOPs, GCP, and applicable regulatory requirements. In this respect, it is one of the fundamental aspects of GCP adherence. Monitoring takes place before a trial starts, during the trial period, and after the trial. It is essential that monitors have the appropriate training and knowledge that is needed to monitor the trial and be familiar with the protocol and other procedures of the trial, but not be directly involved in the study.

Before the trial, the monitor has to verify that:

- Relevant authorizations and documents (e.g., written informed consent) are present
- The study site has the capacity to include the anticipated number of participants in the study
- The study is able to fulfill planned procedures

- The investigational products are available and appropriately stored

During the trial, site monitoring visits will ensure that:

- The study is being performed according to the protocol
- The study is in compliance with GCP and other applicable regulatory requirements
- Every patient signed the written informed consent form and received the planned procedures
- Subject enrollment rates are adequate
- Data are correctly and completely recorded into the clinical report forms (CRFs) and verifiable from the source documents
- Adverse events are handled correctly

After the trial, the monitor ensures that:

- Investigational products are accounted for
- Trial documents are stored properly
- The trial database is properly maintained and secured

Monitoring is an integral part of the quality control of a clinical trial and is designed to verify the quality of the trial. It can also detect gross deviation from the protocol at single sites of large multi-center trials by the use of central monitoring and statistical procedures. These include e.g., check for invalid data, comparison of repeated measurements, calendar check, control of digit preferences or comparison with external sources. These methods can reveal deviations that can simply be due to misunderstandings, but also due to falsification of data. In addition to central monitoring, usually, on-site monitoring before, during, and after the trial is necessary.

21.7 Advantages and Disadvantages of the Implementation of GCP as a Standard

ICH-6 GCP is now the de facto global standard by which clinical trials are run and has been legally implemented in many places, including the European Union, USA, and Japan. Following GCP guidelines increases the need for standardization, training, reporting, and training; therefore, these guidelines are sometimes difficult to implement. In addition, GCP guidelines increase costs and bureaucracy, and they can even make a trial impossible, especially in resource-limited settings. The need for capacity building at all levels (trial coordinators, technicians, data managers, statisticians, monitors, research nurses, and investigators) has already been mentioned. Indeed, many (European) universities now have clinical trial offices that support their academics in conducting clinical trials. In addition, it has been criticized that GCP guidelines have mainly been developed for clinical trials that test new drugs or drug treatments. However, there is a substantial need for trials that investigate disease management, or other investigations that are not new product trials (e.g., exercise intervention trials). The applicability of GCP guidelines is less straightforward in these types of trials. Furthermore, it has to be stated that some core issues of clinical trial design are not appropriately covered by GCP guidelines, e.g., randomization. Since this is a critical issue for the quality of clinical trials, a further development of GCP guidelines seems to be necessary.

However, it is clear that the introduction of common guidelines has increased the standards and quality of clinical trials substantially, along with increased respect for patients' integrity. Further training, experience, and critical thinking will enhance the quality, comparability, and standards of clinical trials in a sustainable way.

This chapter was the last of a series of chapters on study planning and implementation (Part III). The main theme across the series was how to validly, ethically, and efficiently obtain data for analysis. Logically, this brings us to Part IV: Study analysis.

References

Acosta CJ et al (2007) Implementation of good clinical practice guidelines in vaccine trials in developing countries. Vaccine 25:2852–2857

Beauchamps TL, Childress JF (2001) Principles of biomedical ethics, 5th edn. Oxford University Press, Oxford, pp 1–454. ISBN 0195143329

Council for International Organizations of Medical Sciences (2009) International ethical guidelines for epidemiological studies. CIOMS, Geneva, pp 1–128. ISBN 929036081X

Department of Health, Republic of South Africa (2006) Guidelines for good practice in the conduct of clinical trials with human participants in South Africa. Department of Health, Pretoria

EMA/European Medicines Agency. Good clinical practice guidelines. www.ema.europe.eu. Directive 2005/28/EC. Accessed Feb 2013

National Institutes of Health Office for Extramural Research. Protecting human research participants. http://phrp.nihtraining.com/users/login.php. Accessed Sept 2012

Osrin D et al (2009) Ethical challenges in cluster randomized controlled trials: experiences from public health interventions in Africa and Asia. Bull WHO 87:772–779

SATVI. Good clinical practice training courses. http://www.satvi.uct.ac.za/index.php/20101214261/Education-and-training.html. Accessed Sept 2012

World Medical Association. The declaration of Helsinki. http://www.wma.net/en/10home/index.html. Accessed Sept 2012

Part IV

Study Analysis

Statistical Estimation

22

Jan Van den Broeck, Jonathan R. Brestoff,
and Catherine Kaulfuss

Often, even the most accurate data are only used to make educated guesses.

Anonymous

Abstract

This first chapter on study analysis covers statistical estimation methods that are frequently made use of in epidemiological research. It does so at an introductory level only. In the first section we explain important concepts related to statistical estimation: we distinguish estimators from estimates, and we contrast (1) point vs. interval estimates and (2) crude vs. adjusted estimates. In the ensuing sections, we discuss the estimation of outcome frequency in a single group and descriptive comparisons of outcome frequencies in multiple groups. Standardization of estimates falls within that context but is dealt with in a separate section. Finally, estimation of outcome parameters in analytical studies is covered, with special attention to strategies for controlling confounding during analysis.

J. Van den Broeck, M.D., Ph.D. (✉) • C. Kaulfuss, M.Sc.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_22,

22.1 Concepts of Statistical Estimation

22.1.1 Estimators and Estimates

Estimators are a type of statistic designed to capture a frequency or a contrast of frequencies (a causal or acausal statistical relation; Panel 22.1). Among estimators, epidemiologists classically distinguish three classes:

- *Measures of frequency*, e.g., odds, prevalence, incidence risk, or incidence rate
- *Measures of association*, e.g., odds ratio, prevalence rate ratio, relative risk, or incidence rate ratio
- *Measures of causal effect*, e.g., adjusted odds ratio, adjusted prevalence rate ratio, adjusted relative risk or adjusted incidence rate ratio (the adjustments made for the purpose of control of confounding); a general term for all these examples is *causal rate ratios*

Important is also that the last two classes (measures of association and measures of causal effect) are often jointly referred to as measures of effect or *effect measures*.

Estimates are particular values calculated for an estimator based on the study data. They are called estimates only because their purpose is to estimate the true value of the estimator in the target population, not to measure it with absolute accuracy. The need to make such estimates is due to the inability to examine an entire target population. Instead, one normally resorts to examining a statistical or non-statistical sample and uses values derived from the sample only to *estimate* what the true underlying values might be for the entire target population.

Panel 22.1 Selected Terms and Concepts Relating to Statistical Estimation

Adjusted estimate An estimate that is adjusted to correct for confounding, bias and/or to make it independent of the modifying influence of other variables

Crude estimate An estimate that is unadjusted for confounding, bias or modification by other variables

Confidence interval A range of values that is likely to include the true value for the population parameter of interest

Estimate *See:* outcome parameter estimate

Estimator A statistic which is a measure of frequency, association or causal effect

Imprecision Lack of total precision (*Miettinen* 1985) (*See:* precision)

Interval estimate See: confidence interval

Measure of association A statistic expressing a degree of association

Measure of causal effect A statistic expressing the size of a causal (confounding-free) effect

(continued)

Panel 22.1 (continued)

Measure of frequency A statistic expressing the frequency of occurrence of an event

Outcome parameter (in statistical estimation) Estimator

Outcome parameter estimate Particular value for an estimator calculated on the basis of the study data

Point estimate A 'best guess' estimate of the population value of an estimator, inferred from the sample

Percentage Proportion times 100

Precision (- of an estimate) Degree of sampling *and* random measurement error that influenced the estimate

Proportion number with the characteristic/event of interest, divided by the number examined who could have had the characteristic/event of interest

Sampling variation Variation in the distribution of a statistic in a hypothetical very large series of equally sized statistical samples from the same population

Sample size (1) Number of observation units sampled (statistically or non-statistically) to be approached for possible inclusion as participants; (2) Number of observation units with data available for analysis

Target population The type of people about which evidence will be created in the research

Unbiasedness (- of an estimate) closeness to the true value in the target population

22.1.2 Point and Interval Estimation

Estimates obtained from a single sample are subject to uncertainty because of sampling variation. Thus, a single sample can only provide a point estimate, 'a best guess' of the underlying target population's value. For example, a sample mean is a point estimate of the underlying target population mean. A point estimate is thus always surrounded by a margin of error, also known as an interval of uncertainty. This uncertainty can itself be estimated from the sample data and expressed as an interval estimate, such as a 95 % confidence interval (CI). The possibility of estimating a margin of uncertainty is based on the facts that:

1. The estimate obtained from a single sample will be more precise (the confidence interval narrower) with increasing sample size. If the entire target population could be sampled, there would be no sampling-related uncertainty anymore.
2. If the spread of values in the population is narrower (smaller true variance), then the spread of estimates found in repeated samples (the *standard error*) and thus the confidence interval will also be narrower.

As mentioned in Chap. 13, three main methods exist to obtain confidence intervals: the standard error-, bootstrapping-, and likelihood-based methods. In this chapter we will only discuss the standard error-based method (For bootstrapping, *See:* Chap. 13). With this method, once a population distribution is estimated and characterized, one can use this distribution to determine the probability of finding certain values or the probability of finding a value within a certain range.

22.1.3 Crude and Adjusted Estimates

In Chap. 13 we listed a variety of possible reasons why initial crude estimates may need some adjustment. Most important of all is adjustment for confounding, as such adjustments are essential – definitional even – to analytical research in epidemiology. Since this topic is so important, it will be covered in a separate section below.

22.2 Estimation of Outcome Frequency

The estimators described in this section are frequently used in case-series studies, surveys, and epidemic pattern studies. These estimators include the prevalence rate, cumulative incidence, incidence rate, and pseudo-rate. Some of these were mentioned in Chap. 2 in our discussion of basic concepts in epidemiology. Here, we expand that earlier discussion to include the calculation of relevant standard error-based confidence intervals, potential validity issues, and frequently required types of adjustment. Formulas are given only for quick and practical reference.

22.2.1 Prevalence (*Syn.* Prevalence Rate)

The point estimate of a prevalence rate is calculated by dividing the number of participants (or other observation units) with the characteristic of interest by the number who could have had the characteristic of interest, and then multiplying the quotient by 100. We can calculate the 95 % confidence interval as follows:

95 % Confidence Interval of a Prevalence Rate
Step 1:
Calculate the standard error (SE) of the proportion

$$SE_{\text{proportion}} = \sqrt{\frac{p(1-p)}{n}} \tag{22.1}$$

Where:
p = proportion with the characteristic of interest
n = number examined for the characteristic of interest

Step 2:
Calculate the upper and lower limits of the 95 % confidence interval of the proportion and multiply the values by 100:

$$\text{Lower limit} = p - 1.96 * SE$$

$$\text{Upper limit} = p + 1.96 * SE$$

Table 22.1 Weighted prevalence estimation of type 2 diabetes in a survey with quota sampling

Target population (Size 10,000)	Sample size	Type-2 diabetes (%)	Unadjusted point estimate of the overall prevalence	Adjusted point estimate of the overall prevalence
Ethnic **group A** (n = 8,000)	100	10 % (10/100)	5 % (15/300)	8.5 % (850/10,000) (850 composed of: 10 % of 8,000, 2 % of 1,000, and 3 % of 1,000)
Ethnic **group B** (n = 1,000)	100	2 % (2/100)		
Ethnic **group C** (n = 1,000)	100	3 % (3/100)		

In surveys, adjustments of prevalence rates may need to be done to take the sampling scheme into account. As a simple illustration of adjusting for a sampling scheme, consider a survey (n = 300) concerning the prevalence of type 2 diabetes. The survey used quota sampling (*See:* Chap. 9), in which 100 persons were sampled from each of three ethnic groups in the target population: A, B, and C. These ethnic groups represented 80 %, 10 %, and 10 % of the target population, respectively. Table 22.1 shows the findings and the *weighting* that was necessary to arrive at an overall prevalence estimate. For adjustments applicable to various sampling schemes and for finite population adjustments, we refer to statistical and survey handbooks.

The major possible validity issues with prevalence estimation are lack of sensitivity and/or specificity of assessing the outcome. Low sensitivity leads to underestimation and low specificity to over-estimation of the prevalence rate. When information is available on the relative extent of these problems, it may be possible to adjust the estimate.

22.2.1.1 Problems with Dichotomizations

Prevalence estimates may concern the presence of a characteristic that was assessed via categorization of a continuous variable. For example obesity is commonly assessed by categorizing the continuous body mass index variable. Other examples are anemia based on low blood hemoglobin levels, and stunting or wasting in children. In those instances one needs to consider the possible influences of measurement error in the continuous variable on the prevalence estimate. Prevalence tends to be over-estimated in the common scenario of measurement imprecision (non-differential). Indeed, random measurement error will inflate the variance of the variable and increase the proportion falling below a cut-off in the lower tail of the frequency distribution.

Moreover, when terminal digit preference (*See:* Chap. 29) affects the measurement values of the continuous variable, the frequency distribution will show some peaks surrounded by dips. When the cut-off for categorizing and defining the outcome (e.g., body mass index greater than or equal to 30 for obesity) falls on a peak or a dip caused by digit preference, some borderline values will be misclassified. The prevalence will consequently be over- or underestimated. In such instances, it may be better to first smooth the frequency distribution before the cut-off is applied and the prevalence calculated.

22.2.1.2 Period Prevalence

A 'period prevalence' is the proportion of observation units that ever exhibited the state of interest, including those units that already exhibited the state at the start of the period. The derivation of a period prevalence and its interval estimate is, in principle, identical to the process used to derive a point prevalence and interval estimate.

22.2.2 Cumulative Incidence (*Syn.* Incidence Risk)

Cumulative incidence is mostly used in descriptive cohort studies to estimate the risk of new occurrences of an all-or-nothing state in the cohort or a subgroup thereof. A point estimate of incidence risk is calculated as the number of subjects who develop the outcome of interest divided by the number of at-risk subjects being followed, times 100. This applies to a pre-specified period, e.g., cumulative incidence over 2 years, over 5 years, over x years. An interval estimate can be calculated as:

95 % Confidence Interval of a Cumulative Incidence

Step 1:

Calculate the standard error (SE) of the proportion, as in Eq. 22.1.

Where:

p = proportion developing the outcome of interest in the specified observation period

n = number of subjects **at risk** of getting the outcome followed during the observation period

Step 2:

Calculate the upper and lower limits of the 95 % confidence interval of the cumulative incidence, as described in association with Eq. 22.1.

Validity issues of frequent concern include *informative censoring*. This can occur when not all individuals have data for the entire specified time period. The cumulative incidence estimate is biased if loss to follow-up or loss by a competing risk-event (e.g., death) is related to the risk for the outcome of interest. Another possible problem is a lack of sensitivity or specificity of assessing the outcome. When information is available on the extent of these problems it may be possible to adjust the estimate. When there is considerable loss to follow-up, it is usually preferable to use the incidence rate instead of cumulative incidence as the outcome parameter (next subsection).

22.2.3 Incidence Rate (*Syn.* Incidence Density)

Incidence rates are used to estimate the rate of new occurrences of an all-or-nothing state in a cohort or dynamic population. For the calculation of point and interval estimates of the incidence rate, one can use the following approach:

Point Estimate of an Incidence Rate

$$\text{Incidence rate} = \frac{A}{B} \tag{22.2}$$

Where:

A = number of first events among subjects at risk and followed

B = total person time contributed by all at-risk subjects followed; each subject’s contribution censored at, whichever comes first:

- End of planned follow-up
- Or, loss to follow-up
- Or, death
- Or, first occurrence of event

Confidence Interval of an Incidence Rate

Step 1:

Calculate the SE of the natural logarithm of the incidence rate:

$$\text{SE} = \frac{1}{\sqrt{A}}$$

Where:

A = number of first events among subjects at risk and followed

Step 2:
Calculate the upper and lower limits of the 95 % confidence interval for the incidence rate:

$$\text{Lower limit} = \mathrm{e}^{\frac{A}{B}-1.96*SE}$$

$$\text{Upper limit} = \mathrm{e}^{\frac{A}{B}+1.96*SE}$$

Where:
e = the natural number *(~2.71)*
$\frac{A}{B}$ = the point estimate of the incidence rate

Validity issues of frequent concern include the problems of informative censoring and low sensitivity or specificity of outcome detection mentioned above, as well as a number of dubious and controversial variants of the incidence rate encountered in the literature (which relate to ways the numerator and/or denominator are calculated):

- Counting several events per subject, e.g., fractures, disease episodes, recurrences (For a discussion of problems with this, *See:* Windeler and Lange 1995)
- Individual person-time censored differently, e.g., only at death, loss to follow-up and end of follow-up

22.2.4 Pseudo-rates

The concept of a pseudo-rate was introduced in Chap. 6. The following are several commonly used pseudo-rates:

- *Crude birth rate* – Crude birth rate is a pseudo-rate measure of population fertility and is calculated as the ratio of (1) 1,000 times the number of live births to residents in the area in a calendar year, and (2) the estimated mid-year population in the same area in the same year
- *Crude death rate* – This is a pseudo-rate measure of mortality burden in a population and is calculated as the ratio of (1) 1,000 times the number of deaths in the area in a calendar year, and (2) the estimated mid-year population in the same area in the same year
- *Stillbirth rate* – This is the number of fetal deaths in a year divided by the sum of all live births and fetal deaths in the same population in the same year
- *Infant mortality rate* – This is a pseudo-rate measure of mortality in infants in a given area and is calculated as the ratio of (1) 1,000 times the number of death infants younger than 1 year in a given year and (2) the number of live births in the same area in the same year

- *Child mortality rate* – This is a pseudo-rate measure of mortality in children under 5-years-old for a specified area and a specified calendar year and is calculated as the ratio of (1) the number of children under 5-years-old who died in a given year and (2) the number of live births that year
- *Maternal mortality ratio* – This pseudo-rate is a population indicator of the burden of maternal mortality and is calculated as the ratio of (1) the number of women in the area who die during pregnancy or childbirth in a chosen period and (2) the number of live births in the same area in the same period, 'extrapolated' to the number of women who would have died in a period with 100,000 live births

To evaluate validity problems and possibly needed adjustments, one should bear in mind that the numerator and denominator information of the ratios tends to come from multiple sources, e.g., different study design types or studies with different timelines of data collection. For example, data from surveillance systems and registries may be combined with data from surveys. Each of these source studies may have their own problems of selection bias and information bias. Extrapolations are sometimes made from counts in parts of years to estimated counts for entire years, or from smaller areas to wider areas. Justifications for this need to be checked carefully.

In this section, we discussed single frequency estimates and their validity problems. The next section discusses estimation strategies when there is an interest in how outcome frequencies compare among categories of interest, still in the context of descriptive studies.

22.3 Estimation in Comparative Descriptive Studies

In descriptive studies, comparisons of frequencies can be made in several ways that may involve either measures of frequency or measures of association.

Formal methods of comparison include:

- *Estimating measures of association*: the estimator itself captures the frequency contrast between subcategories or is a parameter of an age/time trend or other relationship examined with regression modeling. Examples are the difference in prevalence, prevalence ratio, incidence rate ratio, and beta-coefficient
- *Statistical testing*: the test statistic is calculated to explore hypotheses about the possible existence of effects (e.g., differences, statistical relations, etc.). The topic of statistical testing is discussed in Chap. 23 (Statistical Testing).

Informal and semi-formal comparison methods include:

- Checking for *non-overlap of interval estimates* obtained for different categories
- *Visual smoothing* of trends in the estimates obtained for multiple categories (*See:* Chap. 24)
- *Standardization of estimates* obtained for different groups or the same group at different times (*See:* next section)

In this section we will only further discuss the estimation of measures of association. Some of these (odds ratio, relative risk) have been mentioned briefly in Chap. 2 as basic concepts of epidemiology. We expand here on point and interval estimation and briefly mention some validity concerns. Formulas are given for quick reference, but their derivations are not discussed.

22.3.1 Difference in Prevalence Rate

The difference in prevalence rates is a measure of association often used in surveys to contrast cross-sectional frequencies among subgroups. Point and interval estimates can be calculated using the following equations:

Point Estimate of a Difference in Prevalence Rate

$$\text{Difference in prevalence} = \left(p_2 - p_1\right)*100 \tag{22.3}$$

Where:
p_2 = proportion with the characteristic of interest in group 2
p_1 = proportion with the characteristic of interest in group 1

95 % Confidence Interval of a Difference in Prevalence Rate
Step 1:
Calculate the SE of the difference in proportion using the following equation:

$$SE = \sqrt{\frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}}$$

Where:
n_2 = size of group 2
n_1 = size of group 1

Step 2:
Calculate the upper and lower limits of the 95 % confidence interval of the difference in proportion and multiply the values by 100:

$$\text{lower limit} = \left(p_2 - p_1\right) - 1.96*\text{SE}$$

$$\text{upper limit} = \left(p_2 - p_1\right) + 1.96*\text{SE}$$

As to possible sources of bias, one first considers possible misclassifications of the outcome. If the bias differs for the two groups (e.g., the true prevalence is underestimated by 2 % in one group and over-estimated by 2 % in the other), then the difference in the prevalence rate will be biased. But if the two groups have the same absolute bias (e.g., the true prevalence is under-estimated by 2 % in both groups), then the estimate of the difference in prevalence will be unbiased.

Next, one considers misclassification of the determinant (in this case the characteristic that defines whether one belongs to group 1 or 2). For example, the prevalence of an illness may be compared between groups with higher and lower socioeconomic status (SES). Misclassification of SES will bias the estimate of the difference in prevalence towards the zero value if this misclassification is unrelated to the illness. If the misclassification of the determinant *is* related to the illness (e.g., only high

SES people *with* the illness are misclassified as low SES), then the effect can be to over- or under-estimate the difference in prevalence.

22.3.2 Prevalence Rate Ratio

Like the difference in prevalence rates, the prevalence rate ratio is another measure of association often used in surveys to contrast cross-sectional frequencies among subgroups. Point and interval estimates can be calculated using the following equations:

Point Estimate of a Prevalence Rate Ratio

$$\text{Prevalence rate ratio} = \frac{p_2}{p_1} \tag{22.4}$$

Where:
p_2 = proportion with the characteristic of interest in group 2
p_1 = proportion with the characteristic of interest in group 1

95 % Confidence Interval of a Prevalence Rate Ratio

Step 1:
Calculate the standard error (SE) of the natural logarithm of the prevalence rate ratio using the following equation:

$$\text{SE} = \sqrt{\left(\frac{1}{A_2} - \frac{1}{n_2}\right) + \left(\frac{1}{A_1} - \frac{1}{n_1}\right)}$$

Where:
A_2 = number with the characteristic of interest in group 2
A_1 = number with the characteristic of interest in group 1
n_2 = size of group 2
n_1 = size of group 1

Step 2:
Calculate the upper and lower limits of the 95 % confidence interval of a prevalence rate ratio:

$$\text{Lower limit} = \frac{\frac{p_2}{p_1}}{e^{1.96*\text{SE}}}$$

$$\text{Upper limit} = \frac{\frac{p_2}{p_1}}{e^{1.96+\text{SE}}}$$

Where:
e = the natural number *(~2.71)*

When considering validity problems for the prevalence rate ratio the concern is how outcome misclassifications affect the numerator and/or denominator of the ratio. Other validity problems for the prevalence rate ratio are analogous to those discussed for differences in prevalence.

22.3.3 Relative Risk

The concepts of risk and relative risk were introduced in Chap. 2. This measure of association can be used in descriptive longitudinal studies to contrast the cumulative incidence (*See:* Sect. 22.2.2) in two groups. Point and interval estimates can be calculated using Eq. 22.4 and associated formulae, where p_2 and p_1 represent the proportion with the event of interest in groups 2 and 1, respectively, and A_2 and A_1 represent the numbers of individuals who developed the outcome of interest in groups 2 and 1, respectively.

As to misclassification in outcome assessments, when the errors in the two cumulative incidence estimates are proportionally the same (e.g., when the cumulative incidence is underestimated by half in the two groups), the ratio estimate remains unbiased.

Misclassification of the determinant – if unrelated to prognosis – tends to have an attenuating effect on the relative risk estimate (closer to the value 1).

22.3.4 Incidence Rate Ratio

This measure of association is often used in descriptive longitudinal studies, when individual follow-up times are quite variable. It captures a contrast between the incidence rates of two groups (exposed and unexposed). Point and interval estimates can be calculated using the following equations:

Point Estimate of an Incidence Rate Ratio

$$\text{Incidence rate ratio} = \text{IRR} = \frac{\dfrac{A_2}{B_2}}{\dfrac{A_1}{B_1}} \tag{22.5}$$

Where:

A_2 = number who developed the outcome of interest in group 2 (exposed)
A_1 = number who developed the outcome of interest in group 1 (unexposed)
B_2 = total person time in group 2 (For calculation of total person time, *See:* Sect. 22.2.3)
B_1 = total person time in group 1

95 % Confidence Interval of an Incidence Rate Ratio

The following formula is easy to use for obtaining a good approximation of the upper and lower limits of the 95 % confidence intervals for the IRR:

$$\text{Lower limit} = \text{IRR} \,/\, e^{1.96^{*}\sqrt{\frac{1}{A_2}+\frac{1}{A_1}}}$$

$$\text{Upper limit} = \text{IRR} * e^{1.96^{*}\sqrt{\frac{1}{A_2}+\frac{1}{A_1}}}$$

Where:
e = the natural number *(~2.71)*

Validity concerns for the IRR are analogous to those for relative risk. Also of concern are possible dubious variants of the incidence rates, as mentioned in Sect. 22.2.3.

22.4 Standardization of Estimates

Standardization is a way to enhance descriptive comparisons of rate estimates across multiple groups or within the same group at different time periods. It is used to compare groups that may differ in characteristics that influence the outcome rate. In such circumstances, failure to execute standardization may lead to observed differences that are, to some extent, attributable to the different compositions of the groups. In other words, standardization may help to rid the influence of these background factors. Theoretically, it is possible to adjust for any background factor, but standardization is mostly done for age and sex.

There are two methods of standardization: direct and indirect standardization. We first discuss direct standardization below and thereafter only briefly mention indirect standardization.

22.4.1 Direct Standardization

This method applies rate estimates in two or more samples to a single underlying distribution of determinants, such as an age- or sex-distribution. For example, if the rates are age-dependent, the crude rates will be transformed into *standardized rates* that likely would have been attained were the age structures in the samples the same. Let’s look at a simple example in Table 22.2, in which investigators are attempting to compare mortality rates in two areas. In areas A and B the mortality rates are calculated to be 427 deaths per 100,000 inhabitants and 507 deaths per 100,000

Table 22.2 Age-specific mortality rates in two areas

	Area A			Area B		
Age category	Number of individuals	Number of deaths	Mortality rate per 100,000	Population size	Number of deaths	Mortality rate per 100,000
20–44	400,000	1,120	280	250,000	675	270
44–64	500,000	2,365	473	450,000	2,115	470
65+	100,000	785	785	300,000	2,280	760
All ages	1,000,000	4,270	427	1,000,000	5,070	507

Panel 22.2 Direct Age Standardization in Six Steps

1. Stratify the two study groups into the same age strata
2. Calculate the age-specific rates for each stratum in each group
3. Choose a standard population
4. Calculate the expected number of outcome events for each stratum in each group
5. Calculate the total number of expected outcome events in each group
6. For each group, divide this total expected number of outcome events by the total size of the standard population

inhabitants, respectively. By employing direct age standardization, one can calculate what the mortality rates would have been if the two areas had the same age structure. The procedure follows the steps listed in Panel 22.2.

22.4.1.1 Step-1: Stratify the Two Study Groups into the Same Age Strata

The example illustrated in Table 22.2 is based on imaginary data with only three age strata (to simplify matters). In general, the size and number of the strata have to be concordant with the available data of the chosen standard (*See:* Step-3).

22.4.1.2 Step-2: Calculate the Age-Specific Mortality Rates for Each Stratum in Each Group

For this calculation the numbers of individuals in the population and of the occurred deaths in each of the age strata are needed. In the example in Table 22.2, each of the age-specific mortality rates in Area A is higher than in Area B, but the overall death rate is higher in Area B. Looking at the size of the populations in the strata, it becomes clear the age structure in Area A is weighted towards younger people and that of Area B is weighted towards older people (an age stratum that has a much higher mortality rate).

22.4.1.3 Step-3: Choose a Standard Population

The next step is to choose a *standard population* whose age distribution will be used as a common hypothetical age distribution for all study populations.

Table 22.3 Calculation of expected numbers of deaths in a chosen standard population

		Area A		Area B	
Age-group	Standard population size[a]	Mortality rate per 100,000	Expected number of deaths	Mortality rate per 100,000	Expected number of deaths
20–44	100,149,000	280	280,418	270	270,403
45–64	60,991,000	473	288,488	470	286,658
65+	34,710,000	785	272,474	760	263,796
Total	195,850,000		841,380		820,857

[a] U.S. population from the year 2000 is used. Available at www.cdc.gov

When identifying a standard population, one of the following possibilities is commonly selected:

- One of the study populations
- The entity of the study population groups combined
- An “external” population, i.e., the population from a local area or country
- A hypothetical population

The choice is somewhat arbitrary, and there is no correct or incorrect standard. But it should be kept in mind that the choice has an effect on the standardized rates, i.e., a relatively young standard population gives more weight to rates in younger age groups while a relatively old standard population gives more weight to the rates in older age groups. Therefore the standard should preferably be representative of the study populations being compared.

22.4.1.4 Step-4: Calculate the Expected Number of Deaths for Each Stratum in Each Study Population

For each study population and each age stratum, one calculates the expected age-specific death rates by applying the crude death rate (from step-2, *See:* Table 22.2) to the chosen standard population. This calculation can be executed using the following formula below. The results of this step in our example are shown in Table 22.3.

$$\frac{\text{Size of standard population} \times \text{age specific mortality rate}}{100,000}$$

22.4.1.5 Step-5: Calculate the Total Number of Expected Deaths in Each Study Population

This is done by simply summing the calculated number of expected deaths in each stratum calculated in step-4 (*See:* “Total” in Table 22.3).

22.4.1.6 Step-6: For Each Study Population, Divide the Total Expected Number of Deaths by the Total Size of the Standard Population

This step allows one to obtain the age-standardized mortality rates

$$\text{Adjusted mortality rate area A}: \frac{841,380}{195,850,000} \times 100,000 = 429\ \mathit{deaths\ per}\ 100,000$$

$$\text{Adjusted mortality rate area B}: \frac{820{,}857}{195{,}850{,}000} \times 100{,}000 = 419 \; \textit{deaths per} \; 100{,}000$$

Thus, it is evident that area B had a higher crude mortality rate than area A (Table 22.2), but area B's age-adjusted mortality rate was *lower* than area A's. So it can be stated that the difference in the crude rates were to some extent attributable to the differences in age structure.

22.4.2 Indirect Standardization

For the indirect method, specific rates from the standard population are applied to the populations being compared. For example, the age-specific mortality rates of the standard population might be applied to the corresponding age-strata in areas A and B. In other words, indirect standardization is the inverse of direct standardization. This method is often used when insufficient strata-specific data for the study populations are available (either they are unavailable or too susceptible to random variability).

22.5 Estimation in Analytical Studies

Estimation in analytical studies first involves a *crude analysis,* in which control for confounding is not yet a concern. In cohort studies and trials this step uses methods that are analogous to the methods described in Sect. 22.3, and will typically yield a crude relative risk or incidence rate ratio estimate. In Sect. 22.5.1 we will discuss crude analysis in case control studies, which typically leads to an estimate of a crude exposure odds ratio. We will show that it can also, with proper sampling of the 'controls,' lead to a direct estimate of a crude incidence rate ratio. There are several utilities to crude analyses, including:

- Checking for confounding (by comparing crude versus adjusted estimates)
- Checking for effect modification by that third factor (by comparing crude estimates across strata of the third factor)
- Controlling for confounding (by calculated crude estimates separately in strata of a suspected confounder)

After crude analysis follows an analysis that controls for confounding and leads to the estimation of the *causal rate ratio* (*See:* Sect. 22.5.2 and Chap. 24). From this latter statistic, one can derive *secondary outcome parameters*, e.g., number needed to treat, vaccine efficacy, etc. (*See:* Sect. 22.5.3).

22.5.1 Crude Analyses in Case-Control Studies

The concepts of odds and the odds ratio were introduced in Chap. 2. In crude analysis of unmatched case-control studies, the point estimate of the *crude exposure odds* *ratio* is the odds of exposure in cases divided by the odds of exposure in the controls (data from a simple 2 × 2 table). An interval estimate can be calculated using the following equations:

95 % Confidence Interval of a Crude Exposure Odds Ratio

Step-1:

Calculate the SE of the natural logarithm of the odds ratio using this equation:

$$\mathrm{SE} = \sqrt{\frac{1}{A_2} + \frac{1}{A_1} + \frac{1}{B_2} + \frac{1}{B_1}}$$

Where:

A_2 = number of exposed cases
A_1 = number of unexposed cases
B_2 = number of exposed controls
B_1 = number of unexposed controls

Step 2:

Calculate the upper and lower limits of the 95 % confidence interval of the odds ratio:

$$\text{Lower limit} = \mathrm{e}^{\ln\left(\frac{A_2}{A_1} \div \frac{B_2}{B_1}\right) - 1.96 * \mathrm{SE}}$$

$$\text{Upper limit} = \mathrm{e}^{\ln\left(\frac{A_2}{A_1} \div \frac{B_2}{B_1}\right) + 1.96 * \mathrm{SE}}$$

Where:

ln = natural logarithm
e = the natural number *(~2.71)*

Point and interval estimates of the crude exposure odds ratio can also be obtained from single-predictor logistic regression analysis. This method is explained in Chap. 24 (Statistical Modeling). As a form of quality control, results of 2 × 2 table analyses and coefficients from single predictor models should always be compared to check if they lead to identical parameter estimates (Vandenbroucke 1987).

When the controls are a representative sample of the source population of the cases (*See:* Chap. 6 for explanation of this concept), the crude analysis can directly estimate an incidence rate ratio (IRR) or relative risk (Miettinen 1976). This is to be seen as an advantage because IRR and relative risks have a more intuitive interpretation than do odds ratios. Indeed, the source population can be seen as an underlying cohort. The group of cases that ended up in the case-control study can be seen as a representative sample of cases developing in this underlying cohort, and they should have an exposure distribution (number of cases exposed/number of cases unexposed; i.e., A_2/A_1) identical to the exposure distribution of all cases developing in the underlying cohort. Similarly, since the controls included in the case-control study are a representative sample of the same underlying cohort, the population time distribution (unexposed person time/exposed person time; i.e., B_1/B_2) observed in the controls should be the same as the population time distribution in the entire underlying cohort. Finally, since the cases and controls provide us with A_2/A_1 and B_1/B_2 estimates for the underlying cohort/source population, we can calculate the incidence rate ratio in the source population as:

$$\text{IRR} = \frac{A_2}{A_1} * \frac{B_1}{B_2}$$

Note that this formula is a simple transformation of the familiar IRR formula (Eq. 22.5):

$$\text{Incidence rate ratio} = \text{IRR} = \frac{\frac{A_2}{B_2}}{\frac{A_1}{B_1}} \tag{22.5}$$

B_1 and B_2 represent person-time information for the controls, not simply numbers of exposed and unexposed individuals. This is to account for the possibility that for some controls complete exposure information is unavailable, i.e., it might be available only for part of the etiologically relevant period. In simpler, more frequent instances, however, complete exposure information about the entire etiologic period is available for all controls. This will often be the case when the exposure is merely an 'ever/never' characteristic (e.g., ever traveled to Africa, yes or no) or a dichotomous characteristic in a defined interval in etiognostic time (e.g., traveled to Africa in the period 5–10 years ago, yes or no). The balance of population time (B_1/B_2) then reduces to the simple ratio involving numbers 'unexposed' and numbers 'exposed,' from the familiar 2 × 2 table.

22.5.2 Methods of Adjusting for Confounding During Analysis

In epidemiology, adjustment (control) for confounding during analysis traditionally uses one or more of the following approaches:

- Exclusion from analysis of subjects with rare confounding characteristics
- Regression analysis
- Stratified analysis, with or without calculation of a pooled estimate

All of these methods can be applied in case-control studies, cohort studies and trials. No single method of control during analysis can be considered optimal in every situation; each has strengths and limitations. In most situations, a combination of strategies will provide better insight and control.

Panel 22.3 Advantages and Disadvantages of Stratified Analysis

Advantages of stratified analysis
- Easy to calculate by hand
- Allows evaluation of confounding and effect modification
- Confounding is effectively controlled for

Disadvantages of stratified analysis
- Only a very small number of confounders can be accounted for
- Large sample sizes are needed to have enough numbers in each stratum

22.5.2.1 Exclusions from Analysis

This strategy is the equivalent to restriction at the design stage. If the few subjects with the rare confounding characteristic are excluded from the analysis, the confounder can no longer exert its influence.

22.5.2.2 Regression Analysis

The most popular method of controlling for confounding is to perform regression analyses. Most frequently used are:

- Multiple linear regression
- Multiple logistic regression
- Cox proportional hazards regression
- Poisson regression

These analyses will be discussed in Chap. 24. Briefly, all these methods allow for effective control of confounding by introducing the potential confounders as independent variables (as ‘covariates’ in addition to the exposure variable) in the model. In multiple regression analyses, the slope coefficients (beta-coefficients) of the independent variables and their precision are estimated. When all confounders are introduced in the model, the beta-coefficients for the exposure variable give an un-confounded measure of the effect of the exposure on the outcome.

22.5.2.3 Stratified Analysis

This approach involves calculating the outcome parameter estimate separately for each stratum of the potential confounder. For example, if one anticipates that age is a potential confounder, one calculates the odds ratio separately for each age category to control for that possibility. The separate odds ratios (one for each age category) are by definition free of confounding by age. Reporting these separate odds ratios could be sufficient if an overall pooled estimate of the effect measure (for all age categories combined) is not strictly required or wanted. Controlling for confounding by stratified analyses has some advantages and disadvantages (Panel 22.3).

22.5.2.4 Pooled Estimation

This method can be helpful in studies where only one (or at most two) confounders need to be adjusted for. Pooled estimates are based on the fact that, in a stratified analysis, each stratum-specific estimate (e.g., each odds ratio) is an un-confounded estimate of the overall un-confounded parameter, but the precision of each of those separate estimates depends on the size of the stratum. The principle of pooled estimation is therefore to obtain a weighted average of all the stratum-specific estimates (more weight given to strata with larger sizes). We only give the formula for calculating the pooled odds ratio point estimate using the Mantel-Haenszel method in an unmatched case-control study. For interval estimates and other scenarios, we refer to Kirkwood and Sterne (2003) and other statistical literature.

Point Estimate of the Mantel-Haenszel Odds Ratio

$$\mathrm{OR}_{MH} = \frac{\sum\left(\frac{a*b}{T}\right)}{\sum\left(\frac{c*d}{T}\right)} \tag{22.6}$$

Where:
$\sum$ = sum over all strata
T = size of stratum = a + b + c + d
a = number of exposed cases in the stratum
b = number of unexposed controls in the stratum
c = number of unexposed cases in the stratum
d = number of exposed controls in the stratum

The Mantel-Haenszel method assumes that the stratum-specific estimates only vary because of sampling variation, not because of effect modification by the confounder. To check this assumption, one can do a chi-squared test for heterogeneity. Simple inspection for non-overlap of the confidence intervals in the different strata can also be suggestive of effect modification. Comparison of the Mantel-Haenszel odds ratio with the exposure odds ratio obtained from the crude analysis is informative about confounding: if the Mantel-Haenszel odds ratio is identical to the crude odds ratio, then no confounding is present.

22.5.3 Calculation of Secondary Outcome Parameters

After one calculates the effect measure, it is sometimes useful to derive secondary outcome parameters. These estimates are often helpful in making arguments about the importance of an observed effect measure. Such secondary outcome parameters often include: attributable fraction (AF), number needed to treat (NNT), and vaccine efficacy.

22.5.3.1 Attributable Fraction

All health outcomes are multifactorial, and some factors matter more than others. To estimate the degree to which an exposure contributes to an outcome, one may calculate the attributable fraction (AF). Usually determined in cohort studies, AF is calculated as the difference in the rates of an outcome between the exposed and unexposed (also known as the risk difference or RD), that quantity divided by the rate of the outcome in the exposed.

$$AF = \frac{I_e - I_o}{I_e} = \frac{RD}{I_e}$$

Where,
I_e = incidence of the outcome in the exposed
I_o = incidence of the outcome in the unexposed
RD = risk difference = $I_e - I_o$

Alternatively, the attributable fraction can be calculated as:

$$AF = \frac{CRR - 1}{CRR}$$

Where,
CRR = Causal rate ratio

The value for AF is often interpreted as the expected degree of change in the rate of an outcome had no one been exposed. In the context of assessing etiognostic probabilities (*See:* Chap. 24, Sect. 24.4.3) AF is interpreted as the probability that a particular exposure causally acted in an individual.

22.5.3.2 Number Needed to Treat

This outcome parameter (*See:* Cook and Sackett 1995) is a measure of effect used in intervention-prognostic research, and is interpreted as the average number of individuals that need to be treated to prevent one additional outcome event. For example, if NNT for death is 19, then this estimate is interpreted to mean that, on average, 19 individuals need to be treated to save one life. NNT is calculated as the reciprocal of the absolute value of the risk difference (RD). The NNT is always calculated for a defined risk period in comparison with the control intervention or non-intervention.

$$NNT = \frac{1}{|RD|}$$

Where:
RD = risk difference = rate in the exposed minus the rate in unexposed

Sometimes a similar estimate, the number needed to harm, is calculated if the exposure is associated with an increased risk of an undesirable outcome event (e.g., an adverse event).

22.5.3.3 Vaccine Efficacy

Vaccine efficacy is the difference between the illness rates of the immunized and the non-immunized, expressed as a proportion (or percentage) of the illness rate among non-immunized.

This chapter provided an introduction to frequently employed methods of statistical estimations and their selection. The estimates produced are mostly evidence about frequencies, associations, or causal effects. In the next chapter, we will discuss how evidence about the existence *of associations and effects can be obtained using statistical tests.*

References

Cook RJ, Sackett DL (1995) The number needed to treat: a clinically useful measure of treatment effect. BMJ 310:452–454

Kirkwood BR, Sterne JAC (2003) Essential medical statistics, 2nd edn. Blackwell, Malden. ISBN 9780865428713

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

Miettinen OS (1985) Theoretical epidemiology. Delmar, New York, pp 1–359. ISBN 0827343132

Vandenbroucke JP (1987) Should we abandon statistical modeling altogether? Am J Epidemiol 126:10–13

Windeler J, Lange S (1995) Events per person year – a dubious concept. BMJ 310:454–456

Statistical Testing

23

Jan Van den Broeck and Jonathan R. Brestoff

Nisi crediteritis, non intelligitis.

Saint Augustine of Hippo

Abstract

Statistical testing is used for exploring hypotheses about the possible existence of effects (differences, statistical relations). One chooses a statistical test mainly on the basis of which type of variable or which distributional characteristic of a variable is to be compared and related. Each statistical test has its own type of *test statistic* that captures the amount of effect/difference observed in the sample data. The problem with observed effects in samples is that they are influenced by sampling variation (chance) and may not accurately represent real population effects. *P-values* are therefore attached to the observed values of a test statistic in an attempt to acquire better insight into whether an observed effect is real. P-values are the probability of finding the observed value of the test statistic, or a value more extreme than it, when the *null hypothesis* (that there is absence of an effect or difference) is in fact true. As such, P-values are sometimes but not always a good basis for *accepting or rejecting* a null hypothesis. After discussing the uses of statistical testing in epidemiology and different types of hypotheses to test, we discuss the interpretations of P-values and conclude with a brief overview of commonly used statistical tests.

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_23,

23.1 The Use of Statistical Testing in Epidemiology

Statistical tests are employed to obtain evidence about the possible *existence* of effects. Each statistical test produces a *test statistic* that captures the amount of an effect/difference observed in the study. Given the information conveyed by test statistics, it is not surprising that statistical testing is common practice in epidemiological research, although some have argued that there is no place for it (e.g. Rothman 2010). The arguments against statistical testing are often based on the fact that misinterpretations of P-values are commonplace and that statistical estimation is more informative about effect sizes than statistical testing. Misinterpretations are indeed common (Abelson 1995; Sterne and Davey Smith 2001). However, we argue that statistical testing still deserves a place in epidemiology because evidence about the mere existence of an effect may be needed. The existence of an effect is especially important to establish before estimating effect sizes and studying effect modification, both of which require relatively large sample sizes. For example, large-size trials with low prior probability of detecting a clinically important effect would be deemed unethical and would tend to get little practical or financial support without some prior evidence about the existence of an effect. Statistical testing can be part of a valid approach to generate this prior evidence in a smaller-sized study. Indeed, for *any* research questions about *existence* of effects/differences, testing can be a helpful tool provided that the interpretation of test results avoids some common pitfalls.

Panel 23.1 Selected Concepts and Terms Relevant to Statistical Testing

Alternative hypothesis A statistical hypothesis stating that two or more variables are expected to be statistically related (sometimes also specifying their expected degree of relatedness)

Heteroscedascity Lack of constancy of variance of the outcome over levels of the determinants

Hypothesis A scientific idea (*Adapted from Miettinen* 1985)

Nonparametric test Test involving no assumptions about the shape of the distribution of the variables concerned

Null hypothesis A statistical hypothesis stating that two or more variables are expected to be statistically unrelated, or, that a variable's distribution is not different from a theoretical distribution

Null case The case where the null hypothesis is actually true

Paired samples

1. Two series of data, the second representing re-measurements of the same attribute/experience of the same observation units
2. Matched samples

Parametric test Test involving assumptions about the shape of the distribution of the variables concerned

(continued)

Panel 23.1 (continued)

P-value (of a null hypothesis) Probability of finding a value for a test statistic at least as extreme as the value obtained in the study, in the null case

Sampling distribution Distribution of a statistic in a hypothetical very large series of equally sized samples from the same population

Sampling variation Variation in the distribution of a statistic in a hypothetical very large series of equally sized statistical samples from the same population

Significance level (of a test) A particular a priori P-value α used to label obtained P-values as ‘significant’ if the obtained P-value is smaller than α or ‘non-significant’ if the obtained P-value is greater or equal to α

Significant (Statistically -) Characteristic of a P-value that is lower than the chosen significance level

Statistical distribution The expected frequency of values of a statistic

Statistical testing Computation of a P-value (*Miettinen* 1985)

Table 23.1 Shortlist of statistical packages often used for statistical testing in epidemiological studies

Statistical package	Commonly perceived advantages and disadvantages
R and **R commander**	Free use
	Rapidly increasing range of statistical methods available
	Easy use of R commander for simple statistical operations
	Relatively difficult to use compared to SPSS, Epi-Info, and EpiData
STATA	Very wide range of statistical methods available
	Need to learn syntax language
	Expensive
SPSS	Easy use of the menu system
	Medium size range of statistical methods available
	Expensive
Epi-Info and **EpiData**	Free and easy use
	Limited range of statistical methods

Statistical testing is a study activity that almost always requires the use of *statistical packages*, pieces of software that facilitate data analysis and statistical testing. Table 23.1 is a shortlist of statistical packages that are particularly popular among health researchers. The user of the package can be an investigator, or someone delegated by an investigator to perform the testing, such as a student, professional analyst, statistician or, person working for a contract research organization. Whoever is involved, the key scientific concern is that the operator should not only know ‘which buttons to push’ but also have a good understanding of the validity issues at stake, which include the needs for:

- A cleaned dataset (*See:* Chap. 20)
- The appropriate choice of a test and options within a test (*See:* Table 23.2)

Table 23.2 Selection of statistical tests frequently applied in diagnostic epidemiological studies

Measurement scale of the outcome variable	Determinant levels compared	Applicable statistical tests
Categorical	One group[a]	Chi-squared Goodness-of-fit test
	Two independent groups	Chi-squared test for contingency tables; Fisher's Exact test
	Two related groups	McNemar test
	k independent groups	Chi-squared test for contingency tables; Chi-squared test for trend; Fisher's Exact test
Numerical with Normal distribution	One group[a]	One-sample t test; Kolmogorov-Smirnov one-sample test
	Two independent groups	Student's t test (for difference in means); F-test (for heteroscedascity)
	Two related groups	Paired t-test
	k independent groups	F tests done in the context of ANOVA
	k related groups	F tests done in the context of repeated measurements ANOVA
Numerical with non-Normal distribution	One group[a]	Kolmogorov-Smirnov one-sample test
	Two independent groups	Mann-Whitney test; Median test; Kruskal-Wallis test
	Two related groups	Paired Wilcoxon test
	k independent groups	Kruskal-Wallis test

[a]The observed frequency distribution is to be compared with a theoretical frequency distribution

- Verification of the assumptions underlying the chosen test; knowledge of actions that can be taken when assumptions are violated (*See:* Chap. 13 and below)
- Appropriate handling of missing values (*See:* Chaps. 12, 13, and 20)
- Correct interpretation of outputs produced by the package

This chapter takes the use of a statistical package as a given. A wide range of excellent statistical handbooks is available to the increasingly rare investigator who does not have access to computers or statistical packages and to those who wish to deepen their knowledge of how test statistics are calculated.

23.2 Null Hypotheses and Alternative Hypotheses

There are two types of hypotheses that one can generate: a null hypothesis and an alternative hypothesis. In study design, one usually only specifies an *alternative hypothesis*, which is formulated as a simple one-sentence expression of an expected relation (e.g., one hypothesizes that oral broad-spectrum antibiotic use before the age of 5 years increases the risk of developing allergies in children aged 5–10 years). In statistical testing one usually formulates and examines a *null hypothesis* stating that there exists 'no effect'. Sometimes, however, statistical testing can address specific alternative hypothesis stating that there exists an effect of a particular size. Below we expand on how null hypotheses and alternative hypotheses are addressed in statistical testing.

23.2.1 The Nature of Null Hypothesis Testing

Null hypotheses are statistical hypotheses that there is ‘no effect’ of a determinant on an outcome or ‘no difference’ in compared distributions. They relate the idea that two or more variables are statistically unrelated or that there is no difference between a single variable’s distribution and a theoretical distribution. *Null hypothesis tests* examine to what extent the empirical data support or detract from the null hypothesis. The output of that examination is usually a P-value associated with a test statistic. The P-value expresses an *amount of evidence for or against* the null hypothesis (Miettinen 2009a, b). Nowadays, the wider goal has become to decide whether one sees the P-value as a reason to *accept or reject* the null hypothesis. This is often done simply by checking whether the P-value is lower than an arbitrary pre-set threshold (called the *significance level* of the test, which is usually set by convention at P = 0.05). This practice was unintended by Fisher, the inventor of the P-value (Goodman and Berlin 1994; Miettinen 2009a), but has become pervasive in epidemiology and other scientific disciplines.

23.2.1.1 Test Statistics

A test statistic expresses the amount of effect/difference observed when groups are compared or when comparison is made between an observed and a theoretical summary value or distribution. Statistical tests differ in what exactly is being compared and in the type of test statistic capturing the amount of difference. For example, in a Student’s t test, the mean of a continuous variable is compared between two groups. The test statistic is the t statistic, which expresses the degree of inequality between the two observed means as their difference (mean 1 – mean 2) divided by the standard error of this difference. Another example is the One-sample t test. This test compares the mean of a continuous variable with a theoretical value y. The test statistic in this test expresses the degree of inequality between the observed mean and y as the difference (mean – y) divided by the standard error of this difference.

A further characteristic of a test statistic is that each empirically determined value of it should have a known probability of being observed in instances in which the observed inequality is due to sampling variation only (i.e., there is no real inequality). In other words, assuming that the null hypothesis is true, the test statistic must have a known frequency distribution of values found in a hypothetical, very large number of repeat samples from the same target population. This distribution is the model that underlies the statistical test, and on the basis of it, a P-value can be calculated for any empirical test statistic obtained.

23.2.1.2 P-Values

A P-value for a null hypothesis is the probability of finding a value for the test statistic at least as extreme as the value empirically found, in a scenario where the null hypothesis is true. A *one-sided P-value* (obtained in a *one-tailed test*) is the probability of obtaining a test statistic at least as extreme in one direction away from the null value. A *two-sided P-value* (obtained by a *two-tailed test*) is the probability of obtaining a result at least as extreme in the same or opposite direction away from the null value.

Whether or not the null hypothesis is actually true is always unknown. All we know is that empirical test statistics that occupy extreme positions within the sampling distribution model (with very small P-values) *would be* very rarely seen *if* the null hypothesis were true and the sampling distribution model valid. On the other hand, values of empirical test statistics that are located relatively close to the null value (with high P-values) would be nothing unusual if the null-hypothesis were true and the sampling distribution model valid. In a simplistic approach, these considerations are sometimes taken as an adequate-enough basis for making a judgment about the need to accept or reject the null hypothesis. What is commonly used is a threshold value for the P-value (a 'significance level') to which one attaches the expression of belief in question. We expand on the interpretation of P-values in Sect. 23.3.

23.2.2 Choice of Null-Hypothesis Test and Degrees of Freedom

The choice of an appropriate null hypothesis test concerns which test statistic to use and the sampling distribution of that test statistic in the null case. Tests tend to differ according to:

- The measurement level of the compared variable
- The number of groups to be compared
- The independence of groups to be compared
- The underlying distributions of the data to be compared (if the variable is continuous and, if so, if it is Normally distributed)

Selecting the appropriate type of test is important, and for diagnostic studies Table 23.2 can assist in making that choice on the basis of the features listed above. After selecting the type of test one may have to specify the Degrees of Freedom (DF). DF is an integer number that specifies the specific null distribution that will be most applicable for the chosen test. It is calculated as a function of the sample size and/or the number of statistical parameters involved in the calculation of the test statistic. In practice, when measurement levels and numbers and independence of comparison groups are correctly specified, the statistical package will automatically identify the degrees of freedom based on the number of selected records with information for the variable or model parameters.

23.2.3 Statistical Testing of Alternative Hypotheses

Current practice in medical research is that alternative hypotheses are rarely tested. Alternative hypotheses are hypotheses that the statistical association is non-null, that there *is non-equality* among compared groups *of a hypothesized amount*. This is rarely practiced in spite of the fact that researchers commonly use a hypothesized effect size as a basis for 'sample size and power calculation.' This state of affairs has been often criticized but has never changed. With non-null hypothesis testing, the sampling distribution of the test statistic is of the same shape as with null-hypothesis testing but is shifted in comparison with a null hypothesis test. A framework for using alternative hypothesis testing together with null hypothesis testing has been promoted by Miettinen (1985), who suggested providing a P-value function in research reports, i.e., a function showing the P-values for the null hypothesis as well as for a range of hypothesized effect sizes. We suggest the possibility of reporting at least two P-values: one for the null hypothesis and one for any reasonable alternative hypothesis, perhaps the one used in the study proposal as a basis for sample size calculations.

23.3 Interpretation of P-Values

23.3.1 The Significance Level of a Null-Hypothesis Test

The significance level of a test is an a priori P-value, α, used to label P-values obtained in study analysis as 'significant' if the test-obtained P-value is smaller than α or 'non-significant' if the obtained P-value is greater than or equal to α. This threshold α is called the 'significance level' of the test. It is usually set at $P=0.05$ or $P=0.01$. When multiple tests are carried out an adaptation to lower P-value thresholds (a 'Bonferroni correction') is commonly proposed. Whatever the chosen significance level, 'significant' is seen as 'highly detractive from the null-hypothesis' and 'non-significant' is seen as 'highly supportive' of it. Hence, researchers adhering to the contemporary 'culture of significance' have used this dichotomy to 'accept' or 'reject' null hypotheses as the conclusion of the testing. Let us take a critical look at this approach.

23.3.1.1 Problems Associated with the Use of Significance Levels for Interpretation

A problem with the conclusion-oriented approach is that the amount of evidence produced for or against a null hypothesis depends on other factors than the P-value alone. P-values should be interpreted in light of and together with the amount of information in the data (which is heavily determined by sample size). A non-significant P-value, for example, when based on a very small sample, provides no evidence at all (neither for nor against the null hypothesis). Perhaps the most frequent misinterpretation is the idea that a non-significant P-value implies 'no effect' (Altman and Bland 1995; Evans et al. 2009). Panel 23.2 contains hints that may help in avoiding grave misinterpretations of P-values.

P-values should also be interpreted in light of and together with the prior credibility of the (null-) hypothesis (Miettinen 2009a; *See also:* Sect. 23.3.3). Indeed, whenever the prior probability of a hypothesis is very low, a 'significant' P-value does not weigh heavily as evidence and should be regarded as a possible case of bad/good luck. Evidence and its understanding are inextricably linked to prior credibility, hence the opening quote of this chapter "Nisi crediteritis, non intelligitis" ("If you don't believe it, you won't understand it"), or, to cite Richard Bach, "There is nothing we know until intuition agrees" and this applies to scientific knowledge as much as to any form of knowledge.

Panel 23.2 Avoiding Pitfalls in Interpretation of P-Values from Null Hypothesis Tests When a Significance Level of 0.05 Was Chosen

When the **P-value is < 0.05,** consider that:

- A P-value of <0.05 will falsely appear in 5 % of instances where there is actually no true effect/difference; the smaller the P-value, the stronger is the evidence against the null hypothesis, but this does not warrant complete 'rejection' of it
- An existing true effect may be very small and can have little or no clinical relevance; clinical relevance can be seen as the potential of an effect to bring about a change in clinical or public health practice
- In very large studies, even a P-value of 0.001 can be found for differences of a magnitude that is either true but irrelevant, or untrue and caused by a small bias

When the **P-value is >0.05,** consider that:

- This does not mean that there is no true effect, especially in small studies. The higher the P-value the more the null hypothesis becomes viable, but this does not warrant blind 'acceptance' of it
- For an appreciation of the range of possible true effect sizes, one can look at the confidence interval around the point estimate of the effect size
- The bigger the sample size, the more evidence there is in favor of the null hypothesis

Finally, there is no reason why one should restrict the evidence from testing to P-values from null hypotheses alone. As shown by Miettinen (1985), additional useful evidence can be obtained from P-values of 'alternative' hypotheses as well and can be conveniently presented as a P-value function. P-value functions model the P-values for a range of hypothesized effect sizes (including a zero effect).

23.3.2 Adaptations Towards Lower P-Value Thresholds

P-value adaptations for 'multiple testing' are often advocated when multiple comparisons are made in the same study. A strong tradition exists to make Bonferroni-type adjustments of the P-values in those instances. We join the several authors who have counter-argued that there is rarely if ever a need for such adjustments (Miettinen 1985; Rothman 1990; Perneger 1998; Nagakawa 2004). In every research study the experience of members of a study base is documented by measuring a selected number of attributes among those members. This selection of attributes for measurement is decided before the start of the data collection. It is important to see that, except maybe in the case of a Hawthorne effect, the experience lived by the study base members remains unchanged, whether it was one or multiple occurrence relations that were put forward as being of interest before, during, or after the data collection. Any change in opinion about which relations should be examined, even if data-driven, leaves the documented experience unaffected. In other words, the outcome parameter estimates (causal rate ratios, etc.) and P-values remain unchanged by the investigators' multiplicity of interests or change in those interests. For example, any new additional hypothesis tested *after* data collection, if it had been the only hypothesis proposed *before* data collection, would have led to exactly the same P-value. Thus, adding multiple and/or post-hoc comparisons does not alter the chance of finding a significant P-value for any of them.

23.3.3 Interpretation of P-Values Using the Bayes Factor

Enhanced inference from statistical testing requires taking into account the prior credibility of the null hypothesis. According to Miettinen (2009a), the latter can best be done by translation of P-values into Bayes factors. Briefly, one first determines the Z score corresponding to the P-value using standard statistical tables. For example, if the two-sided P-value produced by the statistical package is exactly 0.05, then Z = 1.96. With Z determined, the Bayes factor is then calculated as follows (Miettinen 2009a):

$$\textbf{Bayes factor} = \mathrm{BF} = \mathrm{Exp}\left[0.5\left(Z^2 - 1\right)\right] \tag{23.1}$$

This Bayes factor can now be used for a better interpretation of the test results, an interpretation that takes into account the prior probability (P^0) of the hypothesis (Miettinen 2009a). The P^0 is the probability that the null hypothesis is true, as appreciated (partly subjectively) before evidence from the study is obtained. A P^0 of 0.5 would correspond to equipoise. The Bayes factor now allows estimating how much the evidence in the data changes the prior probability. This is done by calculating the posterior probability (P^1) according to Bayes' rule as follows:

$$\textbf{Posterior probability} = \mathrm{P}^1 = \frac{1}{1 + \dfrac{1 - \mathrm{P}^0}{\mathrm{P}^0 * \mathrm{BF}}} \tag{23.2}$$

An informal comparison between P^1 and P^0 now allows subjective yet evidence-based interpretation of the test findings. The bigger the increase in probability, the stronger is the evidence in favor of the null hypothesis. As suggested by Miettinen

(2009a), a good way to present the evidence of statistical tests is thus to provide a table that lists selected prior probabilities and the corresponding posterior probabilities. Perhaps the selected prior probabilities for this table could represent a fair sample of expert opinions on the issue, gathered before start of data collection as part of the development of the study protocol.

23.4 Verification of Basic Assumptions of Tests

Recall that Chap. 13 has dealt with the statistical analysis plan at the design stage of the study. The plan may foresee statistical tests based on certain assumptions about the distribution of the prospective data. Once the data have been collected these assumptions need a good fact check, and revision of the initial plan may be needed. Assumptions worth checking depend on the type of test, and we refer to statistical handbooks for a complete treatment. Yet, there are some types of assumptions that are of frequent interest. We mention them briefly:

- All tests are based on an assumption of *stochasticity*; testing a null hypothesis when a null effect is impossible may not make sense
- The *measurement level* of the variables involved (*See:* Chap. 5) should be appropriate for the test
- Many tests have *distributional assumptions*. For example, Normality of the distribution of a continuous variable may need verification before doing a t-test; Some sub-types of t-tests require absence of heteroscedascity (i.e., the variances must be equal)

We will only expand here on the verification of distributional assumptions. Statistical tests exist to verify distributional assumptions underlying other statistical tests. For this purpose, the following tests are commonly used:

- *The Kolmogorov-Smirnov one-sample test* is a test of the goodness-of-fit of a cumulative distribution of measurement values with some specified theoretical distribution (e.g. a Normal distribution). The test statistic was designed on the basis that the largest difference between theoretically expected and observed values of cumulative distributions has a known sampling distribution in the null case
- *The Shapiro-Wilk test* is a general test of Normality and is based on comparing the ordered observed values with those expected if the distribution was Normal
- *The F-test* is a test for the null hypothesis that there is no difference in variance (no heteroscedascity) of a Normally distributed variable observed in two groups or attributable to two sources

23.5 An Overview of Commonly Used Tests

In this section we briefly mention some commonly used tests in epidemiology. For more information we refer to statistical handbooks.

23.5.1 Tests Commonly Used in Diagnostic Studies

Table 23.2 describes the commonly used families of tests in diagnostic research (for a discussion of what diagnostic research is, *see:* Chaps. 4 and 6). In diagnostic studies the choice of a test often starts by determining:

- The number of groups being compared (1 group – for comparison with a theoretical expected frequency distribution-, 2 groups, or *k* groups)
- Whether observations in the groups are independent (unpaired) or not
- The underlying measurement scale of the outcome being compared (categorical vs. numerical)
- The distributional characteristics of numerical outcomes being compared (Normal vs. non-Normal distribution)

For example, when two groups are to be compared for a continuous outcome variable then the choice can be for a (parametric) Student's t-test if the unpaired data have a Normal distribution and perhaps for a (non-parametric) Mann-Whitney U test if they don't have a Normal distribution. If the data are paired, it would be a (parametric) paired t-test or a (non-parametric) paired Wilcoxon test. When two groups need to be compared for a categorical variable, a chi-square test could be chosen in case of unpaired data, a McNemar test for paired data.

23.5.2 Commonly Used Tests in Etiologic and Prognostic Studies

23.5.2.1 Wald Test

The Wald test is perhaps the most frequently performed test in modern epidemiological research. It is a test for the hypothesis that the true causal beta-coefficient of an independent variable in a logistic regression model is zero. The test is based on the expectation that if beta is zero, the observed beta-coefficient in the model fitted divided by the standard error of this beta-coefficient has a Student's t sampling distribution with degrees of freedom equal to sample size minus 1 ($DF = n-1$). Most statistical packages automatically perform and report Wald test results when statistical modeling is done.

23.5.2.2 Chi-Square Tests for Goodness of Fit

These are Chi-square tests for checking if an observed frequency distribution of a categorical variable fits with an expected frequency distribution for that variable ('expected' being based on some theory/model). This type of test is therefore also used in descriptive prognostic studies. With a perfect fit, the number of observation units observed (O) and the number expected (E) in each category should be the same. The greater the O–E differences, the poorer the fit. The chi-square test statistic is the sum, over all categories k, of $(O-E)^2/E$. These values, under the null case, follow a chi-square distribution with $DF = k-1$.

23.5.2.3 Hosmer-Lemeshow Goodness-of-Fit Test

The Hosmer-Lemeshow goodness-of-fit test is commonly used in logistic regression analyses. This test is a Chi-square test for the goodness-of-fit of a fitted logistic regression model, comparing observed and expected numbers in deciles of risk predicted by the risk function derived from the fitted logistic regression model.

23.5.2.4 Log Rank Test

The log rank test assumes that if a hazard ratio is truly constant and equal to 1 throughout follow-up time, then the Mantel-Cox estimate of the hazard ratio has a known Chi-square sampling distribution with one degree of freedom.

In Part IV we have thus far discussed statistical estimation and statistical testing. In the next chapter, we discuss how both of these activities require assumptions on the applicability of specific statistical models.

References

Abelson RP (1995) Statistics as principled argument. Lawrence Erlbaum Associates, Hilsdale, pp 1–221. ISBN 0805805281

Altman DG, Bland JM (1995) Absence of evidence is not evidence of absence. BMJ 311:485

Evans MD et al (2009) Outcomes of resection and non-resection strategies in management of patients with advanced colorectal cancer. World J Surg Oncol 7:28

Goodman SN, Berlin JA (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 121:200–206

Miettinen OS (1985) Theoretical epidemiology. Delmar, New York, pp 1–359. ISBN 0827343132

Miettinen OS (2009a) Up from ‘false positives’ in genetic – and other – epidemiology. Eur J Epidemiol 24:1–5

Miettinen OS (2009b). Ziliak ST and McCloskey DN. The cult of statistical significance. How the standard error costs us jobs, justice, and lives (book review). Eur J Epidemiol 24:111–114

Nagakawa S (2004) A farewell to Bonferroni: the problems of low statistical power and publication bias. Behav Ecol 15:1044–1045

Perneger TV (1998) What’s wrong with Bonferroni adjustments. BMJ 316:1236–1238

Rothman KJ (1990) No adjustments are needed for multiple comparisons. Epidemiology 1:43–47

Rothman KJ (2010) Curbing type I and type II errors. Eur J Epidemiol 25:223–224

Sterne JAC, Davey Smith G (2001) Sifting the evidence – what’s wrong with significance tests? BMJ 322:226–231

Statistical Modeling

24

Jan Van den Broeck, Lars Thore Fadnes,
Bjarne Robberstad, and Bandit Thinkhamrop

All models are wrong, some models are useful.

G.E.P. Box

Abstract

This chapter starts with a brief essay that reminds us of the nature of statistical modeling and its inherent limitations. Statistical modeling encompasses a wide range of techniques with different levels of complexity. We limit this chapter mostly to an introductory-level treatment of some techniques that have gained prominence in epidemiology. For more in-depth coverage and for other topics we refer to other didactical sources. Every health researcher is likely, at some point in her career, to make use of statistical smoothing techniques, logistic regression, modeling of probability functions, or time-to event analysis. These topics are introduced in this chapter (using Panel 24.1 terminology), and so is the increasingly important topic of cost-effectiveness analysis.

24.1 The Nature of Statistical Modeling

Some teachings of the ancient Greek philosopher Plato can be interpreted nowadays as meaning that the essence of learning is modeling. Ideas and concepts are ways and tools of thinking about the complex realities we all try to make sense of.

J. Van den Broeck, M.D., Ph.D. (✉) • L.T. Fadnes, M.D., Ph.D.
B. Robberstad, M.Sc., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

B. Thinkhamrop, Ph.D.
Department of Biostatistics and Demography, Faculty of Public Health,
Khon Kaen University, Khon Kaen, Thailand

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_24,

This is true whether the complex information comes to us in the form of a storm of impressions and sensations, or carefully collected data in a research study. In each of those cases there is a vital need to model i.e. to reduce, recognize a pattern, summarize that pattern and classify it, so as to be able subsequently to communicate rapidly, compare, predict, react and…learn to survive. To stretch the point, the 'professional survival' of the epidemiologist or biostatistician depends on the ability to model data. At the most basic level, each measurement value is a model of the observation. At a further level, still very basic, a median or a mean are nothing but models that describe a set of data.

Panel 24.1 Selected Key Concepts and Terminology in Statistical Modeling

Beta coefficient In a linear regression equation, the slope coefficient of an independent variable, expressing the estimated change in the value of the dependent variable for a unit increase of the independent variable in question; In case the independent variable is categorical, the beta coefficient can be seen as the mean difference of the dependent variable between each category of the independent variable.

Covariates Variables entered in the regression model as independent variables

Dependent variable Outcome variable in a regression model

Explanatory variable A determinant as represented in a statistical model

Extraneous factors Factors that need to be controlled for in the study of a determinant – outcome relation by introducing them as covariates in a model of that relation

Independent variable In a regression analysis, a variable that independently determines ('predicts') the dependent variable i.e. independently from the effect of *other* independent variables in the model

Indicator variable A categorical variable with two categories

Interaction term A product term of two independent variables included in a linear regression model for the study of effect modification

Intercept Value of a dependent variable when all independent variables take the value zero (*Syn:* alpha coefficient)

Least squares regression methods Regression analysis methods that base the estimation of the regression parameters on minimization of the sum of the squared residuals around the regression line

Linear regression Regression analysis using methods that are based on the assumption of a linear relationship between the dependent variable and the independent variables; 'Linear relationship' (*syn.* Linear trend) meaning that each unit increase in each independent variable corresponds to a fixed increase in the continuous dependent variable

Logistic regression Linear regression of the natural logarithm of the odds for the outcome on independent variables

(continued)

Panel 24.1 (continued)

Multicollinearity The presence of highly correlated independent variables in a multiple regression resulting in difficulty in estimating the independent effect of these variables

Multiple linear/logistic regression A linear/logistic regression involving a model with several independent (explanatory) variables (*See:* linear/logistic regression analysis)

Parsimonious regression model A regression model that only includes independent variables that contribute substantially to the explanation of the outcome variable

Poisson regression Linear regression of the natural logarithm of the incidence rate on independent variables

Prevalence modeling Development and validation of a regression model in which the dependent variable is a prevalence (as empirically observed in a series of groups)

Product term *See:* Interaction term

Regression coefficients Intercept and beta coefficients

Residuals Differences between observed values of the dependent variable and those predicted by the regression model

Simple linear regression Regression analysis using methods that are based on the assumption of a linear relationship between the dependent variable and one independent variable

Stepwise multiple regression A multiple regression in which candidate independent variables are entered (forward) or removed (backward) one at a time based on chosen criteria for their contribution to explaining overall variance of the dependent variable

Trend Modeled shape of relation

When reflecting on these very basic forms of modeling one can already understand one of its fundamental characteristics: statistical models are by nature simplifying summaries, ‘reductions’, and there can be various degrees of simplification or ‘smoothing’, with more smoothing implying less well fitting models. At a slightly higher level of modeling we find, for example, the Gaussian distribution (Normal distribution) function, which models the frequency behavior of many phenomena under the form of a simple mathematical function of a mean and a standard deviation. The importance is not that Gauss recognized and described this shape of distribution in the data he was working with. The importance lies in the fact that this general form, this theoretical model, seems to fit reasonably well with frequency data of various sources. It is recognized as being a general shape applicable to many situations, with variable ad-hoc means and standard deviations.

Statistical modeling is done both in statistical estimation and statistical testing. When the same general shape of frequency distribution is thought to fit two different groups of ad hoc data, say data from a group of males and from a group of females, one basis is present for parametric statistical testing. Even in non-parametric ('distributional assumption-free') statistical testing modeling is done, since for each non-parametric test it is assumed that there is 'a statistic' that is able to capture (model) the degree of discrepancy between two frequency distributions. Thus, statistical estimation and statistical testing, the topics of the two previous chapters, can both be seen as based on statistical modeling.

One level up from simple distributions and their comparisons, linear and logistic regression models, frequently used in epidemiology, try to give a useful summary description of the relationship between a dependent and one or more independent variables. And there are higher levels of statistical modeling. At each level the challenge is to find a balance between model parsimony (smoothing) and model fit, a balance to be chosen, among others, upon considerations of usefulness, be it usefulness for summarizing in an understandable way, for testing, or for predicting. This balance may depend on the particular case and purpose. To clarify this point: a single median may be sufficient as a very rough but simple (unsmoothed but parsimonious) model of the distribution of a set of data. In another instance, a more complex, very well fitting, mathematical function may be needed to describe the same data's distribution. Particular paradigms for statistical modeling in epidemiology tend to differ according to the main types of studies.

24.2 Describing and Comparing Trends

In epidemiology one frequently obtains repeated estimates (e.g. repeated prevalence estimates, means and standard deviations) for the same target population over time. Or, one obtains these estimates across levels of another variable such as across successive age categories. For the description of these findings the following questions are commonly asked: Is there a trend and what is the shape of the trend? Is it a linear or a non-linear trend? Is the trend increasing or decreasing? Is the non-linear trend asymptotically flattening off towards a plateau value?

24.2.1 Linear Trends

Simple linear regression is often used to examine and depict a linear trend between two numerical variables. This is illustrated in Fig. 24.1.

24.2.2 Non-linear Trends

Non-linear modeling is typical in pharmacokinetic studies, for example studies looking at plasma levels, clearance rates, and adherence studies of accumulated substance concentrations levels in hair or nails. Non-linear modeling is also common in ecological studies and in case series studies.

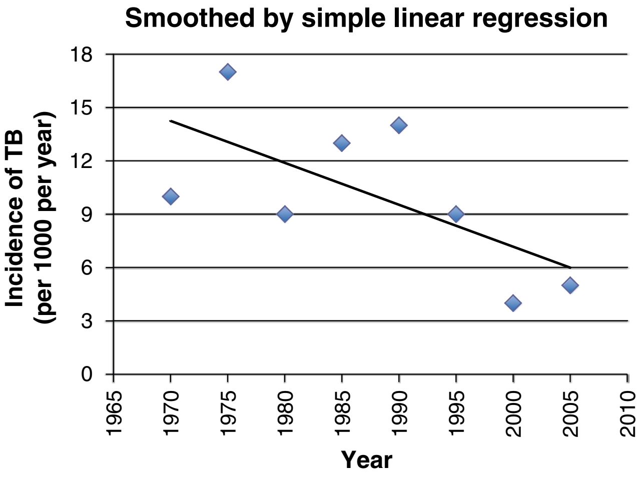


Fig. 24.1 An example of a linear trend line estimated by simple linear regression

A range of smoothing methods is available to depict non-linear trends. With all these methods it is possible to choose a degree of smoothness. Some are now easily accessible even in standard spreadsheet applications such as Excel. Commonly used statistical smoothing methods include:

- *Moving average* methods – This is a family of smoothing techniques characterized by the fact that each observed value is replaced with the weighted mean of a set of values that includes a specified number of successive values preceding and following the original value as well as the original value itself
- Fitting of *polynomials* – This method preserves continuous covariates as continuous (without categorizing) and transforms them to the nth order of power or degree while fixing the current functional forms of all other covariates. All combinations of powers are fitted and the 'best' fitted model is obtained
- *Cubic spline* fitting – Cubic splines are curves consisting of a series of third-order polynomials smoothly attached together

When different smoothing methods are applied to the same data, they can produce rather different looking trend lines, especially when the number of repeat estimates is low. We illustrate this using an example from a hypothetical ecological study that obtained repeated estimates of tuberculosis incidence in a number of calendar years (Fig. 24.2).

The examples in Fig. 24.2 illustrate two important points about smoothing methods and trend lines. First, it is clear that when the raw data are sparse and looking at the pattern ('visual smoothing') does not suggest a smooth underlying pattern (only 8 time points and a hectic visual pattern in the example) great differences in trend lines can be obtained with different smoothing methods. Secondly, when there are missing data points in between estimates, the trend lines shown can be very biased. Also note that the shape of the curve is particularly under the influence of the extreme points on the time scale ('edge effect') and that the flatness of the curve

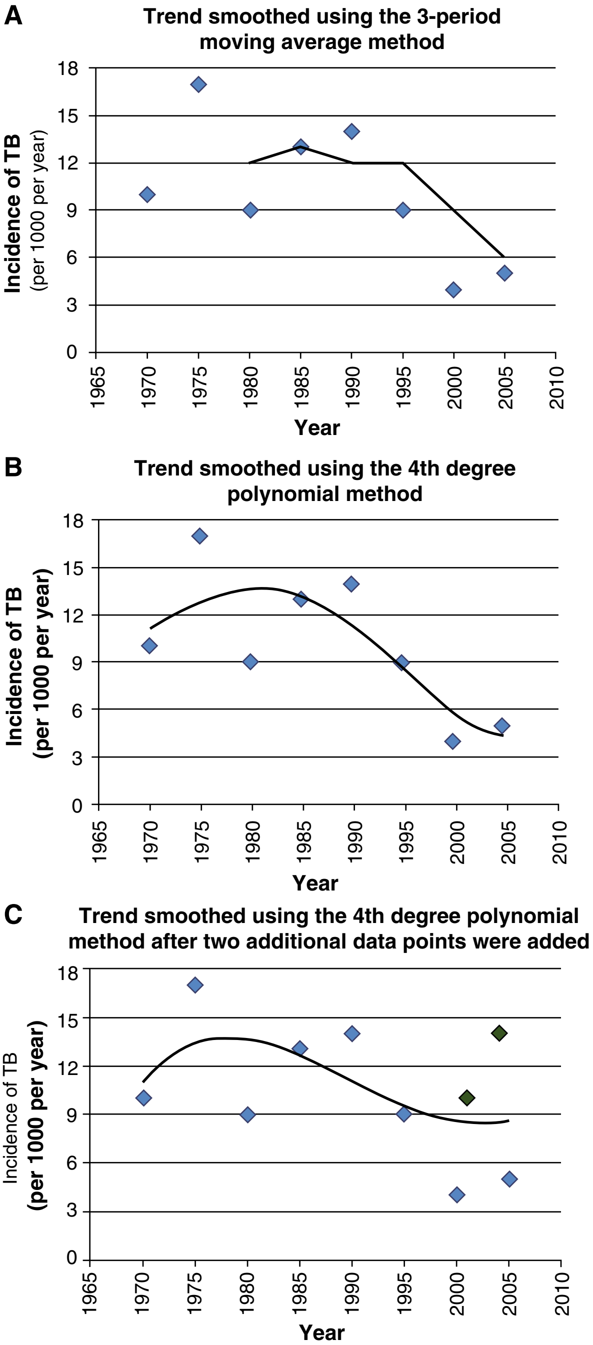


Fig. 24.2 Hypothetical data from a biased ecological time trend study. In *Panel A* and *Panel B* two different smoothing methods are used to describe a time trend in incidence of tuberculosis. Each *diamond* represents incidence in a single year period. Incidence rates in in-between years are unknown but an assumption was made that the pattern seen in the sampled years would be adequate enough to approximately describe the true underlying trend. In *Panel C*, the same 4th degree polynomial fitting method is used as in *Panel B* but after addition of two extra data points

depends not only on the chosen degree of smoothing but also on how far the Y axis is stretched. An additional problem with the ecological study data of Fig. 24.2 is that precision of the estimates in the different years is not shown. In some of the years precision may be much higher than in other years. The use of weighting during smoothing can allow for the fact that precision may be different for different estimates at different age/time points. Finally, when interpreting trend lines from ecological time trend studies one should always consider the possibility of bias resulting from changing measurement methods over calendar time.

24.2.2.1 Growth Diagrams

So far we have mentioned and illustrated situations where a single trend line (e.g. a line representing successive means, prevalence estimates or incidence rates or ratios) is assumed to adequately capture the central tendency of the age or time changes. However, non-linear trends in entire distributions can also be of interest, for example for the construction of age-sex dependent reference values of laboratory or anthropometrical data (*See also:* Panel 24.2). In such studies the challenge is indeed to obtain accurate and precise estimates for the extremes of the distribution, not only the central tendency. Describing age trends in continuous variables can also be of interest merely for the description of sample characteristics.

Two categories of methods exist to describe trends in entire distributions of continuous variables: those that are based on distributional assumptions and those that are not. We will not expand on the latter. Among the former, Box-Cox power-exponential modeling has become one of the major methods and was applied, for example, in the WHO-MGRS study (Borghi et al. 2006). In situations where

Panel 24.2 Terminology Related to Growth Modeling

Growth diagram A graph of the age or time trend in the distribution of a continuous variable representing a constitutional characteristic, displaying lines that represent selected centiles or Z scores

Growth modeling Construction of a model for the distribution of a variable as a function of a time variable

Growth reference A graphical and/or tabulated and/or statistical model based representation of the (usually smoothed) age/time dependent distribution of a continuous variable representing a constitutional characteristic, considered useful as a reference to score and classify individual measurement values as to their position within the distribution

Growth standard A growth reference considered to be normative i.e. representing the limits of what is considered normal or healthy (for example, in anthropometry a growth standard is considered to represent the distribution of growth unconstrained by illness or malnutrition)

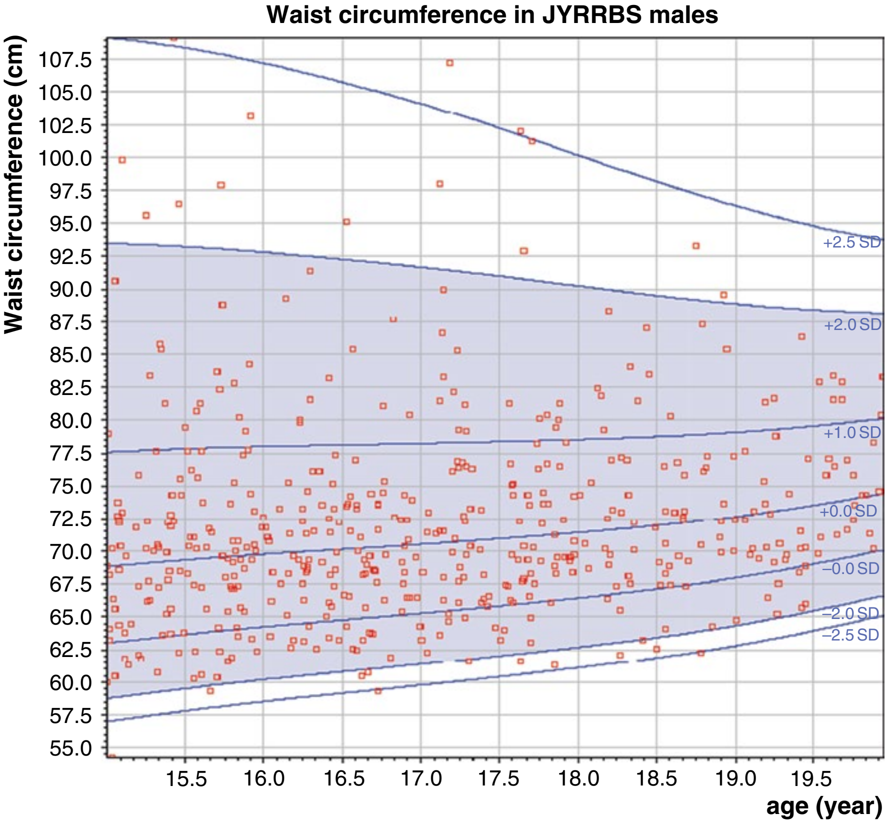


Fig. 24.3 A body mass index-for-age distribution constructed with the LMS method and cubic spline smoothing using the Growth Analyzer package (Reproduced from Francis et al. 2009)

kurtosis is of no concern, which is very often the case in growth studies, the LMS method is simpler to apply than Box-Cox power-exponential modeling (Cole and Green 1992). We recommend using the Growth Analyzer package for easy and widespread implementation of this method (Growth Analyzer 2009), for example for the description of sample characteristics. As an example, Fig. 24.3 shows a body mass index by age diagram constructed using this application.

24.2.3 Comparing Trends

In ecologic time-trend studies one often compares trend lines of exposure levels and outcome frequency by plotting them on the same graph. The construction of this type of graphs requires taking into account the latent period between first exposure and illness detection. The exposure curve needs to be lagged by the average latent period. Figure 24.4 illustrates this. This lagging is also needed before analyzing correlation.

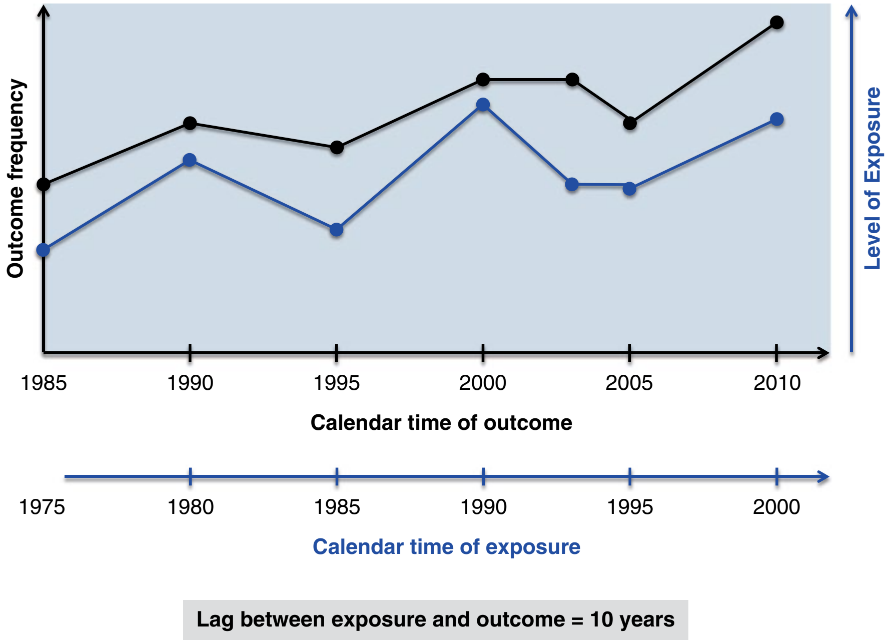


Fig. 24.4 Unsmoothed trend plot with lagged exposure time scale from a hypothetical cancer study in which the induction period between time of exposure and outcome is estimated to be 10 years

24.3 Logistic Regression Analysis

In epidemiology logistic regression is used frequently in the analysis of several types of studies, including and most commonly in cross-sectional studies, case–control studies and other etiologic studies. In this section we first introduce the basic statistical aspects of the commonly used binary logistic regression models (omitting the more rarely used ordinal and polytomous logistic regression). Next we give practical advice on how a typical and simple logistic regression analysis is performed in an etiologic epidemiological study.

24.3.1 Binary Logistic Regression Models

The use of logistic regression requires that the outcome is a categorical phenomenon, and the technique is therefore classified under categorical data analysis techniques (e.g., Thinkhamrop 2001). Indeed, in epidemiology the outcomes are often categorical, more specifically binary (e.g. presence of disease or disease outcome: yes/no). The use of binary logistic regression analysis assumes that exposure variable and covariates are linearly related to the ln-odds of the binary outcome.

That assumption is usually a fair one. The basic model is thus of a general linear form. It is called a simple logistic regression model:

Simple Logistic Regression Model

$$\ln(\text{odds}) = a + bx \tag{24.1}$$

Where:

ln = natural logarithm

odds = odds of outcome = p/(1 − p)

x = exposure variable

a = intercept (also called alpha coefficient)

b = slope coefficient (also called beta coefficient)

Note that in the basic model the dependent variable is a transformation of the outcome variable. That transformation is also called a 'logit-transformation': ln (p/(1 − p)).

In the basic model "a" and "b" are coefficients whose values are to be estimated from the data. Such estimation amounts to "fitting the model" to the data. The "b" coefficient, further called the beta coefficient, represents the size of the effect of the x variable (the exposure variable). It represents the change in logarithm of the odds associated with a one-unit change in x. The coefficient "a" is a fitted constant, also estimated from the data, representing the logarithm of the odds for a person with x = 0 (unexposed).

The main reason for the success of logistic regression in epidemiology is that the estimated beta coefficient can be transformed into an odds ratio by simple exponentiation i.e. by raising the natural number e (~2.71) to the power of the beta coefficient: odds ratio = e^b. The reason for this is explained in Textbox 24.1.

Textbox 24.1 The Exponent of the Beta-Coefficient Is a Point Estimate of the Odds Ratio

Basic model: $\ln(\text{odds}) = a + bx$

For the exposed (x = 1): $\ln(\text{odds}_{exp}) = a + b$

For the unexposed (x = 0): $\ln(\text{odds}_{unexp}) = a$

Thus: $b = (a + b) - a = \ln(\text{odds}_{exp}) - \ln(\text{odds}_{unexp})$

Thus: $b = \ln(\frac{\text{odds}_{exp}}{\text{odds}_{unexp}})$

Thus: $b = \ln(\text{odds ratio})$

Thus: $\textbf{odds ratio} = \mathbf{e^b}$

This odds ratio can be roughly interpreted as a relative risk of the disease outcome for a one-unit change in x (when the disease outcome is not common). This interpretation is more correct the rarer the outcome is. For example, in a case–control study of the effect of previous smoking (yes/no) on lung cancer an odds ratio of 9 could be roughly interpreted as meaning that the risk of lung cancer in smokers is 9 times the risk in non-smokers. When the outcome is common, however, the odds ratio always overestimates the relative risk. In such case, the interpretation of odds ratio as relative risk should be cautious (Zhang and Yu 1998).

When additional independent variables are introduced in the logistic model (these additional variables are called 'covariates'; they represent other exposures and/or confounders), the assumption is still that each of these covariates is linearly related to the logit of the outcome. And, again, this assumption can be considered fair in most circumstances. The model is then called a multiple logistic regression model:

Multiple Logistic Regression Model

$$\ln\left(\text{odds}\right) = \text{a} + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \tag{24.2}$$

Where:
odds = odds of outcome = p/(1 − p)
p = probability of the outcome
x_1 = exposure variable of main interest
x_2 to x_k = covariates, representing additional exposures and confounders
a = intercept (also called alpha coefficient)
b_1 to b_k = slope coefficients (also called beta coefficients) representing the independent effects of the corresponding x

Importantly, each beta coefficient represents the size of the *independent* effect of the corresponding x variable. When fitting the model to the data, each estimated e^b will now be an odds ratio that can be roughly interpreted as a relative increase in risk of disease associated with a one-unit change in the corresponding variable x, *independently of confounders* or other exposures in the model. Therefore any odds ratios obtained from this multiple logistic regression analysis can be called 'adjusted odds ratios.' Indeed, multiple logistic regression analysis is an approach to adjust for confounding effects of other variables (*See*: Chap. 22).

Confidence intervals can be obtained for the (adjusted) odds ratios, based on the standard errors for the corresponding beta coefficients, which the model-fitting method yields automatically. In multiple logistic regression analysis one can test the statistical significance of each variable's contribution to the overall model fit, by testing whether the corresponding beta coefficient is statistically significantly different from zero. A Wald type test is used for this purpose (*See:* Chap. 23). One can also test for modification of the effect of x_1 by x_2. This is done by: (1) creating a new variable,

x_3, as the product of x_1 and x_2; (2) fitting a model which contains the independent variables x_1, x_2, *and* the product term x_3; (3) determining the size and statistical significance of the coefficient b_3, which reflects the magnitude of effect modification. In the situation where the main aim is not assessing the relative outcome risk but prediction, confounding and testing of effect modification are not usually of interest. Thus, the investigator should be clear about the main objective before fitting the model, whether there is a risk factor assessment goal (with or without an interest in effect modification), a prediction goal (Kleinbaum and Klein 2002), or both.

24.3.2 How to Do a Logistic Regression Analysis

This sub-section looks at logistic regression analysis from a process point of view so as to guide the novice analyst through a simple and typical analysis in a step-by-step fashion.

Panel 24.3 lists the commonly needed steps, which are elaborated on below.

Step 1: Dataset creation and initial exploration

The analysis starts with taking a look at the analysis plan as foreseen in the study protocol (*See:* Chap. 13) and carrying out the preparatory steps of data extraction, computation of derived variables, exploration of univariate distributions, final data cleaning (*See:* Chap. 20), and arranging access to a functional statistical analysis package. One reminds oneself of the basic statistical choices that were made e.g. about what type of interval estimates will be obtained; typically, this will be 95 % confidence intervals obtained from the standard error of the estimated beta coefficient. The measurement level of each foreseen study variable should be clearly identified (*See:* Chap. 5). Of uttermost importance is checking if the outcome variable is truly binary.

Panel 24.3 Logistic Regression Analysis in Ten Steps

- Step **1**: Dataset creation and initial exploration
- Step **2**: Crude analysis
- Step **3**: Deciding which confounders to adjust for
- Step **4**: Examining effect modification
- Step **5**: Avoiding model over-fitting and addressing multicollinearity
- Step **6**: Further model reductions
- Step **7**: Assessing model adequacy
- Step **8**: Obtaining the final point and interval estimates
- Step **9**: Summarizing the findings
- Step **10**: Interpreting the findings

Step 2: Crude analysis

The analysis plan foresaw one or several exposures of interest and these may have different measurement levels. Continuous exposure variables can be used in logistic regression but using them as such only makes sense if the variable is more or less linearly related to the outcome. This needs checking at this point, perhaps by categorizing the variable and plotting the probability of outcome for each category. If it appears that the relation has a clear U-shape, J-shape or another non-linear shape, then the inclusion in the model as a continuous variable becomes problematic because the effect estimate of the exposure will tend to be diluted. There are two main options to solve the problem. First, some power transformation may be applied to the exposure variable (*See:* Chap. 13) and this new variable can be used as such or added as an additional exposure variable. Second option is to categorize the continuous variable into a number of categories carrying contrasting outcome frequencies. One category will then be chosen as the reference category and all other categories as index categories of the exposure.

With the exposure variables appropriately defined as above, one proceeds to the stage of crude (i.e. unadjusted) analyses of the relation between exposures and outcome. This step is also called ‘bivariate analysis’ as each time only one exposure variable is related to the one outcome variable and thus only two variates are involved. Crude analysis may use:

- Exposure odds ratios obtained from simple 2×2 tables, interpretable, when the disease is rare, as an approximation of crude relative risk of disease (For the calculation *see:* Chaps. 2 and 22)
- Odds ratios obtained from single-predictor logistic regression analysis. With this method the point estimate of the odds ratio is the exponent of the estimated beta coefficient and confidence intervals are calculated as in the box below
- With case–control designs that use a suitable sampling scheme for the controls and with the single etiologic study it is possible to obtain direct estimates of crude relative risk or crude incidence rate ratio (*See:* Chap. 22)
- Chi-square test are sometimes used for crude analysis of binary exposures in genetic studies

Confidence Interval for the Odds Ratio Obtained by Logistic Regression

$$95\%\ \mathrm{CI} = \mathrm{e}^{b \pm 1.96 * \mathrm{SE}}$$

Where, produced by the statistical package:

b = estimate of the beta coefficient

e = the natural number *(~2.71)*

SE = standard error of the estimated beta coefficient

Step 3: Which confounders to adjust for?

The crude odds ratios or incidence rate ratios obtained in step 2 may in fact represent some mixing of effects with extraneous factors (also called third factors or confounders) i.e. they may be distorted by confounding. The analysis plan foresaw a number of confounders to be measured and used for adjustment during analysis. Now it comes to making a final decision as to exactly what variables should be adjusted for in the multiple logistic regression analysis. This task can be approached by considering the following four questions:

Question-1: – Was the factor considered a potential confounder in the study protocol?

- Some factors are already known to be confounders of the relationship under study, and *must* therefore be further considered
- Some factors are known to be intermediates in the causal chain (mediators) and not confounders, and *must not* therefore be further considered as confounders

Question-2: – Were any potential confounders forgotten in the study protocol?

- New literature on risk factors may have become available since the time the study protocol was written; This is a frequent issue in studies with a prospective study base
- When trying to identify confounders one should look extra carefully for potential confounders that belong to the same 'type of exposure' as the exposure under study. For example: *other* nutritional factors, *other* risk behaviors, *other* environmental contaminants, et cetera…(Miettinen 1985)
- If any 'new' potential confounders are identified, where they measured in the study? If not, is there a proxy available?

Question-3: – Was the suspected confounder not already controlled for by design?

- The general rule is that one should not control for characteristics for which restriction or matching was successfully applied

Question-4: – Did the factor actually act as a confounder?

- There is no need to adjust for factors that did not act as confounders
- Some of the factors that were considered initially may unexpectedly turn out to be very rare. Perhaps only one or a few participants exhibited the potentially confounding characteristic. In this case the confounding effect is likely to be small and it can be an option to assess the effect of excluding these participants' data from the analysis and discuss this in the presentation of the data
- There are two ways of checking whether a variable actually had a confounding effect. The first is to check if the factor complied with the well-known confounding criteria (*See:* Chap. 2). The second way is called 'the distortive impact approach'. Both approaches are described successively below

To check confounding criteria the first question is whether the third factor independently predicts the outcome. To answer this, one can examine the relationship between third factor and outcome in an *analysis among the unexposed* that adjusts for other confounders. This analysis should be done in the unexposed only, to avoid confounding or mediation by the exposure. If a relationship is found then the first criterion for confounding is satisfied. Secondly, to have acted as a confounder, the third factor must be associated with the exposure. If the distribution of the third factor is different across levels of the exposure then the second criterion for confounding is fulfilled. Note that a strong confounder does not need much imbalance to exert its confounding effect and therefore statistical testing is not useful in this situation. If in doubt, one should control in the analysis. The third criterion is that the third factor should not be an intermediate of the effect of the exposure on the outcome. Although methods of mediation analysis exist, the assessment of this third criterion can often be done using common sense judgment considering the prevailing conceptual framework around the topic.

The distortive impact approach checks the extent to which, for example x_2, confounds the association between x_1 and disease outcome, comparing the outcome parameter estimate (the value found for the odds ratio, relative risk or incidence rate ratio) for x_1 in two models: one which includes x_2, and one which omits x_2. If these estimates are similar, then x_2 is not an important confounder. Usually when there is confounding, the effect estimate will become closer to the null effect (closer to 1) after adjustment in analysis. This means there was positive confounding. If the outcome parameter estimate completely returns to the null effect after control, then the crude effect was entirely due to confounding by the third factor. When the outcome parameter estimate shifts further away from the null effect after adjustment in analysis, this means there was negative confounding.

In this section we take it for granted that, after selecting the variables to control for confounding, multiple logistic regression analysis is the chosen method to do so. However, remember that, to control for confounding, an alternative to multiple logistic regression analysis is stratified analysis with pooled estimation (*See:* Chap. 22). That method is now rarely used in practice because of frequent problems of lack of sample size in individual strata. Knowledge of illnesses has progressed so much that more and more risk factors are known for each illness or illness outcome. Consequently, a large number of confounders often have to be adjusted for in etiologic studies and this can make the use of stratified analysis inefficient or impossible; the number of (sub-) sub-strata becomes too large. Consequently, multiple logistic regression analysis has become one of the major statistical methods in etiologic research. Yet, if stratified analysis is feasible it can be good to use it alongside multiple regression analysis and check the consistency of the findings.

Step 4: Examining effect modification

After deciding which confounding factors to adjust for, it can be good to examine effect modification (*See:* Chap. 2: Basic concepts in epidemiology) by some of the covariates. This means examining how the strength of the exposure-outcome relation (the value of the odds ratio) depends on the levels of the covariates. We can examine this by including ‘product terms’, a.k.a. ‘interaction terms’ into the model and examining the beta-coefficients of these product terms. A product term is a multiplicative form of two or more variables. For example, the model can include the following independent variables: x (exposure variable), z (covariate), and $x*z$ (product term). When doing this, the variables that form the interaction/product term are also included in the model as single terms. Interaction terms usually only concern two variables (the exposure and one covariate). Product terms incorporating more than two covariates lead to odds ratios that are very complicated to interpret.

In addition, the more interaction terms are included in the model, the more likely there will be problems of multicollinearity (next paragraph), hence, instability of the model. Thus it is recommended that the number of interaction terms should be kept to a minimum. Examination of effect modification can be considered if there is a rationale that a particular covariate might modify the strength of the exposure-outcome relation and if there is an interest in showing this.

Step 5: Avoiding model over-fitting and addressing multicollinearity

Up to this step, we have defined a full model that contains all covariates that could possibly affect the outcome and may remain in the final model. These covariates include all potential confounders identified in Step 3 and perhaps one or two interaction terms identified in Step 4. It is possible to include too many variables. Two common situations indicate a need to reconsider the number of covariates in the model.

The first is a situation in which the full model contains too many covariates relative to the amount of information (sample size and number of outcome events) in the sample. This is a cause of *model over-fitting*, also called *over-parameterization*. This is characterized by the fact that relatively high values of the dependent variable are over-estimated and relatively low values of it are underestimated. Thus, in studies where regression models are used to derive probability functions (*See:* Sect. 24.4), the predicted values will be biased when there is model overfitting. To recognize whether overfitting is present one can use the following rule of thumb (Miettinen 2011a):

Model Over-Fitting

Models are considered to be over-fitted if:

$$0.05 < \frac{P_s}{N * p(1-p)}$$

Where:

P_s = number of parameters in the model
N = sample size
p = proportion of individuals with the outcome

The approach to handling model overfitting may be a reduction of the number of covariates (for example by a decision to forego examination of effect modification; also see below for model reduction strategies to address multicollinearity). Alternatively, one can apply ‘shrinkage techniques’ to the model (in Step-8), for example by using the leave-out-one method (not further discussed). A third approach could be to aim for pooled analysis with data from similar studies in a meta-analysis.

A second problem can arise when there is *multicollinearity* i.e. strong correlation among covariates. This is a cause of model instability: the beta coefficient can change

Textbox 24.2 Methods for Recognizing Multicollinearity

The following can be signs of **multicollinearity**:

- Two or more covariates measuring conceptually the same thing are included in the model, a.k.a. covariates redundancy
- Large *variance inflation factor* (VIF = 1/(1 − R^2), where R^2 is the coefficient of determination of a beta coefficient of a covariate on all other covariates; A VIF of 10 or above indicates a multicollinearity. This is equivalent to a Tolerance, defined as 1/VIF, of 0.1)
- Drastic changes of the beta coefficients when a covariate is entered or removed or when a subset of data is used
- A large condition number (an index of the global instability of the beta coefficients). A condition number of 30 or more is an indication of model instability
- A high correlation coefficient as indicated in a pair-wise correlation matrix among all covariates; A coefficient of 0.8 and higher requires further investigation for multicollinearity
- Significant results of an overall Chi-square test whereas none of the individual Wald test were significant

dramatically in response to small changes in the model or in the data. The more multicollinearity there is the greater will be the standard errors of the beta-coefficients. Multicollinearity misleadingly inflates the standard errors. The confidence interval of the beta coefficients will tend to be too wide and the P-value of Wald tests falsely large. In practice the existence of multicollinearity is often examined by checking if any two covariates have a correlation that exceeds a correlation coefficient of 0.8. However, that is only a rough method. Textbox 24.2 describes a more detailed examination method.

The following are approaches to handle multicollinearity:

- Carefully select covariates to enter into the model; avoid redundancy
- It may be possible to derive a new variable based on two or more related variables then use this variable for modeling. For example, in the model we may want to use only body mass index which is derived from height and weight
- Use factor analysis or some other methods such as propensity scoring to create one scale from many covariates. Then use this variable for modeling
- In the process of model fitting, add multicollinearity-potential covariates into the model, then consider dropping the one that is less important in terms of the standardized beta-coefficient. If all covariates are important to the model, it is better still to keep it in the model but one needs to realize that multicollinearity is present and be aware of its consequences
- If an interaction term was added into the model 'centering methods' can reduce multicollinearity. By centering, we transform the original continuous variable to be the difference between its mean and the original value

Step 6: Further model reductions

So we have defined a *full model* that should be free from overfitting and less likely to have multicollinearity issues. Before assessing model adequacy and deriving the final estimates there can be another preparatory step. There may be a need to reduce the number of covariates so that a more ‘parsimonious’ model is achieved. That is, a model that only contains the covariates that provide important information about the outcome.

In most circumstances, the full model is the final model. All covariates are essential for the model and no further reduction is required. This is common in etiologic studies and in randomized controlled trials where there is a factor of main interest: the exposure or the treatment, and a well-defined set of factors to adjust for. However, such studies may have an additional interest in developing ‘prediction models’ (*See:* next section). In such analyses there is no factor of main interest, and the full model, when it contains many covariates, may benefit from further model reduction. The same applies to prognostic studies primarily set up with the specific aim of developing a prediction model or forecasting model (*See:* Chap. 6, Sects. 6.6.1 and 6.6.2).

The procedure of model reduction involves a strategy for comparing relevant models, based either on testing significance of the covariates, or on a comparison of estimates of the error variances, or on a comparison of the changes of the beta coefficient between the model with and without the covariates under assessment. There is no single method that is overall satisfactory; hence, a combination of these methods is recommended.

With ‘backward elimination’ variables are sequentially removed from the full model. At each step, the variable showing the smallest contribution to the model is removed or eliminated. To illustrate this, a backward elimination is described in Textbox 24.3.

Step 7: Assessing model adequacy

It is essential to assess model adequacy to assure that valid inferences can be made from the estimated beta coefficients. To do this, regression diagnostics need to be obtained first. This includes examination of residuals, leverage, and influence statistics, which we will not discuss further. Note that regression diagnostic plots can be useful for identifying observations that cause a problem to the model fitting. After examining these diagnostic statistics, one should assess the model goodness-of-fit. This examines how well the model describes the observed data and can inform about how sensitive the model is to certain individual observations. The most common method is the Hosmer-Lemeshow goodness-of-fit test (*See:* Chap. 23).

Step 8: Obtaining the final point and interval estimates

Once the final model is achieved, one can obtain the estimated values of the odds ratios. This can be obtained directly from the output of the statistical package. In a situation where there is an interaction effect, the odds ratio needs to be calculated separately according to subgroup of the effect modifier.

Step 9: Summarizing the findings

One should report the results of bivariate analyses to inform readers regarding evidence about potential confounders, multicollinearity, departure from linear

Textbox 24.3 A Backward Elimination Method

1. Fit the full model and obtain the log-likelihood of the current model, to be used as the reference for comparison with a subsequent reduced model.
2. Examine the Wald statistics of all interaction terms, if any, and select the term with the highest P-value for removal.
3. Fit the reduced model (without the term having the highest Wald test P-value), then calculate the Likelihood Ratio test. A rule of thumb is that if the term is significant with a P-value of less than 0.05 then it cannot be deleted, otherwise, the reduced model without that term is the one to be used for the next step. Repeat the process until no more interaction term is eligible for removal. If any interaction term is to be retained, all variables forming such term are not eligible for removal.
4. Examine the Wald statistics for all individual terms that are not a component of an interaction term retained in the previous step, if any, and then select the variable that is the least contributing to the model, i.e., the one with the largest Wald test P-value, to be the candidate for removal.
5. Fit the reduced model without the selected variable and calculate the likelihood ratio test. Follow the process of variable elimination described under 3. Repeat the processes until no more variable is eligible for removal.

Note: For the Likelihood Ratio test, one needs to examine the sample size being used to estimate the coefficients in each model while performing model reduction. If the sample size of each model being compared differs by a large amount of observation due to missing values, then the Likelihood Ratio test is not valid.

trend, and number of observations for each category of the covariates. Results of multivariable analysis are then presented. Usually these are presented as tables. Presentation as a forest plot is also possible.

Step 10: Interpreting the findings

For common pitfalls of interpretation we refer to Chap. 27. Also note that many interpret the odds ratio as if it is a risk ratio without taking into account that the odds ratio tends to overestimate the risk ratio when the outcome is common or when there is a strong association between the covariates and the outcome. When internal validity is high it can be relevant to interpret the magnitude of the odds ratio by comparing it with a meaningful level of association. As the rule of thumb, an odds ratio of 3 indicates a strong association (Tugwell et al. 2012). In the situation where there is an interaction effect, we report that the association between the variable of main interest on the outcome depends on the third variable – the effect modifier. Then we interpret the odds ratio separately for each subgroup of the effect modifier.

24.4 Probability Functions

24.4.1 Types and Uses of Probability Functions in Medicine

Probability functions are constructed frequently in support of clinical prognostication and for forecasting of population burdens. These are the main and best known current uses of probability functions in medicine (*See:* Chaps. 4 and 6). This alone would be a sufficient reason to devote a section on probability functions in the present chapter. However, the role of probability functions extends well beyond its main current use. They also have some role in diagnostic, etiologic and intervention-prognostic studies, both in the realm of clinical medicine and community medicine (*See:* Chap. 6). In support of clinical practice, for example, Miettinen (2011b) proposes the construction and use of three types of '*Gnostic Probability Functions*' (Miettinen 2011b: Up from Clinical Epidemiology and EBM). The three types of functions address the probability of an untoward event happening in the course of an illness, of an illness being present, or of a risk factor having played a causal role, as a function of prognostic, diagnostic or etiognostic indicators, respectively. They are therefore called:

- Diagnostic Probability Functions (DPF)
- Etiognostic Probability Functions (EPF)
- Prognostic Probability Functions (PPF)

In simple terms, the main tasks of a doctor are to tell the patient about the diagnosis, how the illness likely came about and what the prognosis is, with or without treatment. Appropriately fitting probability functions can be helpful for each of these tasks, not the least because probability functions constitute a quantitative approach that allows taking the specific individual patient profile into account when estimating the probabilities. Therefore, the modeling of DPF, EPF and PPF deserves some further introduction.

24.4.2 Diagnostic Probability Functions

Diagnostic probability functions model the probability of a particular illness being present in a presenting patient as a function of diagnostic indicators, which include components of the risk profile (e.g., socio-demographic factors) and of the manifestation profile (e.g., symptoms, signs, and test results). DPF can be incorporated into software that allows users to estimate diagnostic probabilities for presenting patients (Miettinen 2011b). Such functions also provide for the evaluation of new diagnostic tests. This can be done, among others, by comparing the predictive abilities of functions with and without the test result added as a predictor variable (Miettinen 2011a, b).

A DPF must be constructed with the purpose to apply to a specified type of presenting patients. In other words, there must be a restriction to a particular target population. For example, a DPF may help with the diagnosis of pneumonia *in children presenting with cough and fever*. The modeling uses multiple logistic regression, after which the fitted logistic regression model is transformed into a 'risk function':

Construction of a Diagnostic Probability Function

Step-1:
Using multiple logistic regression we model:

$$\ln\left(\text{odds}\right) = \text{a} + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \tag{24.2}$$

Where:
odds = odds of the illness being present = P/(1 – P)
x_1 to x_k = indicators of the risk profile and the manifestation profile
a = intercept and b_1 to b_k = beta coefficients of the corresponding x

Step-2:
We transform the fit model into a probability function:

Probability of the patient having the illness = P

$$\text{P} = \frac{1}{1 + e^{-\left(a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k\right)}} \tag{24.3}$$

Where:
e = the natural number *(~2.71)*

Two types of practical approaches are available for the fitting of DPF. A first option is to study a large sample of presenting patients whose profile fits with the definition of the target population (e.g. children presenting with cough and fever), gather information about risk profile and manifestation profile of each patient, and also about the later rule-in diagnosis made with gold standard methods. Another, often more efficient, type of approach, is to use a method similar to the method of Miettinen et al. (2008), further developed in Miettinen (2011b), which is based on giving expert diagnosticians a variety of hypothetical patient profiles and asking them to attach a probability of the illness to each of these fictitious scenarios.

24.4.3 Etiognostic Probability Functions

In clinical medicine, etiognostic probability functions express the probability of a particular exposure having played a role, given the patient's etiognostic profile. The etiognostic profile of the patient includes (1) other known risk factors than the exposure of main interest, and can also include (2) non-causal factors acting as effect modifiers and (3) specifics about the subtype of illness (Miettinen 2011b).

These functions can be constructed using data from observational etiologic studies. In essence, etiologic studies produce causal rate ratios (e.g. adjusted incidence rate ratio, adjusted odds ratio, adjusted relative risk or adjusted prevalence rate ratio). However, effect modification can alter these causal rate ratios. To make statements about what the probability was that a particular exposure causally acted in a particular patient, knowledge is needed about how the causal rate ratio depends on factors in the etiognostic profile. Once the applicable causal rate ratio for the particular patient is identified, the etiognostic probability for an individual patient is calculated as an attributable fraction (Miettinen 2011b):

Probability of the exposure having played a causal role in the particular patient = P_c

$$P_c = \frac{CRR - 1}{CRR} \tag{24.4}$$

Where:
CRR = **Causal rate ratio** (value applicable to the particular patient to be estimated using the model described below)

The modeling of the causal rate ratio itself is also described in Miettinen (2011b). Briefly, in studies where an incidence risk or prevalence rate is compared between exposed and unexposed the modeling involves the following:

<u>Step-1:</u>
Using multiple logistic regression we model:

$$\ln(\text{odds})_{Exposed} = \text{a} + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

Where:
odds = **odds of the outcome** being present **among the exposed**
x_1 to x_k = indicators of the etiognostic profile
a = intercept and b_1 to b_k = beta coefficients of the corresponding x

<u>Step-2:</u>
Using multiple logistic regression we model:

$$\ln(\text{odds})_{Unexposed} = \text{a} + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

Where:
odds = **odds of the outcome** being present **among the unexposed**
x_1 to x_k = indicators of the etiognostic profile
a = intercept and b_1 to b_k = beta coefficients of the corresponding x

Step-3:

From steps 1 and 2 we define a **function for the causal rate ratio** (CRR) as:

$$\text{CRR} = \frac{1+e^{-\ln(Odds)exp}}{1+e^{-\ln(Odds)unexp}}$$

Where:

e = the natural number *(~2.71)*

In studies where the outcome is an incidence density the modeling involves:

Step-1

Using Poisson regression we model:

$$\ln(\text{rate})_{Exposed} = \text{a} + b_1x_1 + b_2x_2 + \ldots + b_kx_k \tag{24.5a}$$

Where:

Rate = numerical value of the **incidence density among the exposed**

x_1 to x_k = indicators of the etiognostic profile

a = intercept and b_1 to b_k = beta coefficients of the corresponding x

Step-2

Using Poisson regression we model:

$$\ln(\text{rate})_{Unexposed} = \text{a} + b_1x_1 + b_2x_2 + \ldots + b_kx_k \tag{24.5b}$$

Where:

Rate = numerical value of the **incidence density among the unexposed**

x_1 to x_k = indicators of the etiognostic profile

a = intercept and b_1 to b_k = beta coefficients of the corresponding x

Step-3

From steps 1 and 2 one we define a **function for the causal rate ratio** (CRR) as:

$$\text{CRR} = \text{e}^{\ln(rate)exp - \ln(rate)unexp}$$

Where:

e = the natural number *(~2.71)*

24.4.4 Prognostic Probability Functions

Prognostic probability functions model the occurrence of a defined outcome event as a function of prognostic profile indicators. In clinical medicine the outcome event can be the occurrence of an illness or a particular outcome of an illness. The prognostic profile includes client/patient characteristics present at the time the prognosis is made, and aspects of personal history up till that moment. The occurrence probability of the outcome is studied:

- Either for a single defined period of time, or, for multiple points in prognostic time (smooth or segmented)
- Conditional on surviving

When the aim is to know about the event 'ever happening' in a defined time span, the modeling can use multiple logistic regression analysis, after which the fitted model is transformed into a prognostic probability function (a 'prediction model'):

Construction of a Prognostic Probability Function

<u>Step-1:</u>
Using multiple logistic regression we model:

$$\ln(\text{odds}) = \mathrm{a} + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

Where:
odds = odds of the event happening during the defined period = P/(1 − P)
x_1 to x_k = indicators of the prognostic profile
a = intercept and b_1 to b_k = beta coefficients of the corresponding x

<u>Step-2:</u>
We transform the fitted model into a probability function:

Probability of the event ever happening during the defined period = P_E

$$P_E = \frac{1}{e^{-(a+b_1x_1+b_2x_2+\ldots+b_kx_k)}}$$

Where:
e = the natural number *(~2.71)*

When the aim is to know about the event's cumulative occurrence at a multitude of time points, two main options are open for the modeling. Either one bases the prognostic models on survival analysis/Cox regression, for which we refer to the next section, or, one aims for the construction of smooth-in-time risk prediction functions, for which we refer to Hanley and Miettinen (2009).

24.4.5 Validation of Probability Functions

Distinction can be made between internal and external methods of validation (*See also*: Sect. 6.6.3 of Chap. 6).

24.4.5.1 Internal and External Validation Methods

Internal validation methods commonly made use of in prognostic studies include:

- Split-sample validation
- Cross-validation
- Bootstrap validation

In *split-sample validation* the sample is randomly divided in two groups, usually 50 % / 50 %. One group will be used to create the model, the other group to verify model performance. This classical approach has many drawbacks:

- Not all data are used for development/performance assessment
- Perceived 'unlucky splits' may lead to a temptation to do repetitions until a 'lucky split' arrives

In the *cross-validation* method the sample is randomly divided into equal size small groups e.g. 10 times 10 %. Successively, each of the small groups serves to validate a model developed with all other groups together. This is repeated about 50 times until stable results are obtained. It is a more cumbersome procedure, but has fewer drawbacks than the split-sample method.

In *bootstrap validation* one estimates the standard error and confidence interval of the outcome parameter estimates, in this case of the coefficients in the risk prediction model, obtained in a large number of random samples with replacement (of size N) drawn from the original sample (of size N). The samples are different because, each time, some of the original sample items are selected more than once and others are not selected. A risk prediction model, usually about 150–200 in total, is derived for each sample and its performance evaluated both on that bootstrap sample and on the original sample. This method allows (1) assessing whether the performance of the original model was overly optimistic, (2) correction for optimism. Many of the commonly used statistical packages can do bootstrapping.

External validation methods are based on the application of the risk model on subjects whose data were not used for model construction, for example more recent patients, patients from another site e.g. from another country, from another study.

24.4.5.2 Prediction Model Performance Parameters and Model Updates

Quantifying the performance of a prediction model may involve:

- Estimating the variance explained by the model and the size of the standard errors of the coefficients
- Accuracy of prediction can be assessed by comparing goodness of fit of predicted and observed risks, which can be done separately in – for example, quartile-defined – categories of predicted risk
- When the predicted risks can be meaningfully dichotomized into high e.g. requiring some form of action, or low, the area under the ROC curve is frequently taken as a parameter

Competing models are compared in respect of the above parameters. In such comparisons, however, the performance parameters need to be ultimately judged in the light of the relative parsimony of the competing models: when two models have about the same performance, the one with the least number of predictor variables tends to be preferred.

Prediction models, once applicable to a population, may require updating after some time. Making a model applicable to another population also requires re-calibration, mainly because the intercept tends to differ between populations. An extensive overview of issues and practical methods of developing validated clinical prediction models can be found in Steyerberg (2009).

24.5 Time-to-Event Analysis

24.5.1 Time-to-Event Analysis (or Survival Analysis)

Time-to-event analysis (or survival analysis) is an analytical strategy originally designed to assess time to death and can be used in e.g. prognostic studies. However, the analysis has much broader applications than to assess time to events such as diseases, time to vaccination etc., and is an essential statistical tool when the follow-up time of the participants are different and when the time of an event is of importance. Time-event-event analysis takes the time each person (or case) contributes into the denominator to assess the risk of an event (which could be death in a traditional survival analysis) at each time point. In time-to-event analyses (*See:* Panel 24.4), each person will at end of follow-up either have had a predefined event (e.g. death or studied disease) or be censored (not having had the predefined event yet). It is important to keep in mind that censored cases might experience the event at any time after the observation time is completed. This is however unknown to the analysis.

Panel 24.4 Terminology Related to Time-to-Event Analyses

Censoring Term used in time-to-event analysis to indicate that a participant did not experience the studied event (e.g., disease or death) at end of follow-up or when the participant was lost to follow-up

Cox regression A regression method using time-to-event analysis with the assumption of proportional hazard (see text)

Hazard ratio The parallel term to risk ratio in time-to-event analyses/ hazard regression

Kaplan-Meier table A table presentation of survival data

Survival function A fitted model for the probability of not having the outcome of interest (e.g., death) as a function of individual follow-up time and sometimes other variables as well

Time-to-event analysis (or survival analysis) An analytical strategy originally designed to assess time to death and can be used in e.g. prognostic studies

24.5.2 Life Tables and Kaplan-Meier Plots

Time-to-event analysis could be presented in e.g. a life table or with a Kaplan-Meier plot.

A life table is a table presentation of survival data where the time variable is first divided into smaller pieces, and the risk of event at each time is calculated by the number of events during the time period and dividing it by the number of cases at the start of the period. The Kaplan-Meier estimator is a commonly used time-to-event analysis, similar to a life table but using exact survival times among the cases to make time stratification, and can be used to make Kaplan-Meier plots. This will generally give slightly more precise estimates than the conventional life tables. See Fig. 24.5 for an example in how data can be presented with a Kaplan-Meier plot.

24.5.3 Time-to-Event Analysis for Retrospective Information Collected Cross-Sectionally

It is also possible and sometimes beneficial to utilize time-to-event methods for cross-sectionally collected data with retrospective information. As an example, when collecting data on breastfeeding duration during a vaccination visit e.g. at 15 months, not taking censoring into account will give sub-optimal estimates of the duration. Let us make an example to illustrate this with a group of 30 children. Ten of these children breastfed for 6 months, 10 children breastfed for 12 months and ten children breastfed for 18 months. If all these children were assessed at the age of 15 months (e.g. nested to a vaccination program), an analysis excluding children not having experienced the event would result in an estimate of median and mean breastfeeding duration of 9 months (95 % confidence interval of the mean is 7.6–10.4) due to the exclusion of the third with the longest duration. Thus, the duration estimates would have been biased to a shorter duration.

The true median and mean duration is 12 months. When using Kaplan-Meier survival analysis, the estimate for mean breastfeeding duration will be 11 months (95 % confidence interval 9.7–12.3) and 12 months for median breastfeeding duration. We can see that the 95 % confidence interval of the mean with *restricted analysis without* using *survival analysis* does not include the true value while survival analysis gives a good estimate. For such a study, the reliability and validity of the reported information is essential as recalled information might be a source of information bias.

24.5.4 Assumptions on Timing of the Events and Censoring

Even in a prospective study, it is often necessary to make some assumptions regarding time of the events. We might know that in a given study visit e.g. at 12 months of age, the studied event has not happened. Further, we might know that in the next

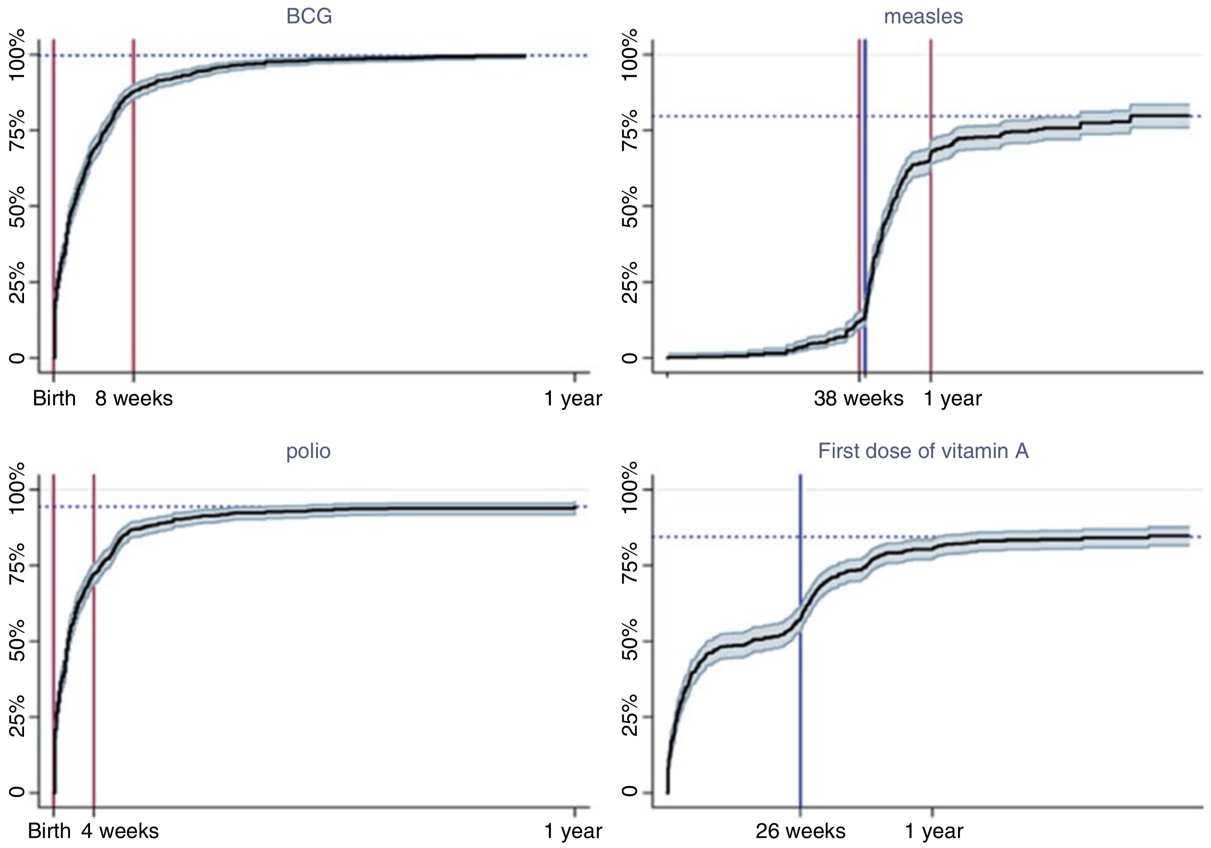


Fig. 24.5 An inverse-Kaplan-Meier plot used to assess timing of vaccinations. The y-axis indicates the proportion having received the vaccines at each time point. The x-axis indicates the age of the child when vaccine was given (For more details, *see:* Fadnes et al. 2011)

study visit e.g. at 16 months of age, the studied event has happened, but not exactly when it happened. In other words, it could be at any time between 12 and 16 months. This is often referred to as interval censoring (but will be indicated as an event and not as censored in the analysis). We have different choices for which time point the analysis should assume that the event happened on. One option is to use the mid-point assumption and assume that events in average have occurred in the mid of unknown time ranges, in other words at 14 months. Another option is to make a right-point assumption, assuming it took place at end of the range, in other words at 16 months in the example above. Both of these will often be regarded as imputation techniques, and might cause biased estimates. A third option is to estimate the timing based on a model taking information at the visits into account (e.g. linear trends). The larger the range of uncertainty is (which is often related to the frequency of visits for data collection), the larger will the importance of the choice how it should be handled be.

24.5.5 Time-to Event Regression Models

It is also possible to use regression analysis to assess determinants/predictors which are associated with the time-to-event. The most commonly used analysis methods are based on a proportional hazard model such as the Cox regression. One of the key assumptions is that the risk of event (hazard) is proportional at each time point for each value label of the assessed co-factors. E.g. when assessing the hazard (risk) for a cardiac event and comparing this between women and men, the hazard ratio (which is parallel to risk ratio in time-to-event data) between women and men at e.g. 50 years of age should be relatively similar to the risk-ratio between women and men at 70 years. If this assumption is severely violated, the Cox regression model might not be preferred, and other strategies such as log rank test could be considered. A log rank test is a chi square based test which takes the order of the events in the different strata into account, but not the exact timing. In the example above comparing gender and cardiac events, the rank of the age in each gender would be calculated for all cases in both men and women (the youngest age at the first cardiac event would be ranked 1; the second youngest age would be ranked 2 etc.). If the average ranks in each gender are significantly different from what is expected by chance, the time to cardiac event would be regarded as significantly associated with gender. In a Cox model, the time variable should also be on a continuous scale and censoring should occur randomly (which can be assessed with e.g. Martingale residual plots). Cox regression utilizes a gamma distribution. In addition, general assumptions for regression analyses are applicable also for time-to-event models.

24.5.6 Alternative Models for Time-to Event Regression

In some cases, time-to-event follow a more predictable pattern that can be modeled more exactly than the techniques described above. This could be when the time-to-event can be modeled with e.g. Weibull (e.g. gradually decreasing hazard), Exponential (constant hazard), Gompertz or Log-logistic distributions. These techniques will not be covered here. It is also possible to use time-to-event analysis to take several events into account in the same analysis. This book will not cover that.

For further introductory reading about time-to-event analyses, we refer to Breslow (1975), Lee and Go (1997), and Leung et al. (1997).

24.6 Cost-Effectiveness Analysis

A wide range of methods for economic evaluation are used, and the common feature is that they simultaneously consider both intervention costs and associated health outcomes. These concepts were introduced separately in Chap. 10, but in this chapter we consider them jointly and introduce techniques for addressing whether health care interventions can be considered to be cost-effective (*See also:* Panel 24.5).

24.6.1 Taxonomy of Methods for Economic Evaluation of Health Interventions

Assessment of cost-effectiveness of an intervention must always be made with reference to a specified *comparator* intervention. In other words, an intervention can only be cost-effective, or not, compared to something else. The comparator is typically the *current standard of care*, but *best alternative practice* or *no treatment* alternatives are also commonly used as comparators.

Panel 24.5 Selected Terms Relevant to Cost-Effectiveness Analysis

Average Cost Effectiveness Ratio (ACER) The ratio of costs and effectiveness of an intervention compared to an implicit alternative intervention (often "no treatment").

Comparator The intervention being included for comparison in an economic evaluation (e.g., current standard of care, best alternative practice, no intervention)

Cost Benefit Analysis (CBA) Economic evaluation in which both costs and outcomes are expressed in monetary terms

Cost Effectiveness Analysis (CEA) Economic evaluation involving the use of simple natural units as outcome measures

Cost Minimization Analysis (CMA) Economic evaluation where outcomes are assumed to be identical for the compared interventions

Cost Utility Analysis (CUA) Economic evaluation involving the use of an outcome measure combining mortality and morbidity, usually quality adjusted life years (QALYs) or disability adjusted life years (DALYs)

(continued)

Panel 24.5 (continued)

Cost effectiveness acceptability curve (CEAC) Output from PSA, indicating the probability that an intervention is cost-effective relative to its comparator

Deterministic cost effectiveness analysis A CEA that use point estimates for parameter values, while uncertainty can be explored using sensitivity analyses (one-way, two-way or multi-way)

Explicit budget A situation where the available amount of funds for a priority decision is known

Extended dominance A situation where an intervention has a higher ICER than the next more effective alternative, implying that the intervention is strongly dominated by a combination of two alternatives

Implicit budget A situation where the exact amount of funds available for a priority decision is undefined, in which case priority setting can be based on willingness to pay for health assessment

Incremental costs The cost difference between two mutually exclusive intervention alternatives

Incremental Cost Effectiveness Ratio (ICER) The ratio of incremental costs and incremental effectiveness between two intervention alternatives that are mutually exclusive

Incremental effectiveness The difference in effectiveness between two mutually exclusive intervention alternatives

Monte Carlo simulation A process where a model is evaluated by making a large number of random draws from a set of distributions, and where expected values are calculated for each simulation

Mutual exclusiveness Situation where costs and effectiveness of an intervention is influenced by the other intervention alternatives being compared. An implication is that only one of the interventions should be given to the patient at the same time (i.e. one treatment regime against a disease)

Mutual independence Situation where the costs and effectiveness of an intervention are independent of the other intervention alternatives being compared. Several interventions can be given at the same time without influencing each other

Probabilistic sensitivity analysis (PSA) An analytical approach to consider potential impact of parameter uncertainty, involving defining distributions for uncertain parameters, combining them in a model using Monte Carlo simulation, and presenting the results using e.g. CEACs

Strong dominance A situation where an intervention is more costly while at the same time being less effective than its comparator

Willingness to pay (WTP) for health The maximum amount of money decision makers are willing to pay for an additional unit of health

While the identification and estimation of various types of costs are simsilar across economic evaluations, the nature and measurement of health outcomes may differ considerably (Drummond et al. 2005). It is the choice of outcome measure that classifies studies into different types of economic evaluations.

The simplest economic evaluation technique is *cost minimization analysis* (CMA). A prerequisite for cost minimization is that the health outcomes of the alternative programs are identical. If the health outcomes, or health effects, are identical for all the alternatives, the cheapest alternative should naturally be preferred from an economic standpoint. In *cost effectiveness analysis* (CEA) both costs and outcomes may differ between intervention alternatives. Typically, a currently used intervention is compared with an alternative with better health outcomes, but which is also more costly. But comparison with less costly and/or less effective alternatives can also be done. The outcome measure in CEA is a single measure, such as cases averted, life years saved, or deaths averted. CEA is foremost useful to compare interventions targeting the same condition.

A simple measure, like deaths averted, is insufficient as the outcome measure when other factors matter, such as patients' health-related quality of life. *Cost utility analyses* (CUA) utilize measures of health that combine mortality with morbidity. The quality adjusted life year (QALY) and disability adjusted life year (DALY) methods combines ill health with mortality into a single numerical expression. This makes it possible to compare interventions targeting different types of health conditions.

Although CUA has a wider applicability range than CEA, the former method can still only be used to compare projects with health-related outcomes. This is unsatisfactory for projects with outcomes across different sectors. For example, water and sanitation projects typically improve population health through improved water quality and a more hygienic disposal of waste. At the same time, such projects typically make living easier for people and save a lot of time that can be used for income-generating activities. Thus, sometimes employing only health measures underestimates the true benefits of the project. It is therefore sometimes convenient to compare all program benefits in monetary terms, the approach that is taken in *cost benefit analysis* (CBA).

Despite the fact that CBAs are easy to interpret, and despite their usefulness in cross-sector comparisons, the technique is not common in economic evaluation of health programs mainly because the valuation of human life and disability in monetary terms is quite challenging (many also find it unethical to place a value on human life). CMA on the other hand, is rarely applicable due to the requirement of identical outcomes. CEA and CUA have therefore become the most influential techniques for economic evaluation in health care.

We have discussed different effect measures in economic evaluations, and observed that the choice actually determines whether we are dealing with CMA, CEA/CUA, or CBA.

24.6.2 Decision Analytical Modeling

Economic evaluation can, by principle, be designed around clinical trials. Information such as efficacy, treatment compliance, long term treatment benefits, health state preferences, and various costs may be collected simultaneously for both the study intervention and its comparators. With this technique, cost-effectiveness estimates can be calculated directly using standard mathematical and statistical approaches. However, coherent and comprehensive information on all of these factors are rarely available; therefore, economic evaluations usually combine different types of evidence from several sources to develop decision analytical models.

Modeling provides great flexibility regarding the availability of evidence, adaptability to different settings and situations, and opportunities to explicitly model and express uncertainties in input variables and in the overall findings. In so-called deterministic evaluations, the parameter values are modeled as point estimates, and in *ex post* calculations they are allowed to vary in *sensitivity analyses*. In probabilistic analyses, on the other hand, the parameter values are treated as distributions, meaning that inferences from trials and other data sources are more directly incorporated into the models. These two approaches will be briefly introduced below.

24.6.3 Deterministic Cost-Effectiveness Analysis

Average cost-effectiveness analysis (ACEA) is a type of deterministic cost-effectiveness analysis based on a very simple mathematical formula, namely the cost-effectiveness ratio (CER).

$$\mathrm{CER} = \frac{\mathrm{Costs(C)}}{\mathrm{Effects(E)}} \tag{24.6}$$

The CER estimates how much it costs to obtain a unit of health outcome using some specified intervention (e.g., the amount of dollars it costs to avert a malaria case by providing bed-nets to the population). This measure, also called the *average CER*, can be used to choose the optimal mix of interventions in a particular optimization problem (Weinstein 1995). This optimization problem contains a set of rather restrictive assumptions, including that the interventions (i = 1, 2, …., N) are not repeatable, that their costs and benefits reflect full implementation, and that costs and effectiveness of any program are independent of which other programs are adopted (‘mutual independence’). The optimal resource allocation, or the allocation that gives the highest attainable aggregate health effect within the budget, is obtained by rank ordering the programs according to increasing CER (C_i/E_i) and adding interventions from the top of the list until the budget is exhausted.

The above decision rule is simple to work out and to communicate to decision makers and to medical professionals. However, one of the assumptions of the basic model is violated in the most typical application of CEA, namely the assumption of mutual independence. CEA is typically used to compare competing alternatives to target the same condition, e.g., alternative drugs to treat malaria. In these cases the alternative interventions are not mutually independent, since the use of one malaria drug will affect the effectiveness of other malaria drugs. In these cases, average CERs cannot be used to maximize the health for a given amount of resources (Karlsson and Johannesson 1996). Mutual exclusiveness requires modifications of the decision rules of the basic CEA model, and can be dealt with by calculating *incremental cost-effectiveness ratios* (ICERs).

24.6.3.1 Incremental Cost-Effectiveness Analysis

When we are dealing with a menu of interventions that are mutually exclusive, the decision of which alternative should be given priority depends on the outcome of two processes. The first of these processes is to rule out interventions that are dominated, either by *strong dominance* or by *extended dominance*. The second process is to consider which of the remaining programs maximizes our health objective within the limits of the budget. Before we return to these processes, it is useful to make some modifications to the basic cost-effectiveness model.

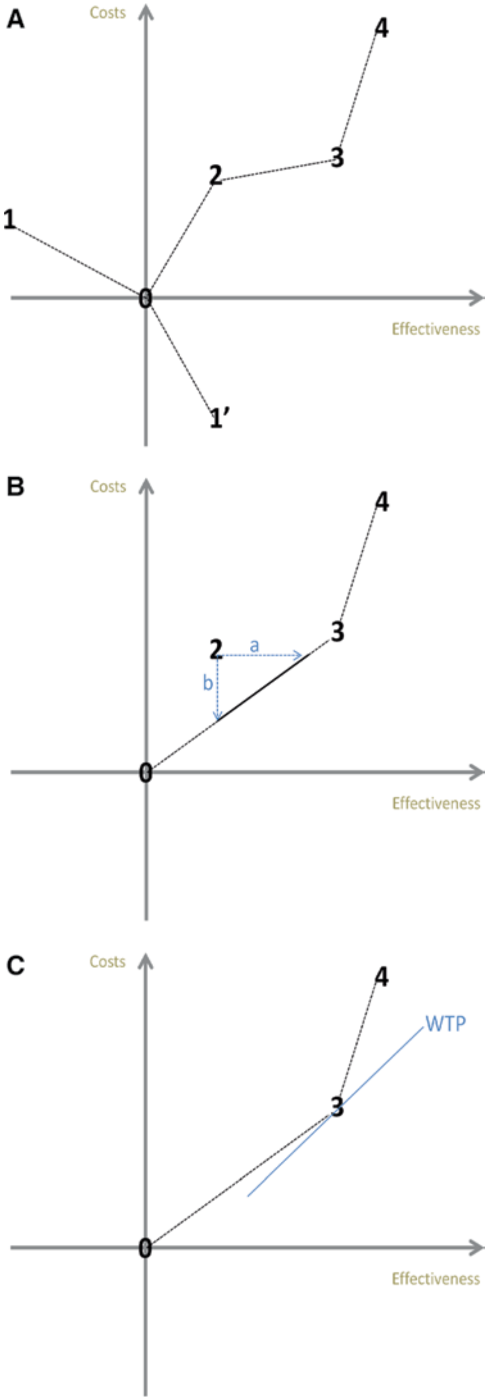
The question typically being investigated in incremental CEA (ICEA) is whether or not it is worthwhile to switch from some current standard treatment to an alternative treatment that is more effective but typically also more costly. In other words, we are interested in comparing the additional, or *incremental*, costs and health effects by switching from the old regimen to the new one. The ICER of moving from program 1 to program 2 can be expressed as:

$$\mathrm{ICER}_{1-2} = \frac{C_2 - C_1}{E_2 - E_1} = \frac{\Delta C}{\Delta E} \qquad (24.7)$$

We will use a *cost-effectiveness plane* (Fig. 24.6A–C) to illustrate the steps of ICEA, with intervention effectiveness on the x-axes, and costs on the y-axes. Let's assume a situation in which we are considering to replace current standard of practice (intervention 0) with one out of four treatment alternatives (interventions 1–4). The alternatives are mutually exclusive.

First, consider alternative 1 compared to 0 (current practice). Figure 24.6A illustrates that 1 is less effective than 0 and more costly. This can be denoted as $C_1 - C_0 > 0$; $E_1 - E_0 < 0$. Alternative 1 is therefore undesirable as replacement for 0, and the correct terminology is that 1 is dominated by 0 by strong dominance. In other words,

Fig. 24.6 *Graph A* illustrates how intervention 1 is strongly dominated by 0, while 1′ strongly dominates 0. *Graph B* illustrates the concept of extended dominance, while *graph C* illustrates how willingness to pay for health affects the choice between two mutually exclusive alternatives 3 and 4



1 should be rejected. Alternatively, let's imagine that alternative 1 is cheaper and more effective (denoted 1′) compared to current practice; or $C_1 - C_0 < 0$; $E_1 - E_0 > 0$. In this case the new alternative is preferable, and we say that 1′ dominates 0 by strong dominance.

Extended dominance exists when an option has a higher ICER than that of the next more effective alternative (Karlsson and Johannesson 1996). In this situation, the option is less effective and more costly than a linear combination of two other strategies (McGuire 2001). This is the case for intervention 2 in Fig. 24.6A, where the ICER for 0–2 is higher (steeper), than the ICER for 2–3.

Rather than giving the same intervention to all patients, it is possible to give intervention 3 to some patients, and to continue with current practice to the rest (alt 0). By considering Fig. 24.6B, it becomes clear that a combination of interventions 0 and 3 (0–3) can produce better outcomes at identical cost to intervention 2 (arrow "a"), or the same effectiveness at lower cost (arrow "b"). Intervention 2 is therefore *extendedly dominated* by 0–3.

After having ruled out interventions that are either strongly or extendedly dominated, we are left with two alternatives to current practice (3 and 4) that have increasing ICERs when they are sorted according to effectiveness. When all dominated alternatives have been ruled out, we are ready to select which of the remaining interventions should be funded given the available budget. If we have an *explicit budget*, where the amount of available funds is defined, it is common to list the remaining interventions in a *league table* sorted according to increasing ICERs. Since the interventions are mutually exclusive, we start on the top of the list and replace interventions till the resources are exhausted (as opposed to ACEA explained above, where mutually independent interventions were added).

Often the exact amount of money available for a new intervention is undefined, in which case assessment must be made based on assessment of an *implicit budget*. Examples are government agencies like the National Institute of Clinical Excellence (UK), which consider economic evidence before (dis)approving funding for new drugs.

Implicit budgets involve defining a maximum acceptable value for ICERs, sometimes called the *willingness to pay* (WTP), and sometimes *value of ceiling ratio* or *cost-effectiveness threshold*. WTP can be illustrated as straight lines in the CE-plane (Fig. 24.6C). The decision rule is to fund the best possible intervention with an ICER lower than the slope of the WTP curve. In our example this is intervention 3, since the ICER of moving from 3 to 4 is steeper than the WTP curve. One could also say that intervention 3 is cost-effective compared to the WTP threshold, whereas intervention 4 is not.

24.6.3.2 Sensitivity Analyses

Above we have assumed that costs as well as health effects are certain and can be described by point estimates. This is hardly ever the case. In fact, the presence of uncertainty is sometimes mentioned as a reason for why economic evaluations are important in the first place (Drummond et al. 2005). CEA can be used to quantify different sources of uncertainty and analyze whether or not they are likely to influence decisions to implement health interventions. If the findings and conclusions are sensitive to changes in parameter values of key variables, then the analysis is not robust, and the firmness of the conclusions must be modified accordingly.

The simplest and most commonly performed sensitivity analyses are so-called *one-way analyses*, in which one parameter value is varied at a time, while all the others are being kept constant. For example, in an economic evaluation of prevention of mother to child transmission in Tanzania, one-way analysis demonstrated that the interventions were more cost-effective in areas with high maternal HIV prevalence than in low prevalence areas (Robberstad and Evjen-Olsen 2010).

One-way analyses are simple to perform and provide output that is easy to interpret, but the information is insufficient because usually there is uncertainty in many variables. It is therefore more realistic to undertake *multi-way analyses*, where two or more variables are varied at a time (Drummond et al. 2005). In *two-way analyses* the output can be illustrated and interpreted in figures or tables, but as more parameters are included multi-way analyses soon become difficult to illustrate and even more difficult to interpret. *Scenario analyses*, like best-case, base-case, and worst-case analyses, illustrate potential consequences of combining the most optimistic, realistic, and pessimistic estimates, respectively.

24.6.4 Probabilistic Sensitivity Analysis

The classical sensitivity analyses described above are important to explore parameter uncertainties, but they are not well suited to model interactions between sources of uncertainty and to illustrate overall model uncertainty. In *probabilistic sensitivity analysis* (PSA) probability distributions replace point estimates as input into the decision analytical model (Briggs 2001). A range of alternative distributions can be used, and the choice should reflect the nature of the data. Gamma distributions are, for example, common choices for cost parameters, because they, like costs, are constrained on the interval 0 to positive infinity. Beta distributions, on the other hand, better reflect the binomial nature of probabilities (Briggs et al. 2006).

The decision analytical model combines values and calculate outcomes based on random draws from the distributions in a process called *Monte Carlo simulation*. The simulation generates a large number of cost/effect pairs that can be plotted in a cost-effectiveness plane (Fig. 24.7A). Mean ICERs, standard deviations and confidence intervals can be calculated using standard statistical methods. Such “clouds” of cost-effectiveness observations are well suited to illustrated overall model uncertainty, but interpretation and selection of interventions becomes challenging because the clouds often overlap both with each other and across the axes in the diagram.

Cost effectiveness acceptability curves (CEACs) provide a solution to the challenges of interpreting scatterplots from Monte Carlo simulation, and have become a standard method to report results in cost effectiveness and cost utility analyses (Fig. 24.7B). In Fig. 24.6C above intervention 3 was cost-effective compared to the WTP, while intervention 4 was not. The CEAC simply illustrates how many of the ICER pairs from a Monte Carlo simulation that falls below the WTP threshold, or in other words, the probability that an intervention is cost-effective (Briggs 2001). Figures 24.7A, B illustrate the relationship between the cost-effectiveness plane and

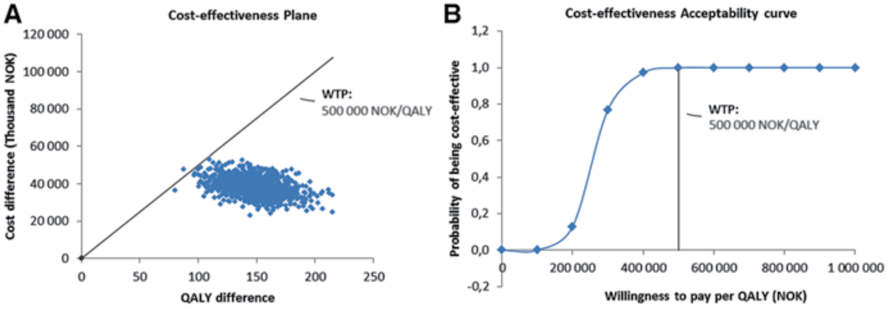


Fig. 24.7 *Graph A* provides the output from a Monte Carlo simulation in which a pneumococcal vaccine is compared to no vaccination in Norway (Robberstad et al. 2011). The 1,000 cost-effectiveness pairs from the simulation appear as a “cloud”. *Graph B* shows how the probability of the vaccine being cost-effective increases with increasing willingness to pay

the CEAC, and how the probability of an intervention may change with different levels of WTP. Probabilistic sensitivity analyses with CEACs produce output that are directly applicable to decision makers, while at the same time being explicit about the level of uncertainty surrounding the advice. The method provides decision makers with direct advice about which interventions to finance, based on the decision makers’ own perception of the societal willingness to pay for health. Since the readers actively have to assess the willingness to pay in order to interpret the findings, the danger of the analyst imposing personal values on the decision maker is smaller. This makes acceptability curves less value-laden compared to classical incremental CEA, where the definitions of cost-effectiveness are more implicit.

In this chapter we discussed a selection of frequently employed methods of statistical modeling. A special topic in modeling is the meta-analysis, in which data from multiple studies (i.e., meta-data) are analyzed. A meta-analysis is almost always preceded by a systematic literature review to identify all relevant studies containing meta-data. We discuss both systematic reviews and meta-analyses in the next chapter.

References

Borghi E et al (2006) Construction of the World Health Organization child growth standards: selection methods for attained growth curves. Stat Med 25:247–265

Breslow NE (1975) Analysis of survival data under the proportional hazards model. Int Stat Rev 43:45–57

Briggs A (2001) Handling uncertainty in economic evaluation and presenting the results. In: Drummond M, McGuire A (eds) Economic evaluation in health care. Merging theory with practice. Oxford University Press, Oxford, pp 172–214. ISBN 0192631764

Briggs AM, Sculpher M, Claxton K (2006) Decision modeling for health economic evaluation. Oxford University Press, Oxford, pp 1–237. ISBN 9780198526629

Cole TJ, Green PJ (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. Stat Med 11:1305–1319

Drummond MF et al (2005) Methods for the economic evaluation of health care programmes. Oxford University Press, Oxford, pp 1–396. ISBN 0198529457

Fadnes LT et al (2011) Vaccination coverage and timeliness in three South African areas: a prospective study. BMC Public Health 11:404

Francis D et al (2009) Fast-food and sweetened beverage consumption: association with overweight and high waist circumference in Jamaican adolescents. Public Health Nutr 12:1106–1114

Growth Analyser, version 3.0 (Application) (2009) Dutch Growth Foundation, Rotterdam. www.growthanalyser.org. Accessed Feb 2013

Hanley J, Miettinen OS (2009) Fitting smooth-in-time prognostic risk functions via logistic regression. Int J Biostat 5:1–23

Karlsson G, Johannesson M (1996) The decision rules of cost-effectiveness analysis. Pharmacoeconomics 9:113–120

Kleinbaum DG, Klein M (2002) Logistic regression: a self-learning text, 2nd edn. Springer, New York, pp 1–520. ISBN 0387953973

Lee ET, Go OT (1997) Survival analysis in public health research. Annu Rev Public Health 18:105–134

Leung KM, Elashoff RM, Afifi AA (1997) Censoring issues in survival analysis. Annu Rev Public Health 18:83–104

McGuire A (2001) Theoretical concepts in the economic evaluation of health care. In: Drummond M, McGuire A (eds) Economic evaluation in health care: merging theory with practice. Oxford University Press, Oxford, pp 1–21. ISBN 0192631764

Miettinen OS (1985) Theoretical epidemiology. Delmar, New York, pp 1–359. ISBN 0827343132

Miettinen OS (2011a) Epidemiological research: terms and concepts. Springer, Dordrecht, pp 1–175. ISBN 9789400711709

Miettinen OS (2011b) Up from clinical epidemiology & EBM. Springer, Dordrecht, pp 1–175. ISBN 9789048195008

Miettinen OS et al (2008) Clinical diagnosis of pneumonia, typical of experts. J Eval Clin Pract 14:343–350

Robberstad B, Evjen-Olsen B (2010) Preventing mother to child transmission of HIV with highly active antiretroviral treatment in Tanzania–a prospective cost-effectiveness study. JAIDS 55:397–403

Robberstad B et al (2011) Economic evaluation of second generation pneumococcal conjugate vaccines in Norway. Vaccine 29:8564–8574

Steyerberg E (2009) Clinical prediction models. A practical approach to development, validation, and updating. Springer, New York, pp 1–497. ISBN 9780387772431

Thinkhamrop B (2001) A handbook of categorical data analysis in health science research. Khon Kaen University Press, Khon Kaen, pp 1–274. ISBN 9746540947

Tugwell P, Knottnerus A, Idzerda L (2012) Is an odds ratio of 3 too high a threshold for true associations in clinical epidemiology? J Clin Epidemiol 65:465–466

Weinstein MC (1995) From cost-effectiveness ratios to resource allocation: where to draw the line? In: Sloan FA (ed) Valuing health care. Costs, benefits, and effectiveness of pharmaceuticals and other medical technologies. Cambridge University Press, Cambridge, pp 77–97. ISBN 0521576466

Zhang J, Yu KF (1998) What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA 280:1690–1691

Systematic Literature Review and Meta-analysis

25

Jonathan R. Brestoff and Jan Van den Broeck

> *The search for truth is in one way hard and in another way easy, for it is evident that no one can master it fully or miss it wholly. But each adds a little to our knowledge of nature, and from all the facts assembled there arises a certain grandeur.*
>
> Aristotle

Abstract

Many clinical decisions, risk assessments, and public health policy decisions depend on the results of epidemiologic studies. Very rarely, however, do all studies on a particular subject reach the same conclusion. Consequently, there is a need in epidemiology to synthesize differing study findings objectively. Two related approaches to synthesizing findings from multiple studies are the *systematic literature review* and the *meta-analysis*. The purposes of this chapter are to introduce each of these useful approaches, to highlight their assumptions and limitations, and to illuminate practical aspects of conducting your own systematic literature reviews and meta-analyses.

J.R. Brestoff, MPH (✉)
Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck, M.D., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_25,

25.1 Systematic Literature Reviews

Literature reviews are important elements of intellectual discourse in epidemiology. They are one of the most efficient ways to orient oneself to a new topic or emerging theme, and they serve as a platform on which knowledge is found or consolidated. Generally speaking, there are two types of review articles: narrative and systematic reviews. A narrative review often discusses the broader context of a phenomenon and includes selected research papers on a topic (their selection being somewhat subjective). Systematic reviews, on the other hand, tend to focus on more narrowly defined topics and employ a pre-specified (*a priori*), objective method to identify all research papers on a topic. For selected terms and concepts relevant to systematic literature reviews and meta-analyses, See Panel 25.1.

Panel 25.1 Selected Terms and Concepts Relevant to Systematic Literature Review and Meta-analysis

Fixed effect meta-analysis A meta-analysis in which the estimation of a summary estimate for the population parameter is based on a model that assumes that the effect measured in individual studies is fixed and thus that the variation in estimates among individual studies is due to sampling variation only

Forest plot Graph representing the point and interval estimates of the effect measures of studies included in a meta-analysis (sometimes including representation of the overall summary estimate or summary estimates for relevant subgroups of studies)

Merging Combining datasets, either on the basis of them having common variables ('appending records'), or, on the basis of them having common observation units ('match-merging')

Meta-analysis A quantitative approach to contrasting and combining the findings of previous studies on a research topic in order to arrive at summary conclusions about the body of research done on the topic

Meta-regression A regression of study findings on study characteristics in a meta-analysis done to explore heterogeneity and to estimate differences between findings in subgroups of studies

Pooled analysis Analysis of a dataset that combines data from several individual studies or several sources[§]

Publication bias Bias in the overall pool of evidence around a research question due to selective publication based on the magnitude or direction of the study findings

Random effects meta-analysis A meta-analysis in which the estimation of a summary estimate for the population parameter is based on a model that assumes that the effect measured in individual studies is not fixed but that there is an underlying variation in effects among individual studies that has a Gaussian distribution

[§] *Definition contributed by Dr. M. Chhagan*

Table 25.1 An ordered approach to performing systematic literature searches

Order	Task	Examples
1	Define a focused **question**	Based on longitudinal studies only, what is the risk of developing colon cancer in patients with a history of ulcerative colitis?
2	Identify core **search terms**	Colon cancer, ulcerative colitis, longitudinal study
3	Create a list of all **synonyms** of core search terms	Colon cancer, colorectal cancer, colon adenocarcinoma, …
4	Select **bibliographic databases** to be searched	MEDLINE, EMBASE, …
5	Create search **query text**	("colon cancer" OR synonyms) AND ("ulcerative colitis" OR synonyms) AND ("longitudinal study" OR synonyms)
6	**Document** the search and store results	Export and save a list of all search results from each database

Consequently, in a systematic review article, there is always a Methods section that describes:

1. The *primary search* used to identify all potentially relevant articles on the topic
2. The *preliminary inclusion screen* used to narrow down the often very long list of primary search results into a short list of a manageable number of research studies to consider more carefully for inclusion
3. The *a priori inclusion criteria* used to formally include studies, and
4. The way in which the quality of included studies will be rated, e.g., by a *grading scheme*
5. The way in which the evidence from included studies will be summarized and presented

This typical organization is used to structure this section of the chapter.

25.1.1 The Primary Search

The primary search is employed to identify all potentially relevant articles that meet pre-specified criteria. Therefore, *before beginning the primary search, one must first define a very focused question that the systematic literature review aims to address* (Table 25.1, Task 1). It is wise to begin with a very simple question and then to add qualifiers that limit the scope of that question. For example, initially one might ask "What is the risk of developing colon cancer in patients with a history of ulcerative colitis?" In Table 25.1, we have added one qualifier – specifically, to base the answer to our question on longitudinal studies only – in order to limit the number of relevant studies and make the systematic review more feasible. Limiting the systematic review to only one general study design has practical implications when a meta-analysis is planned as part of the systematic review, as it can be very difficult or impossible to derive comparable or appropriate statistical estimates from different study design types. Further specifications in the question help to narrow the domain of the review even further and may include factors such as age or biological sex.

After clearly defining and documenting the question to be addressed in the systematic review, one must complete at least four other tasks to prepare for the primary search (Table 25.1, Tasks 2–5). First, one must use the stated question to identify the core search terms (Task 2). In our example, there are three core search terms: colon cancer, ulcerative colitis, and longitudinal study. Each of these terms, however, has a number of synonyms that also need to be included in the search. The term colon cancer, for example, is also commonly referred to as colorectal cancer and colon adenocarcinoma. The next task is therefore to identify all synonyms of the core search terms (Task 3). As you are developing a list of synonyms, you may encounter bibliographic databases that contain information on original journal articles. Keeping notes on these databases may be useful for the next task, to identify the databases that you plan to search (Task 4). Popular databases are MEDLINE and EMBASE. Relevant specialty databases may also exist, such as PsychINFO for psychology-related topics. For searching databases, one may make use of 'search engines' or systems, such as PubMed and Scopus.

The information generated in Tasks 2 and 3 is used to generate *query text* (Task 5), or the text that you will enter in the database's search field. Search queries rely on Boolean logic, in which one uses a set of hierarchical true-false relationships to define a search. In Boolean logic-based searches, each word is a search term, although multiple words can be designated as a single term using quotation marks (e.g., colon cancer becomes "colon cancer"). If one is attempting to enter multiple synonyms as a single search term, they are joined using parentheses and the word *OR* [e.g., ("colon cancer" *OR* "colon adenocarcinoma" *OR* "colorectal cancer")]. Using the fully constructed *query text*, search the databases and store records of all search results (Task 6). Most databases have an advanced function to display and/or export search results with both the citations and their full abstracts. This is the most useful search result output, as the preliminary inclusion screen involves manually screening titles and abstracts for articles with the potential to meet the pre-determined inclusion criteria.

25.1.2 The Preliminary Inclusion Screen

The primary search often yields a very long list on the order of several thousands of articles. In order to narrow this list down to a manageable number of articles that warrant more careful consideration, one employs a *preliminary inclusion screen*. This process involves reviewing all of the titles and abstracts for articles that appear to meet or have even the slightest potential to meet a set of pre-determined inclusion criteria (to be applied later). Often, there are no formal preliminary inclusion criteria for the manual screen, as all studies that pass the screen will enter a careful review process with strict pre-determined inclusion criteria. However, there are often exclusion criteria at this stage to remove narrative review articles, commentaries, editorials, book reviews, etc.

While performing the preliminary inclusion screen, there is a danger of introducing subjectivity and bias, though two commonly employed practices are aimed at combating these issues. The first is to have a very low bar for passing the screen. Any primary search results without enough evidence in the title and abstract to clearly exclude the reference should be retained for careful review. The second is to have two or more persons manually screen the primary search results independently. The screen results from all individuals are merged and advanced to the next round of review.

Hint

During the preliminary inclusion screen of a systematic literature review, keep records about the numbers of studies excluded and included at each step. Standard practice is to report the reasons a study has been excluded, or at least to report the number of studies that were excluded for a given reason.

25.1.3 Applying *a Priori* Inclusion Criteria

The preliminary screen should reduce the number of reports by 70–90 %, though the exact number will depend on your particular study and search. The full-text versions of all the studies that pass the screen must be obtained (electronic or paper media are acceptable) and shared among at least two independent reviewers. Separately, each reviewer should read every paper that passes the screen and determine which meet the pre-determined inclusion criteria, which will be discussed briefly in the next paragraph. The reviewers compile their own lists of included studies, and any discrepancies among these lists are reviewed again to reach a consensus decision regarding whether the study successfully meets the pre-determined inclusion criteria.

The inclusion criteria are ideally established before the primary search has been conducted. At minimum, these criteria must describe studies that provide information on the parameters of interest as well as any statistical estimates and/or test results that may later be necessary for meta-analyses (if meta-analysis is planned for). If such information is not available, one may contact the authors to request the needed information before excluding the article. The inclusion criteria must identify studies that are helpful in addressing the question at issue. Based on the example in Table 25.1, a study on colon cancer risk in adult patients with a history of inflammatory bowel disease would be excluded from the analysis because there are two forms of inflammatory bowel disease: Crohn's disease and ulcerative colitis. Such a study would not allow us to infer an association between colon cancer and ulcerative colitis. If, however, the results section contains analyses that specifically evaluate the risk of colon cancer in adult patients with a history of ulcerative colitis *or* Crohn's disease (where both associations are reported separately), then the study would be included as long as it was longitudinal in design, as this type of study was specified in the question found in Table 25.1. For much more comprehensive advice on building inclusion criteria, please see *Cochrane Handbook for Systematic Reviews of Interventions* (2011).

Table 25.2 A representative table showing characteristics of studies in a meta-analysis

References	Country, study name	Participants	No. of case subjects	Years of follow-up	Assessment of anthropometric measures	Adjustments
Lee and Paffenbarger (1992)	USA, Harvard Alumni Health Study	17,595 men	290 (colon cancer)	1962–1988 (ave 5.6)	Self-reported	Age, family history of cancer, physical activity
Bostick et al. (1994)	USA, Iowa Women's Health Study	32,215 women aged 55–69 years	212 (colon cancer)	1986–1990 (ave 4.8)	Self-reported	Age, height, parity, vitamin A supplement use, intakes of energy and total vitamin E

Table contents extracted from Larsson and Wolk (2007) in accordance with *Am J Clin Nutr* use policies. Only the first two studies were extracted. *USA* United States of America, *No.* number, *Ave* average

25.1.4 Assessing the Quality of Included Studies

After selecting studies that meet pre-determined inclusion criteria, one is faced with a challenge: how does one compare the studies' strengths and weaknesses in a meaningful way? Lower quality studies are less informative than higher quality studies and the evidence they provide should not weigh as much in the overall synthesis of evidence on the topic. Out of the need to address this question, appraising the quality of research studies has become an important practice. Though various quality appraisal approaches have been employed, only the two most common will be discussed briefly here.

The first approach, and by far the most useful, is to construct a table that allows side-by-side comparison of specific study characteristics. These tables can be difficult to make because so many factors contribute to a study's overall quality. An example of a well-constructed summary table is shown in Table 25.2 (modified from Larsson and Wolk 2007). These tables typically include the relevant sample size, key study-base characteristics (age, sex, catchment area), and major strengths and limitations beyond those typical of the study design type. If multiple study design types are included in the review, then this feature should also be included in the summary table.

The first approach is often taken in conjunction with a second one, to employ a grading scheme in an attempt to assess the overall quality of included studies. Various grading tools have been proposed, such as GRADE (GRADE Working Group 2004). These schemes rely on select study characteristics to assign a letter or number score that is intended to reflect a study's overall quality. However, we caution against the use of grading tools because many factors beyond those included in the grading scheme can contribute to a study's overall quality. Moreover, there is a temptation to directly infer from a study's quality score its strength of evidence, an association that can be false.

25.1.5 Presenting and Synthesizing Study Findings

The next step is to present an overview of the findings of each included study. This can be done in a separate table. In this overview table of study findings, the main information consists of details about parameter estimates, standard errors, P-values, and (for etiognostic studies) confounders. In some cases, systematic literature reviews are performed without further statistical analyses of the findings from the included studies. Such reviews amount to an overview of the evidence on a topic accompanied with qualitative synthetic inferences that subjectively give more weight to evidence from large or good-quality studies. However, much more commonly, systematic reviews are accompanied by statistical meta-analyses, a term that literally means “analysis of analyses.”

25.2 Meta-analysis: Objectives and Limitations

25.2.1 Objectives of Meta-analysis

Meta-analyses involve combining the results of previous studies on a research topic using statistical methods. The two objectives of a meta-analysis are nearly always:

1. To explore heterogeneity and reasons for the differences between studies
2. To provide summary estimates of outcome parameters, taking study precision into account

Both of these purposes are quite nicely illustrated in the *forest plot* shown in Fig. 25.1. This forest plot shows results from a meta-analysis of the relative risk

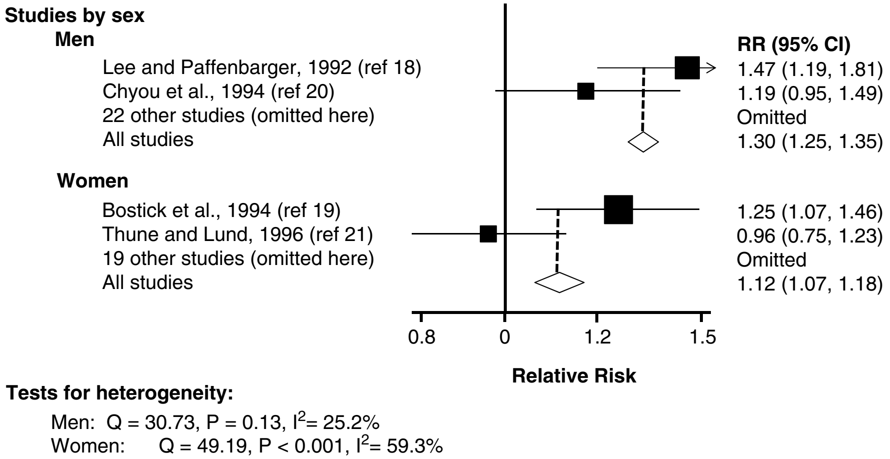


Fig. 25.1 Example of a meta-analysis forest plot. See the text body for descriptions of the elements shown in this figure. (Adapted from Larsson and Wolk (2007) in accordance with *Am J Clin Nutr* use policies.) This figure is an approximation only. Reference numbers in figure correspond to the reference list in Larsson and Wolk (2007). *Ref* reference, *RR* relative risk, *CI* confidence interval

(RR) of colon cancer for each 5 kg/m^2 increase in body mass index (BMI) relative to normal BMI. The square boxes indicate the RR estimate for each study listed on the left, and the box's sizes reflect the weight that a study has in the calculation of an overall RR estimate. The horizontal lines represent the 95 % confidence intervals (CI) of those RR estimates. On the right-hand side, the actual numerical values of the RR and 95 % CI are shown. The dashed vertical line marks the overall RR estimate for the studies being meta-analyzed and terminates on the center of an open diamond, the width of which represents the overall 95 % CI for the overall RR estimate. In this meta-analysis, there were significant increases in the overall RR for both men and women; however, the strength of the association was greater in men than women. Thus, this meta-analysis provides summary estimates of an effect measure *and* provides insight into the heterogeneity among studies.

25.2.2 Limitations of Meta-analyses

Although meta-analyses are potentially very useful and informative, there are some major limitations to this type of study. We focus here on the three most paramount limitations seen today.

First, the summary statistic can give the false impression of consistency among studies, a point that is well illustrated in the forest plot in Fig. 25.1. At face value, the overall RR and overall 95 % CI make it appear as though all of the included studies were similar to each other. However, not shown in this figure are the substantial methodological differences and differences in quality-aspects across the included studies. Three common methodological aspects that almost always vary from one study to another are the study base, the sampling procedure, and the methods by which study variables are measured. Consequently, it is critical to view forest plots in the context of summary tables that characterize the included studies.

Second, there may be *publication bias*, a phenomenon in which studies that show a significant association are more likely to be published than studies that show a non-statistically significant association. Publication bias arises in various ways, as will be discussed in-depth in Chap. 31. Briefly, investigators sometimes perceive that their work will be viewed as unimportant if a hypothesized association is not observed, leading to lower submission rates for studies with non-significant findings. In addition, some journals, editors, and peer-reviewers tend to be more likely to accept studies that show a significant association between two factors. These forces and others combine to produce a publication landscape that is artificially skewed in one direction or another. Unfortunately, although methods to detect publication bias exist (*See:* below), they tend to be useful only under conditions that are not commonly met.

Third, the statistical models that are commonly used to generate the summary effect measures make assumptions that are not always met or that cannot be tested. These models – the fixed effect and random effects models – are therefore considered to be dubious, and findings resulting from their use need to be interpreted carefully.

25.3 Steps of a Meta-analytical Project

Here, we discuss the approach to meta-analysis after identifying and assessing the quality of relevant studies, as was discussed in the previous sections on systematic literature reviews. A controversial point is whether one should exclude studies with poor quality. This needs to be decided on a case by case basis

A meta-analysis using pooled datasets (with a similar structure - see Chap. 12) proceeds as follows:

- Collect datasets from each study
- Add a study identification variable to each dataset
- Ensure common naming and coding of variables in each dataset
- Merge the datasets
- Perform pooled analysis, using the study identification variable as a potential effect-modifier

More commonly, however, a meta-analysis will not use pooled datasets but proceeds through the following steps, each of which is discussed further below:

- Extracting data from published articles for meta-analysis
- Optimizing comparability of extracted effect estimates
- Exploring heterogeneity among study findings
- Calculating summary estimates
- Assessing possible publication bias

25.3.1 Extracting Data for a Meta-analysis

Based on the research question developed in preparation for the systematic literature review, one should be able to foresee the types of findings that will be encountered in the included studies and that will need to be synthesized in the meta-analysis. Two critical pieces of information that always need to be obtained are (1) the point estimate of the outcome parameter and (2) the standard error (which represents the margin of uncertainty surrounding the point estimate). Both of these data can be provided in a number of formats. For instance, point estimates of effect measures are often reported as an odds ratio, relative risk, incidence rate ratio, of beta-coefficient. The standard errors of these estimates are sometimes directly reported but more commonly than not need to be derived from other information that is almost always present: from a confidence interval or a P-value (note: only P-values with two or more significant digits are useful for deriving standard errors). The extracted point estimate and its standard error are sufficient to perform a meta-analysis, but sometimes adjustments are necessary (next subsection) to enhance their comparability. For this purpose, additional data should be extracted from each article about which confounders were adjusted for and how exactly they were measured. Also, one should extract both crude and confounding-adjusted effect estimates. We recommend developing a data extraction form that can be used to record and store data that will be used for the meta-analysis later.

25.3.2 Optimizing Comparability of Extracted Effect Estimates

After extracting the effect estimates and their standard errors some thought needs to go to issues of comparability. A first problem can be that a particular study used a different definition for a confounding variable or that the study adjusted for a set of confounders that was slightly different from other studies. In that case, one should try to make an adjustment of the estimate produced by that study. External adjustment for no or incomplete accounting for confounding can be based on a correction factor estimated from an external study that gave crude as well as adjusted estimates. A correction factor can also be estimated from internal information on (1) the joint distribution of the confounder and exposure, and (2) external info on the confounder-outcome relation (Rothman and Greenland 1998).

A second problem of comparability can occur when a particular study is suspected or known to suffer from selection bias. For studies with (suspected) selection bias, one tries correcting the estimates using info on over- or under-sampling of determinant variable levels. If such information is not available, sensitivity analysis can be performed. Similarly, if there is suspected misclassification bias, one may attempt to correct the estimates using information on the over- or under-classification of variable levels, or one may consider performing sensitivity analyses of various potential misclassification scenarios (Rothman and Greenland 1998).

25.3.3 Exploring Heterogeneity Among Study Findings

When the estimates and standard errors of all studies have been obtained and optimized, one can proceed to addressing the two explicit objectives of meta-analysis, starting with the exploration of heterogeneity. In Fig. 25.1, a forest plot was shown, and in it effect estimates were grouped according to the suspected source of heterogeneity, biological sex. Inspection of this forest plot strongly suggested that estimates tended to differ according to sex, but the figure also mentions a *Chi-square test for heterogeneity* that supports the credibility of this hypothesis, given the data. Yet another way to explore heterogeneity is *meta-regression*, which is a regression of study findings (effect measures) on study characteristics. Both these methods are further discussed in Sect. 25.4.1. Note that in tests for heterogeneity and in meta-regression, it is possible to down-weight the contribution of studies with poorer quality.

25.3.4 Calculating Summary Estimates

Contrary to common perception, the calculation of summary estimates is not strictly necessary in a meta-analysis. Both heterogeneity exploration and estimation of an overall summary estimate are nearly always performed, but it is possible to limit the meta-analysis to the former only.

To make a summary estimate, two approaches are used frequently. *Fixed effect meta-analysis* assumes that the effects measured in individual studies are fixed, thus implying that any differences in estimated effects are due to sampling variation only.

Textbox 25.1 A Theoretical Argument in Favor of Random Effects Meta-analysis

When pondering whether fixed effect or random effects meta-analysis is the more justifiable approach, consider that the magnitude of measured effects often depends on underlying study population characteristics that can not always be measured and taken into account completely. For example, study populations can differ in the distribution of individual susceptibility factors in their members and in the distribution of environmental effect modifiers. Even if these 'background factors' or 'underlying risks' were perfectly balanced among comparison groups (as may be the case in some randomized trials) observed effects will be different among studies.

To illustrate this point with a purely hypothetical example, consider a drug x which only works well in blue-eyed persons and this individual susceptibility factor is not taken into account in the analysis of a number of trials in different populations in which the effect of x is studied. A trial in a population with a large majority of blue-eyed persons (randomized successfully among treatment arms) will estimate a large effect of x. In contrast, a trial done in a population with very few blue-eyed persons (even though this characteristic is well-randomized) will find only a small or no effect.

Measuring a true scientific effect would require taking all these types of effect-modifying background factors into account. This is generally unfeasible, which has to be taken as a given, and thus underlying variation in effects as assumed in random effects meta-analysis can also be taken as a given. Future meta-analysis may increasingly use meta-regressions on background factors which can possibly be assessed by additional data collection on ecological variables.

Random effects meta-analysis assumes that effects measured in individual studies are not fixed but have an underlying Gaussian distribution. As mentioned above, these assumptions cannot be tested empirically. Theoretically, however, we consider the random effects model to be a generally more credible assumption. Our main argument for this stance is described in Textbox 25.1.

In practice both fixed and random effects meta-analysis are often done in the same meta-analytical project. The basic calculations are mentioned in Sect. 25.4. Other statistical methods exist for the calculation of summary estimates. For example, Empirical-Bayesian methods have been used (Stijnen and van Houwelingen 1990), but they fall outside the scope of this text.

25.3.5 Assessing the Likelihood of Publication Bias

A funnel plot is a plot of a study precision parameter (usually 1/SE) against a parameter of effect size (usually an odds ratio, a relative risk, or an incidence rate ratio). When there is no publication bias, a funnel plot should show a symmetrical inverted funnel shape around a vertical line that indicates the overall effect estimate. Asymmetry around that line may indicate publication bias. A limitation is that many studies are needed to distinguish real from imagined patterns, so funnel plots are not always helpful. Statistical tests have been developed to test for funnel plot asymmetry, such as the rank correlation test of Begg and Mazumdar (1994). Possible causes of funnel plot asymmetry include:

- Publication bias
- The smaller studies may have used a prognostically different sub-population, such as high-risk patients

25.4 Statistical Aspects of Meta-analysis

25.4.1 Statistical Aspects of Exploring Heterogeneity

25.4.1.1 Meta-regression

Also known as 'effect size modeling,' meta-regression is used to study sources of heterogeneity in a meta-analysis. The estimates obtained in different studies form the dependent variable (e.g., odds ratios), and the study characteristics form independent variables (e.g., study size, study design type, study population characteristics, etc.). Because the number of studies in a meta-analysis is limited, the number of study characteristics that can be entered into a meta-regression is also limited.

In meta-regression the results of one study or a group of studies can be compared directly with the remainder of studies by adding an indicator variable that identifies study or group membership. The beta-coefficient of this indicator variable represents the difference between the effect measured by the study or group and the effect measured by the remainder of studies. Several such indicator variables can be added (Rothman and Greenland 1998).

25.4.1.2 Chi-Square Test for Heterogeneity

This type of test is suitable for exploring sources of differences between study findings in connection with fixed effect meta-analysis. The test involves the calculation of a Chi-square value (also called a Q-value). To the obtained Q value, one can attach a P-value considering that the degrees of freedom (DF) are equal to the number of studies minus 1. Q can be calculated using the following equation:

Calculation of the test statistic Q of a **Chi-square test for heterogeneity**

$$Q = \sum w_i \left[ln\left(RR_i\right) - \ln(RR_F) \right]^2 \tag{25.1}$$

Where:

RR_F = Fixed effect summary estimate (*See:* Sect. 25.4.2)

RRi = Effect estimates of the individual studies

ln = natural logarithm

w_i = weights, calculated for each study as in fixed effects meta-analysis (*See:* Sect. 25.4.2)

25.4.2 Calculating Fixed Effect Summary Estimates

In fixed effect meta-analysis, the assumption is that all studies estimate the same effect RR_F. Observed differences among *i* studies (among RR_i) are assumed to be due to sampling variation only. The summary estimate RR can be calculated as:

The **summary estimate** in **fixed effect meta-analysis**

Point estimate of fixed effect summary estimate RR_F:

Step-1:

$$\ln\left(RR_F\right) = \frac{\sum w_i * \ln(RR_i)}{\sum w_i} \tag{25.2}$$

Where:

$\ln(RR_i)$ = the natural logarithm of the effect estimates

w_i = weights, calculated for each study as:

$$1/\left(1/a_i + 1/b_i + 1/c_i + 1/d_i\right)$$

Where:

a_i = number exposed and with the outcome

b_i = number unexposed and with the outcome

c_i = number exposed and without the outcome

d_i = number unexposed and without the outcome

Step-2: Point estimate of $RR_F = e^{\ln(RRF)}$

Where:

e = the natural number *(~2.71)*

Confidence interval around RR_F:

Step-1: Calculate the SE of RR_F using the following formula:

$$\ln\left(RR_F\right) = \sqrt{\frac{1}{\sum w_i}}$$

Step-2: Calculate the upper and lower limits of the 95 % confidence interval around RR

$$\text{Lower limit} = e^{\ln\left(RR_F\right) - 1.96 * SE}$$

$$\text{Upper limit} = e^{\ln\left(RR_F\right) + 1.96 * SE}$$

25.4.3 Calculating Random Effects Summary Estimates

In contrast to fixed-effect model, the random effects model does *not* assume that studies measure the same effect. The assumption is that effects measured by studies arise from an underlying Normal distribution with variance S_b. This S_b is the between-study variance, which can be estimated from the data and is then added to each within-study variance S_i to calculate adjusted weights w_i. *The rest of the calculation of the summary estimate is the same as for fixed effect meta-analysis*; therefore, below we only show how to calculate the adjusted weights for use in a random effects meta-analysis:

Adjusted weights w$_i$ to be used in **random effects meta-analysis**

$$w_i = \frac{1}{S_i + S_b} \tag{25.3}$$

Where:
S_i = within-study variance, calculated as:

$$S_i = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

Where:
a_i = number exposed and with the outcome
b_i = number unexposed and with the outcome
c_i = number exposed and without the outcome
d_i = number unexposed and without the outcome
S_b = between-study variance, calculated as:

$$S_b = \max\left[0, \left(\frac{Q-(K-1)}{W}\right)\right]$$

Where:
Q = Chi-square for heterogeneity (*See:* Sect. 25.4.1)
K = number of studies
W = $\sum w_i - \sum w^2_i)/\sum w_i$ where w_i are the unadjusted weights, as calculated for fixed effect meta-analysis (Sect. 25.4.2)

Notable features of random effects meta-analysis, in comparison with fixed effects meta-analysis, are that:

- It has a wider confidence interval than the fixed-effect summary estimate
- It gives relatively greater weight to smaller studies and is therefore more prone to publication bias

Values obtained by the two approaches tend to be quite similar.

In random effects meta-analysis, the assumption that the true underlying effects are heterogeneous implies that one should be particularly careful with interpretation of the summary estimate. Specifically, the *interpretation should maintain the assumption of underlying heterogeneity,* as follows: The larger the summary effect and the farther its confidence interval is located away from the null effect, the more justification one has to interpret this as evidence for the existence of an effect 'in most circumstances/populations'. On the other hand, when the confidence interval estimated with random effects meta-analysis overlaps with the null value, this does *not* imply that there are no circumstances or populations where the determinant or treatment has an effect. Finally, the finding of a wide confidence interval in a random effects meta-analysis can generally be interpreted as indicating the need for more research into effect modification.

In this penultimate chapter of Part IV: Study Analysis, we described methods of synthesizing evidence from different sources in a meta-analysis. In fact, every epidemiological study will eventually arrive at a stage where the quantitative evidence produced needs to be compared with other existing evidence on the same topic. In these comparisons the final quality and credibility of one's own studies and those of others are important. In the next chapter (Chap. 26: The Ethics of Study Analysis) we argue that the ethical conduct and quality of statistical analyses is one of the important pre-conditions for achieving appropriate final quality and credibility.

References

Begg CB, Mazumdar M (1994) Operating characteristics of a rank correlation test for publication bias. Biometrics 50:1088–1101

Bostick RM, Potter JD, Kushi LH et al (1994) Sugar, meat, and fat intake, and non-dietary risk factors for colon cancer incidence in Iowa women (United States). Cancer Causes Control 5:38–52

GRADE Working Group (2004) Grading quality of evidence and strength of recommendations. BMJ 328:1490

Larsson SC, Wolk A (2007) Obesity and colon and rectal cancer risk: a meta-analysis of prospective studies. Am J Clin Nutr 86:556–565

Lee IM, Paffenbarger RS Jr (1992) Quetelet's index and risk of colon cancer in college alumni. J Natl Cancer Inst 84:1326–1331

Rothman KJ, Greenland S (1998) Modern epidemiology, 2nd edn. Lippincott Williams & Wilkins, Philadelphia, pp 1–738. ISBN 0316757802

Stijnen T, van Houwelingen JC (1990) Empirical Bayes methods in clinical trials meta-analysis. Biom J 32:335–346

The Cochrane Collaboration (2011) Cochrane handbook for systematic reviews of interventions. http://www.cochrane.org. Accessed Sept 2012

The Ethics of Study Analysis

26

Jan Van den Broeck and Jonathan R. Brestoff

If you think it's expensive to hire a professional to do the job, wait until you hire an amateur.

R. Adair

Abstract

From the medical literature one might gain the impression that data analysis is not much more than a matter of choosing the right estimators and making the appropriate adjustments for confounders. Yet the process of data analysis has important ethical and practical dimensions that are less apparent from reading research papers. In analyzing data, there are interactions among multiple persons, and critical decisions must be made for the calculation and selection of outputs for reporting. Even if the analysis plan is very well designed, each step in the analysis process is liable to becoming a source of bias, irreproducibility, inefficiency, poor documentation, disrespect for confidentiality, and even fraud. This chapter explores the causes of analytical deviances and strategies to prevent them. These topics are contextualized with discussions of the importance of ethical data analysis and the responsibilities of analysts involved in a research study.

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_26,
© Springer Science+Business Media Dordrecht 2013

26.1 The Importance of the Ethics of Study Analysis

Most critical to epidemiology is the validity of the findings it produces. Without appropriate data analysis conforming to ethical principles, the validity of epidemiological studies stands on shaky ground. Poor data analysis practices undermine the discipline of epidemiology and may lead to decisions, based on biased or erroneous findings, which harm patients or the public. Unfortunately, poor analysis practices are more commonplace in epidemiology than one might surmise (Altman 1980; Swazey et al. 1993; García-Berthou and Alcaraz 2004; Jeng 2006; Martinson et al. 2010; Horner and Minifie 2010).

The medical research community still needs to come to grips with a serious problem of dubious data integrity, a commonly denied problem that has been lingering for decades. Many players are involved in maintaining the integrity of data analyses, not just the investigative team. Other players include educators, journal editors, research institutions, mentors, peer reviewers, ethics committees, DSMB's, sponsors, regulatory bodies. The major players and their typical relationships in terms of maintaining ethical data analyses are depicted in Fig. 26.1, and terms and concepts relevant to the ethics of study analysis are listed in Panel 26.1.

Panel 26.1 Selected Terms and Concepts Relating to the Ethics of Study Analysis

Data analyst Person performing data transformations and statistical analyses

DSMB Data and Safety Monitoring Board. Independent oversight committee installed by a research sponsor in support of a particular ongoing study, charged with the regular review of data quality and participant safety, and advising investigators and sponsor on these

Fraud (– in analysis) `Fabrication of analysis findings

Misbehavior (– in analysis) Neglect, carelessness, or subtle choices to tweak procedures slightly in the hope of obtaining a more favorable result

Misconduct (– in research) Data fabrication, data falsification, plagiarism or other activities that seriously deviate from accepted practice by the scientific community for conducting and reporting research

Parsimony Avoidance of using unnecessary elaborate concepts (hypotheses, models)

Statistician Professional skilled in sampling from sampling frames, quantitatively summarizing and presenting data, estimation of population parameters and hypothesis testing

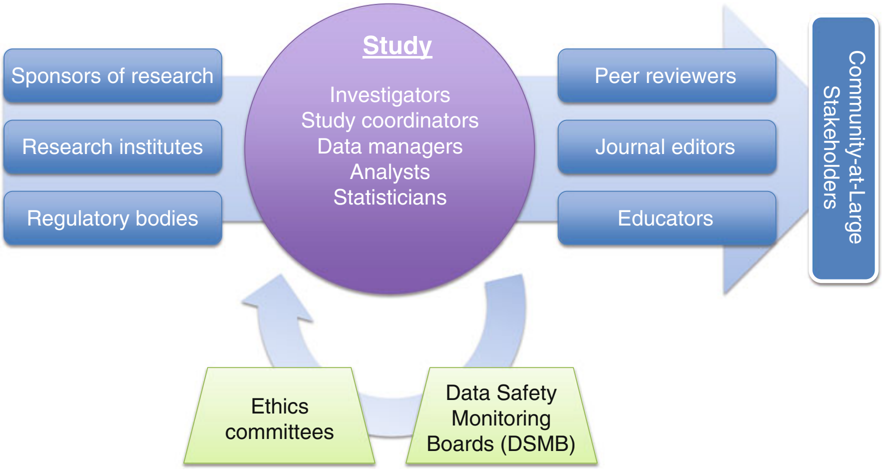


Fig. 26.1 Schematic representation of the major participants involved in ensuring ethical study analysis. Principle investigators approach sponsors, research institutes, regulatory bodies, and perhaps other stakeholders for funding and support. These entities review proposals and their associated analysis plans. The ability to assess the integrity of these plans depends on the level of thought and detail provided by the principle investigators. The study (circle) is administered by investigators, study coordinators, and technicians/support staff, and all of these persons work with data managers, analysts, and statisticians to analyze recorded data. The DSMB and ethics committees are critical elements of the study's approval and review, and members of the study team must interact with the DSMB and ethics committee to support ethical data analysis. The integrity of data analyses should be checked by peer reviewers and journal editors before dissemination of study findings to the community-at-large and other stakeholders, and educators should champion examples of successful data analysis so that the future epidemiologists are well prepared to avoid analysis errors and to perform high quality analyses

26.2 Errors and Unethical Behavior During Data Analysis

Mistakes, misbehavior, and misconduct can occur at any stage of research and tend to result in erroneous data and biased statistical evidence. Studies have been done on the frequency of occurrence and the determinants of mistakes, misbehavior, and misconduct in research, but the focus of those studies has rarely been on deviances occurring specifically during statistical analysis. An exception is the often-discussed topic of misguided choice of statistical procedures (Altman 1980), a decision that can occur when designing an analysis plan or during actual study analysis. However, many other types of bad choices can be made when preparing data for analysis or during the analysis itself. Table 26.1 contains some procedural aspects of data analysis amenable to intentional or unintentional error.

Tiered distinctions can be made among misconduct, misbehavior, and mistakes (Martinson et al. 2010). The most serious tier is misconduct. Examples include data fabrication, data falsification, and plagiarism. Misbehavior (the next most serious tier)

Table 26.1 Sources of intentional and unintentional error during data analysis

Aspect of analysis	Potential sources of error
Preparation of data for analysis	Inexact definitions of analysis variables
	Choices of (adjustment) variables not described or incompletely described in the analysis plan
	Extraction of analysis datasets
	Exclusions from analysis
	Handling of suspect data values discovered in analysis datasets
	Calculation of derived variables for analysis
	Data transformations to meet statistical assumptions
Statistical operations	Choice of statistical package or method of calculation
	Syntax writing
	Options for scoring, rounding, interval estimation
	Options for iterations, model inclusion criteria, imputations, output content and format
	Selection of outputs to report on

involves neglect, carelessness, or subtle choices to tweak procedures slightly in the hope of obtaining a more favorable result. These misbehaviors can also be viewed as mild forms of misconduct. Less serious than misconduct and misbehavior, mistakes are unintentional mishaps but are expected to be made more frequently by those who are careless, unskilled, inexperienced, acutely ill, tired, distracted, or under time pressure.

26.2.1 Procedural Errors During Study Analysis

That procedural errors are commonplace in statistical practice was suggested by García-Berthou and Alcaraz (2004), who showed that a considerable proportion of articles in two major medical journals contained P-values that were inconsistent with the corresponding test statistics. It has also been shown that assumptions needed to use particular tests are commonly not met (Altman 1980; Jeng 2006; Horner and Minifie 2010). Both of these potential problems can be mitigated by making concerted efforts to know in detail the study protocol before carrying out an analysis, to take enough time to execute the analysis, to double-check results, and to discuss provisional results with collaborators. These efforts of the investigative team must be supported by institutions that foster a culture of intense collaboration, mutual support, and quality in research.

26.2.1.1 Small Studies and Student Research Projects

Data analysis is a learned skill that increases with experience. Universities and Departments differ substantially in the quality of education in data handling and statistical analysis, and not all students are well versed in these skills when they get involved in research projects. In addition, students may not have clear insight into the limits of their own analysis skills and therefore may not always ask for statistical support or supervision when doing so would be advisable. The ready availability of analysis support is important and should be a main organizational concern for educational and research programs involving students. If support is lacking or delayed, the result can be low quality student research projects with especially poor statistical analyses. This, in turn, can contribute to a failure of research education and to the perpetuation of a culture of poor statistical knowledge and practice among investigators in health research.

26.2.2 Misbehavior and Misconduct During Data Analysis

Although there are no data available on the frequencies of most forms of misbehavior and misconduct during data analyses, mild-to-moderate forms of both are probably not uncommon. Examples include rounding P-value, replacing one case definition with a very similar one, and changing covariates in a model to achieve a desired result. Martinson et al. (2010) surveyed faculty of 50 major universities in the United States and found that more than a quarter reported not always taking proper care of data integrity and confidentiality. About 60 % admitted to forms of neglect and carelessness during research conduct, and 8 % reported their own plain misconduct (fabrication, falsification, or plagiarism). It was not clear how much of this happened specifically during data analysis, but it is reasonable to assume that if one is neglectful or carelessness at one stage of the research process that they are more likely to exhibit those same behaviors during analysis procedures.

Martinson et al. (2010) also found that misbehavior and misconduct were more frequently reported by those who felt they were treated unfairly in their immediate work environment and by those who felt they were 'over-committed' to their work. Davis et al. (2007) reviewed 92 cases of serious misconduct in the United States using qualitative research methods and identified several factors contributing to misconduct, including:

- Pressure from the publish-or-perish culture in academia
- Personal stressful situations
- Absence of a work climate fostering research integrity
- Job insecurity
- Character weaknesses

The available evidence suggests that misbehavior and misconduct occur not only because of personal factors or fallout from competition in academia but also from systems-level factors. Systematic factors are more amenable to intervention and modification. In fact, systems-level sources of error have been treated with quality assurance checklists in other operational situations, such as surgical procedures. Many hospitals now employ checklists before and after surgical procedures to reduce the frequency of surgical objects (especially sterile sponges) left inside patients. Similarly, quality assurance checklists in data analysis may prove to reduce the frequency of error in the data analysis process.

Hint

We do not wish to suggest that initial analysis plans should never be changed. In earlier chapters we mentioned that the assumptions underlying chosen statistical parameters may prove invalid after data exploration and that this may lead to necessary and legitimate revisions to the statistical analysis plan. At stake here is whether there are any ethical concerns regarding any changes made to the initial analysis plan. Using new tests, exclusions, adjustments, or imputations should never be done if they are inspired by a desire to achieve statistical significance or a certain result.

26.3 Quality Assurance in Data Analysis

Baerlocher et al. (2010) surveyed data integrity safeguards used by investigators who published their research findings in four major medical journals. More than 10 % admitted to never applying any measures to safeguard data integrity in any study phase. About 35 % used independent verification of the analysis for their last published paper, but only about 7 % opined that this practice should become a general requirement. The most common suggestion was that a simple declaration from the main author that ‘data integrity had been ensured’ should be the only requirement; a practice that we believe is an insufficient safeguard.

Investigators can avoid most analytical errors and fraud at the analysis stage by employing some simple quality assurance measures. Panel 26.2 provides a list of quality assurance recommendations, all of which are based on the expectations that errors and fraud occur less when the analysis involves interactions among multiple persons, good planning, proper support, and proper documentation of procedures. In clinical trials the Data Safety Monitoring Board (DSMB) also plays a role in quality assurance of data analysis by carefully checking statistical results in interim reports.

Seeking competent statistical support is important in all studies, not just in clinical trials; however, professional support is not a guarantee against poor analyses. For example, DeMets (1997) reported that, in some trials, statistical centers did not perform adequately and needed to be replaced. Competent statistical support in the form of analysis supervision, analysis duplication, or analysis conduct can often be found among data analysts, epidemiologists, or any competent colleagues. A professional statistician is not always necessary for this purpose.

Adjustments of the initial statistical analysis plan should preferably receive scientific and ethical oversight or guidance of some sort. In clinical trials a DSMB may oversee amendments to the statistical analysis plan (DAMOCLES Study Group 2005). Unfortunately, such committees nowadays tend to dissolve once all data are collected, which is the stage when the need for protocol changes relevant to analysis often becomes apparent. For other types of studies it is less clear how the oversight should be organized. Ideally, the initial analysis plan should describe what exactly will be done if certain assumptions fail or difficulties arise, but obviously not every scenario can be foreseen.

Panel 26.2 Checklist of Simple Measures to Prevent Analytical Errors and Data Fraud

- Inform data analysts/statisticians about study procedures and subject matter but try not to inform them about the investigator's or sponsor's expectations about direction or size of effect
- Establish and maintain frequent and good quality communication between investigators and analysts/statisticians
- Always involve more than one person in data analysis, even in small studies by single investigators; consider duplicating analyses
- Discuss provisional results with colleagues
- Seek assistance from experienced statisticians and statistical units; organize proper supervision of junior or relatively inexperienced data analysts
- Emphasize that all persons involved should recognize the limits of their own knowledge and skills and ask for advice when appropriate
- Reserve enough time for analysis; do not put excessive time pressure on analysts/statisticians
- Consider analyzing only after a firm and explicit decision to publish has been taken within the group and an internal agreement has been reached that this decision should not be altered by the results of the analysis, especially not by the significance of results or sizes of effects
- Make it a policy to save all analysis syntaxes and outputs, including those that relate to checking assumptions underlying statistical procedures
- Require firm justification for any data editing at the analysis stage
- Report results with and without deleted outliers or imputations
- Do not let any choices in procedures be inspired by a hunger for desired results
- Explicitly plan for data sharing after study completion

26.4 Interactions Between Data Analysts and Investigators

A study may involve many data analysts working in one or more teams, and members of any given team may be located in different parts of the world. One team might be working on the main analysis or separately focusing on different occurrence relations. Consequently, an integral part of any data analyst's job is to interact with and report to investigators and other analysts/statisticians on a regular basis.

Before analyses are initiated, the analyst or team(s) should work with the investigators to define how, exactly, proposals to modify the existing database should be processed. Each analyst or team may detect errors in the database or have useful proposals for deletions, corrections, or additions of derived variables to the database.

These changes may benefit other analyses, and failure to deal with proposed changes in a timely manner can delay analyses unnecessarily (and consequently waste time and resources). An example is the discovery of an outlier with high influence on a study variable that is also used in parallel by another analyst or team. Unless the entire research group reaches consensus on how to handle the change, both analysis teams may be precluded from proceeding. Thus, in any study, all analysts and investigators need to be promptly alerted to proposals for database edits.

Hint

One should be careful not to mistake analysis datasets for database files and should always make this distinction clear in the file name and in correspondence.

Not all analysts/statisticians are involved in study planning; some are employed as statistical consultants at a later stage. All analysts/statisticians, especially those who are not involved in study planning, have the duty to verify independently the soundness of analytical plans or suggestions, to discuss any misguided choices, and to propose better ones even if the analysis plan was previously supported by the investigators, ethics committee, steering committee, or sponsor. In some small studies, there may be no formal analysis plan in place, and the investigator may leave the correct choice of tests or outcome parameters to the statistician. And in some cases, analysis plans approved by ethics committees may be inappropriate. If no plan exists or if changes to a pre-existing plan are necessary, it is advisable to consult with the investigators and, if necessary, to work with investigators to submit necessary amendments or notifications to the ethics committee, steering committee, or sponsor.

Hint

If an investigator expresses a desire for a particular outcome of the analysis, such as a significant P-value, the approached analyst/statistician should, in accordance with the ethical guidelines from the American Statistical Association (1999), advise the investigator to recognize that valid statistical results cannot be guaranteed to conform to expectations (Horner and Minifie 2010).

26.5 Ethical Responsibilities of Data Analysts

Outlined in this section are some responsibilities pertinent to any persons involved in statistical analyses of epidemiological studies, with special attention to roles in the data analysis process. As mentioned in Textbox 26.1, general ethics guidelines from statistical professional associations are available, and in this section some of the advice in those guidelines is extracted and translated into operational responsibilities of data analysts. Concrete responsibilities are best understood in the context of the general principles of epidemiology outlined in Chap. 1 (*See:* Panel 1.1), so the advice in this section is organized accordingly:

- *Minimize risk of avoidable, unacceptable harm*
  - If information in the analysis dataset suggests that this principle may have been taken too lightly in the case of a particular participant, seek clarification from the investigator for each suspected incident on a case-by-case basis
- *Respect for autonomy of participants*
  - Before beginning analyses, one should check with the investigator whether informed voluntary consent was obtained from all participants; if documentation is unavailable for a participant, that individual should not be included in analyses
  - In one's analyses, take care not to include data on subjects who have requested that they be excluded from the analysis (the right to withdraw at any stage of the research includes the right to request not to be analyzed after data are collected)
- *Respect the privacy of participants and confidentiality of their data*
  - Avoid analyzing data from studies that violate this principle
  - Return to the sender datasets that contain personal identifying information (and, of course, do not analyze these datasets until identifying information is removed)
- *Minimize burden, preserve safety, and maximize benefit for participants*
  - Do not analyze data from a study that, by design, has intentionally inflicted avoidable harm, such as starvation research or experimental research on the health effects of weapons or potential weapons
  - Avoid using studies in meta-analyses that have violated this principle
  - Remember that people could be harmed if incorrect analyses produce results that lead to wrong decisions about diagnosis, treatment, prognostication, or policy development
- *Maximize societal relevance*
  - Avoid presenting statistical findings in cryptic or overly sophisticated ways so that they are unlikely to be understood
- *Contribute minimally biased evidence to the overall pool of evidence on an issue*
  - Take measures to prevent analytical errors, as listed in Panel 26.2
  - Refrain from tampering with data or fabricating data
  - Gain the necessary competence before commencing an analysis
  - Ask for advice when appropriate
  - Discuss with the investigators any concerns about the study design, especially when developing the analysis plan
  - Decisions made during the analysis process should not be inspired by hunger for statistical significance or career advancement. Therefore one should be mindful not to exclude records, trim data distributions, or exclude outliers from the analysis if that decision is partly or fully inspired by the expectation of more favorable results (DeMets 1997; Eliades et al. 2005)
  - Data cleaning involves the examination of analysis datasets (*See:* Chap. 19); before running the analysis, analysts should screen the analysis dataset for

remaining suspect values and suspect patterns; strange patterns may be caused by fraud (Buyse et al. 1999; Eliades et al. 2005), and such a discovery needs to be dealt with responsibly
  - If there is a need for editing of data values, it is good to do so only after discussion with the investigator and with appropriate annotation of the database
- *Maximize completeness of data for analysis and archiving*
  - Make back-up copies of analysis datasets after extraction from the database; this helps to restore inadvertent deletions or mishandling of data
- *Guarantee verifiability of study procedures*
  - Make available the analysis syntaxes ('macros') used in connection with the statistical package; this includes saving, naming, dating, describing, and backing up macros
  - Properly document in an audit trail any data value edits deemed necessary during analysis
- *Pursue parsimony*
  - Avoid making statistical models more complex than necessary

Textbox 26.1 Ethical Guidelines on Statistical Analysis

Valuable ethical guidance documents for statisticians are available from professional organizations, such as the International Statistical Institute (2010) and the American Statistical Association (1999).

These documents deal with the involvement of statisticians from design to publication and are a useful resource about general responsibilities, not only for professional statisticians in and beyond health sciences but also for data analysts, investigators, and students taking on a role in data analysis. Important principles contained in these documents are extracted and synthesized in this chapter; however, a more detailed set of guidelines is found in these primary sources.

This chapter marks the end of Part IV: Study Analysis. The result of study analysis is a set of outputs from statistical packages, perhaps further processed into an analysis report. Next, one is faced with the challenge of making these findings known to scientific and other stakeholders. Doing so will hopefully contribute to the achievement of new knowledge, upon which actions can be based. This activity is called study reporting (Part V) and is covered in the remainder of the book.

References

Altman DG (1980) Statistics and ethics in medical research. V: analysing data. BMJ 281:1473–1475

American Statistical Association (1999) Ethical guidelines for statistical practice. http://www.amstat.org/about/ethicalguidelines.cfm. Accessed Sept 2012

Baerlocher MO et al (2010) Data integrity, reliability and fraud in medical research. Eur J Int Med 21:40–45

Buyse M, George SL, Evans S et al (1999) The role of biostatistics in the prevention, detection, and treatment of fraud in clinical trials. Stat Med 18:3435–3451

DAMOCLES Study Group (2005) A proposed charter for clinical trial data monitoring committees: helping them to do their job well. Lancet 365:711–722

Davis MS, Riske-Morris M, Diaz SR (2007) Causal factors implicated in research misconduct: evidence from ORI case files. Sci Eng Ethics 13:395–414

DeMets DL (1997) Distinctions between fraud, bias, errors, misunderstanding, and incompetence. Contr Clin Trials 18:637–650

Eliades T, Athanasiou AE, Papadopulos JS (2005) Ethics and fraud in science: a review of scientific misconduct and applications to craniofacial research. World J Orthodont 6:226–232

García-Berthou E, Alcaraz C (2004) Incongruence between test statistics and P values in medical papers. BMC Med Res Methodol 4:13

Horner J, Minifie FD (2010) Research ethics I: historical and contemporary issues pertaining to human and animal experimentation. J Speech Lang Hear Res 54:S303–S329

International Statistical Institute (2010) Declaration on professional ethics http://www.isi-web.org/about/ethics-intro. Accessed Sept 2012

Jeng M (2006) Error in statistical tests of error in statistical tests. BMC Med Res Methodol 6:45

Martinson BC et al (2010) The importance of organizational justice in ensuring research integrity. J Empir Res Hum Res Ethics 5:67–83

Swazey J, Anderson M, Lewis K (1993) Ethical problems in academic research. Am Scientist 81:542–552

Part V

Study Reporting

Interpretation of Findings

27

Jan Van den Broeck, Jonathan R. Brestoff,
and Ingunn Engebretsen

However beautiful the strategy, you should occasionally look at the results.

Winston Churchill

Abstract

Every study is designed and carried out with the expectation that it will have value. The expectation at the early stages of the study is that scientifically valid and ethically collected evidence about the research questions will contribute to increased knowledge, and that the study findings will be important for decision making about further research, public health policy, or patient care. Near the end of the study, the time has come for all stakeholders to check if these early expectations have been met. It is time to interpret the obtained statistical evidence in the light of the achieved internal validity, and to reflect on the generalizability of the findings and on possible lines of action supported by them. In this chapter, we

J. Van den Broeck, M.D., Ph.D. (✉)
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

I. Engebretsen, M.D., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway

Department of Child and Adolescent Psychiatry, Haukeland University Hospital,
Bergen, Norway

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_27,

argue that this evaluation is mainly the task of peer reviewers and other critical readers or listeners. Investigators should provide the necessary objective and unambiguous information to make the task possible. That information should consist of correctly described statistical results and a complete account of issues relevant to study validity.

27.1 The Role of Investigators in Study Interpretation

27.1.1 Interpreting One's Own Study Results

The main role of authors in reporting study findings is to provide readers with all the information needed to make their own interpretations. This can be done by providing a correct description of statistical results and an objective summary of issues pertinent to internal and external validity (discussed below; *See also* Panel 27.1). These pieces of information will help readers to appraise the evidence provided, to place it in a wider context, and to consider possible lines of action. In reality, however, interpreting one's *own* study results is commonly allowed, expected or even required in scientific papers. This is inevitably a partly subjective activity, especially when the researcher tries to come to an evaluation of the *importance* of her/his own study.

Subjectivity can also come in when the researcher makes decisions on the presentation of findings to achieve 'maximal impact.' An enthusiastic and ambi-

Panel 27.1 Selected Terms and Concepts Around Interpretation of Study Findings

Association A statistically significant relationship

Causal association A statistically significant and confounding-adjusted relation between an exposure and an outcome. *Syn.* Causal effect

Causal criteria List of conditions that must be satisfied before inferring that an association is causal

Clinical relevance Potential to bring about a change in clinical practice

Difference (1) Inequality (2) Result of subtraction

Evidence (provided by a study) Outcome parameter estimates together with information on internal validity of the study

Internal validity Freedom from biases resulting from deficiencies in study design or implementation

Significance (1) (Statistical -) Size of P-value relative to the chosen significance level of the test (usually expressed with the labels 'significant' or 'non-significant') (2) Importance

Trend Modeled shape of relationship

tious researcher cannot easily avoid thinking about how important the results would *appear to others* conditional on a suggested interpretation and a chosen wording and presentation. Indeed, researchers are under pressure to promote their work, emphasize its impact by publishing in high-impact journals, and disseminate their results through various activities (*See:* Chap. 30: Dissemination to Stakeholders).

Additionally there is an obligation to quickly and objectively inform stakeholders of new evidence, preferably in such a way that they will become attentive and motivated about it. Stakeholders must be enabled to use the work for adapting health care policies and practices, but this inevitably induces their 'subjective' appreciation of the importance of the new evidence and what it adds to knowledge. For many intellectuals these tasks are perceived as dilemmas and constitute a very challenging aspect of being a researcher. Part of a solution could be to involve external experts in contributing to the discussion sections of scientific articles, or, more generally, to involve independent auditors and external experts in study reporting.

27.1.2 Helping Stakeholders to Make Their Own Interpretations

In epidemiology new evidence comes, in its purest form, as estimates and P-values. In scientific papers and presentations this statistical evidence is reported in the results section. A reader cannot interpret these statistics without also assessing internal validity of the study. Therefore results sections must also contain an account of study implementation difficulties as possible sources of bias. Results sections should also be clear, if relevant, about reasons why certain findings are reported in detail and others are not shown or shown in less detail. The reader should critically weigh all this information together with her/his critical assessment of the study design. In current publication practice, investigator-authors are allowed to make their own subjective interpretations, but these are supposed to be mentioned only in the discussion section of the paper (*See:* Chap. 28). The authors' description of the statistical evidence in the results section should be free of interpretation so that readers can make their own assessments and are not led only to a particular interpretation.

The next sections of this chapter highlight issues in the presentation and interpretation of statistical evidence and of information on internal validity.

27.2 Direct Interpretation of Statistical Evidence

This section highlights two selected pitfalls in the direct interpretation of statistics, assuming that they have arisen from an internally valid study and express unbiased evidence.

27.2.1 Observed Versus Estimated Values

To enable interpretation, descriptions of statistical findings should be unambiguous. Confusion is often created by poor choice of terms and expressions, for example, by failing to make a distinction between observed data and estimates of underlying population values. For instance, 'We observed an increase' and 'We estimated there was an increase' can mean totally different things. Indeed, it is quite possible to observe an increased value in a study sample whereas this observed change, as a point estimate of a population value change, may be surrounded by such a wide confidence interval that it encompasses a null change. In the latter case it would be inappropriate to state, 'We estimated there was an increase.' In order to avoid this type of confusion, it is better to forego descriptions using 'We observed' in results sections of papers, though this phrase is not strictly 'off-limits.'

Another example of ambiguity occurs in statements of the form: 'the frequency was greater in group *x* than in group *y*.' This may falsely suggest that there was a statistically significant association when that is not the case. For this reason, point estimates must be presented with accompanying interval estimates (confidence intervals).

27.2.2 Association Versus Causal Effect

Another common source of confusion for readers (and for authors themselves) results from failing to make a clear distinction between descriptive and analytical (causal-oriented) research questions. Terminology used to describe statistical associations may suggest that there is evidence of a causal link when this is not the case. The following expressions are commonly used in descriptive as well as in analytical research.

- 'was influenced by'
- 'was determined by'
- 'depended on'
- 'had an effect on'
- 'was associated with'

Effects and associations can be either causal or non-causal, but confusion may arise unless there is absolute clarity about the descriptive versus causally-oriented nature of the research question and study design. The readers should be appropriately reminded of this in the results section, and the reminders can be repeated through appropriate use of adjectives and adverbs such as 'descriptive,' 'causal,' 'descriptively,' and 'causally.' This type of confusion can be aggravated by the equally common 'not significant = no effect' fallacy (*See:* Chap. 23):

- 'was not influenced by'
- 'was not determined by'
- 'had no effect on'

27.3 Internal Validity Assessment

New evidence is rarely indisputably true (within the prevailing paradigm). There are levels of uncertainty around most new evidence, and this uncertainty level can be difficult to evaluate objectively because of unmeasured confounders or other suspected biases of unknowable size. This makes evaluation of internal validity partly subjective, even if investigators provide extensive and correct information relevant to validity. According to the official definition, internal validity is the degree to which a study is free from bias and depends on the soundness of study design, conduct, and analysis (Porta et al. 2008). The present section will only draw attention to selected pitfalls in assessing internal validity.

27.3.1 Errors in Outcome Variables

Errors in assessing outcome variables can greatly affect a study's internal validity by generating bias (*See:* Chaps. 22 and 29). However, when the outcome parameter is a ratio or a beta-coefficient, its estimate can be unbiased even if the outcome estimates for the separate exposure categories were all seriously biased. For example, consider a scenario in which the true *relative risk* of case fatality for male versus female hospitalized patients is 2 (males had twice the case fatality rate of females), where the true case fatality rate is 12 % in males and 6 % in females. In a study it may happen that for the males a seriously biased estimate of 6 % was made and an equally seriously biased estimated of 3 % in females. Yet, in spite of the enormous biases in the separate estimates, a correct ratio estimate of 2 was made in the study. Generally, if bias was present in an incidence risk/rate estimate but authors can argue that the extent of this bias is *proportionally* the same across levels of exposure, then an argument is made in favor of an unbiased incidence rate ratio/relative risk estimate. However, if it is clear that outcome rate estimation bias was proportionally greater in one exposure group than in another, the only way to salvage internal validity is to assess the magnitudes of bias in the separate groups followed by an adjustment in the calculation of the ratio. Thus, the degree of bias in common outcome parameter estimates in epidemiologic research (e.g., odds ratios, rate ratios, and rate differences) cannot be judged simply based on evidence of error in outcome detection.

27.3.2 Random Errors in Determinant Variables

Imprecision, or random error, in determinant variables tends to attenuate (make closer to the null value) estimates of odds ratios, rate ratios, and regression slopes. When the extent of imprecision in the determinant variable is unknown but suspected to be considerable, the outcome parameter estimate can often be interpreted as being an underestimate (i.e., biased towards the null value). *Effect size attenuation*

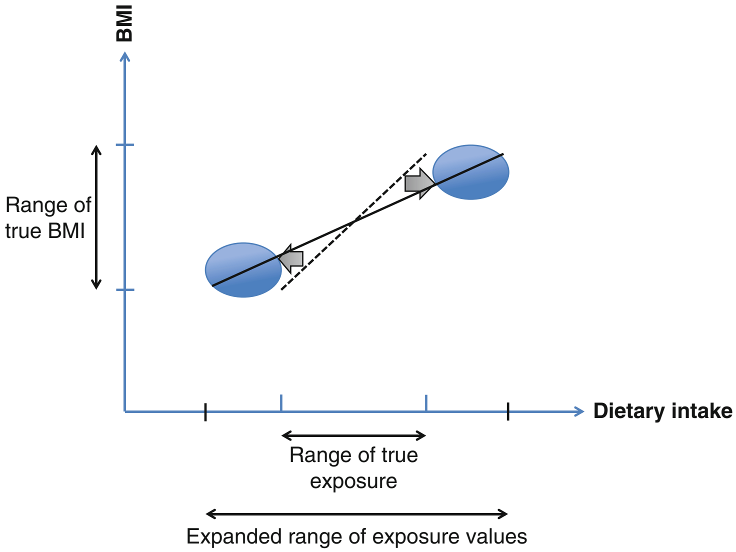


Fig. 27.1 Illustration of *regression dilution*. The *dotted line* depicts the true regression line and true ranges of outcomes values and exposure values. Erroneous measurement values for the exposure (*blue clouds*) expand the range (*arrows*) of observed exposure values (they increase the variance of the exposure). The result is a regression line (*full line*) with a decreased slope, and thus an underestimation (attenuation) of the effect of the exposure

applies to imprecise continuous determinants as well as to randomly misclassified categorical determinants. The reason for this attenuation phenomenon, also known as 'regression *dilution*,' is illustrated in Fig. 27.1, in which an association is being drawn between body mass index (BMI) and dietary intake. The dashed line shows the regression line in a scenario with no errors in the study variables, meaning that all measured values will fall within the true range of outcome (y-axis) and exposure (x-axis) measurements and that the regression line between this exposure and outcome represents a true regression line. The ovals denote imprecision in the measurements of dietary intake, the consequence of which is that some measurement values will fall outside the true range of exposure. Such imprecision tends to expand the variance of exposure, the effect of which is to decrease the slope of the regression line (shown by a solid line). Thus, imprecision in the determinant/exposure tends to lead to an underestimated effect size.

In Table 27.1, an equivalent scenario is used with a categorical determinant, where relative risk is the outcome parameter. Here, attenuation of the crude relative risk estimate is shown to be the result of random error in determinant measurement. In this example, 1,000 individuals are classified as exposed and 1,000 as unexposed. In each case, 10 % of the individuals were misclassified (e.g., of the 1,000 individuals classified as exposed, 100 were truly unexposed, and of the 1,000 individuals classified as unexposed, 100 were truly exposed). The true risk of the outcome is 10 % in those who are truly exposed and 1 % in those who are truly unexposed. Therefore,

Table 27.1 Illustration of attenuation of a crude relative risk estimate by random misclassification of the exposure

Classified exposure status *(number of participants)*	True exposure status *(number of participants)*	True risk (%)	True relative risk	Estimated risk (%)	Estimated relative risk
Classified as exposed *(N=1,000)*	Truly exposed *(N=900)*	10	**10.00** *(10 % of 1,000 truly exposed divided by 1 % of 1,000 truly unexposed)*	9.1 *(90 + 1 out of 1,000 classified as exposed)*	**4.79**
	Truly unexposed but misclassified as exposed *(N=100)*	1			
Classified as unexposed *(N=1,000)*	Truly unexposed *(N=900)*	1		1.9 *(9 + 10 out of 1,000 classified as unexposed)*	
	Truly exposed but misclassified as unexposed *(N=100)*	10			

the true relative risk is 10.00. But because of the misclassifications, the estimated risks for the exposed and unexposed groups were 9.1 % and 1.9 %, respectively. The corresponding estimated relative risk is 9.1 divided by 1.9, which equals 4.79. This excise suggests that if one out of ten participants have their exposure status misclassified, a true relative risk of can get attenuated by more than 50 % (e.g., from 10.00 to 4.79).

27.3.3 Systematic Error in Determinant Variables

Both Fig. 27.1 and Table 27.1 use – for didactical purposes – simplified scenarios where there is only random error in the measurement of the determinants. *Systematic error* can counteract or even reverse attenuation. For example, in Fig. 27.2, it is illustrated how underreporting low dietary intake and over-reporting of high dietary intake tends to result in an overestimation of the strength of association with the outcome.

Taking the messages of Figs. 27.1 and 27.2 together, the net result of random measurement imprecision combined with some systematic misreporting of true extreme values (towards the mean) may be a correctly estimated regression slope, but it may also lead to a biased slope estimate in either direction. As an exercise, one may wish to consider a scenario in which there is systematic error in values around the mean (i.e., under-estimated or over-estimated) but correct values at the extreme exposure levels. Whether a systematic error in the determinant will lead to an over-estimate or under-estimate of the determinant-exposure association depends on the particular pattern of systematic error.

In Table 27.2 a similar scenario is described that, again, shows that systematic error can lead to overestimation of the strength of association. In this example, individuals

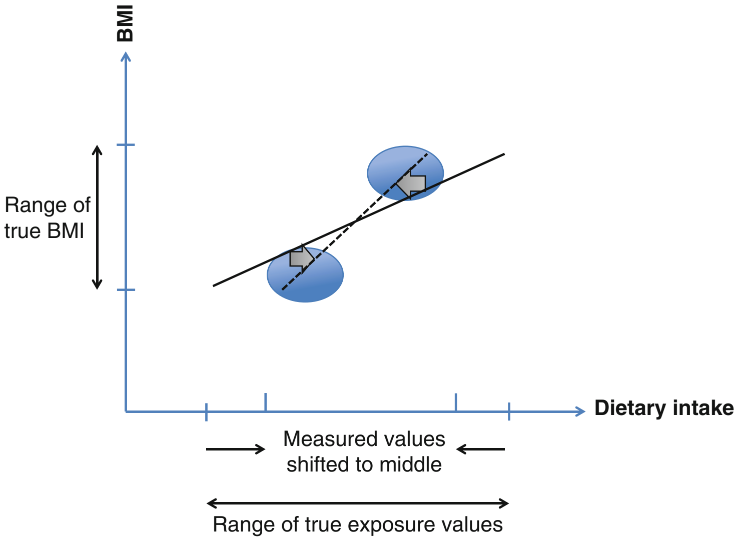

BMI
Range of true BMI
Dietary intake
Measured values shifted to middle
Range of true exposure values

Fig. 27.2 Illustration of a case of overestimated strength of association by systematic errors in measurement of a continuous determinant. The *dotted line* represents the true regression line. *Arrows* indicate the shift of outcome values (*blue clouds*) to the middle by a systematic trend for underestimation of high values and overestimation of low values. The result is a regression line (*full line*) with an increased slope and thus an overestimation of the strength of the relation

Table 27.2 Illustration of overestimation of a crude relative risk by a systematic error in measurement of a multi-level categorical determinant

Classified determinant level *(number of participants)*	True determinant level *(number of participants)*	True risk (%)	Estimated risk (%)	True relative risk for level	Estimated relative risk for level
Level 1 *(N=80)*	Truly level 1 *(N=80)*	20	20	4.00	4.00
Level 2 *(N=120)*	Truly level 2 *(N=100)*	10	11.7 (*14/120*100*)	**2.00**	**2.34**
	Truly level 1, mis-classified as level 2 *(N=20)*	20			
Level 3: = reference level *(N=100)*	Truly level 3 *(N=100)*	5	5	1.00	1.00

are categorized into three levels of exposure: 1, 2, and 3. All of the individuals in Levels 1 and 3 are correctly classified, but 20 of the 120 individuals in Level 2 were misclassified as such and truly should have been classified as Level 1. The true risks for Levels 1, 2, and 3 are listed as 20 %, 10 %, and 5 %, respectively. Since Level 3 is listed as the reference group, this means that the true relative risk for Level 1 is 4.00 and for Level 2 is 2.00. However, the misclassification in Level 2 produced an estimated risk of 11.7 % (rather than the true 10 %). The effect of this error is to exaggerate the estimated relative risk in Level 2 as 2.34. The distorting effect of systematic errors depends on the levels of the determinant at which the errors occur and in what directions they exert their effects. For example, if the error consisted only of a proportion of Level 3 subjects misclassified as Level 2, the result would have been an underestimation of the relative risk for Level 2, rather than an overestimation, as in the tabulated scenario.

When the exposure has only two levels, as is often the case in epidemiologic studies, systematic error can consist of one of the two levels being more frequently misclassified than the other. In those common scenarios, importantly, the net result is systematic error superimposed on random error. This situation tends to further attenuate the ratio estimate compared to random error alone. Exceptions sometimes occur in scenarios where the systematic error in determinant assessment is prognosis-related (in cohort studies) or case status-related (in case–control studies). This point is illustrated in Table 27.3 using some hypothetical cohort study scenarios. Scenario C in this table shows that over-estimation can occur.

Table 27.3 Illustration of the effect of systematic misclassification of exposure on the crude relative risk estimate in cohort studies

Level of determinant, as classified *(number)*	Good or bad prognosis[a] *(number)*	True risk[a] (%)	New cases developing	Total cases developing	Observed cumulative incidence (%)
Scenario A: No errors in study variables:					
Level 1	G$_1$ *(N=20)*	5	1	9	9
(N=100)	B$_1$ *(N=80)*	10	8		
Reference level	G$_0$ *(N=80)*	5	4	6	6
(N=100)	B$_0$ *(N=20)*	10	2		
True crude relative risk = 1.50					
Scenario B: Systematic error: 20 level-1 subjects are misclassified, non-differentially as to prognosis:					
Level 1	G$_1$ *(N=16)*	5	0.8	7.2	9
(N=80)	B$_1$ *(N=64)*	10	6.4		(7.2/80)
Reference level	G$_0$ *(N=80)*	5	4	7.8	6.5
(N=120)	B$_0$ *(N=20)*	10	2		(7.8/120)
	G$_1$ *(N=4)*	5	1.6		
	B$_1$ *(N=16)*	10	0.2		
Estimated crude relative risk = 1.38 *(under-estimation)*					
Scenario C: Systematic error: 20 level-1 subjects are misclassified, differentially as to prognosis: only subjects with good prognosis are misclassified:					
Level 1	G$_1$ *(N=0)*	5	0	8	10
(N=80)	B$_1$ *(N=80)*	10	8		(8/80)
Reference level	G$_0$ *(N=80)*	5	4	7	5.83
(N=120)	B$_0$ *(N=20)*	10	2		(7/120)
	G$_1$ *(N=20)*	5	1		
Estimated crude relative risk = 1.72 *(over-estimation)*					

(continued)

Table 27.3 (continued)

Level of determinant, as classified *(number)*	Good or bad prognosis[a] *(number)*	True risk[a] (%)	New cases developing	Total cases developing	Observed cumulative incidence (%)
Scenario D: Systematic error: 20 level-1 subjects are misclassified, differentially as to prognosis: only subjects with bad prognosis are misclassified:					
Level 1 *(N=80)*	G_1 *(N=20)*	5	1	7	8.75
	B_1 *(N=60)*	10	6		(7/80)
Reference level *(N=120)*	G_0 *(N=80)*	5	4	8	6.7
	B_0 *(N=20)*	10	2		(8/120)
	B_1 *(N=20)*	10	2		
					Estimated crude relative risk = 1.31 *(under-estimation)*

[a]Among subjects with the reference level of the determinant ('unexposed') there is variation in susceptibility for the outcome, arbitrarily taken to be such that 20 % of subjects (Bad prognosis 'B'-subjects) have, together, a risk that is twice (10 % risk) the average risk for the less susceptible remaining ones (5 % risk for Good prognosis 'G'-subjects); The effect of causal exposure (determinant level 1) is taken to be an increase in the proportion of subjects, arbitrarily up to 80 %, with higher susceptibility

When the determinant is *positively* associated with the outcome (a causative exposure), as in all scenarios in the Table 27.3:

- A systematic trend for misclassification of subjects in the index level of the determinant ('missed exposures') always leads to further attenuation, except when those missed exposures are preferentially subjects with good prognosis
- A systematic trend for misclassification of subjects in the reference level of the determinant ('false exposures,' not shown in table) also always leads to further attenuation, except when those false exposures are preferentially subjects with bad prognosis.

Conversely, when the determinant is *negatively* associated with the outcome (a protective exposure):

- A systematic trend for misclassification of subjects in the index level of the determinant ('missed exposures') always leads to further attenuation, except when those missed exposures are preferentially subjects with bad prognosis.
- A systematic trend for misclassification of subjects in the reference level of the determinant ('false exposures') also leads to further attenuation, except when those false exposures are preferentially subjects with good prognosis.

27.3.4 Making Judgments About Biasing Effects of Measurement Error

Sometimes one can be confident about the existence and direction of bias by taking into account the mechanisms discussed in the previous paragraphs (Textbox 27.1).

Ideally, the investigators reporting on the study would provide the necessary information on error rates, imprecision of observers, factors associated with systematic

Textbox 27.1 Case Scenario Illustrating How a Biased Relative Risk Estimate Resulted from Errors in Exposure Measurement

A descriptive study included newly hospitalized adult patients with hip fracture. The study aimed to estimate the relative risk of dying in the hospital, comparing patients presenting with and without additional fractures. The outcome assessment can be assumed to be without error. However, there have probably been some errors in the assessment of the binary determinant (additional fractures, yes/no). First, some random imprecision from the part of the radiologists and from others involved in assigning exposure levels, will attenuate the relative risk estimate. Secondly, there may have been a systematic trend to misclassify patients *with* additional fractures more so than the other way around, in which case the relative risk estimate would be further attenuated, unless these missed additional fractures mainly concerned patients with good prognosis. In the example, it could well be the opposite, where the missed additional fractures were more likely to be patients with bad prognoses, which would even more attenuate the relative risk estimate. Thus, overall, it seems highly likely that in this study the relative risk estimate was attenuated; therefore, associations coming from this study are likely to be under-estimates of the true association.

errors, etc., to allow such intuitive judgments as in Textbox 27.1 to be supported with some data. In general, any fair judgment of possible over- or under-estimation of outcome parameters needs to be based on:

- Information provided by investigators about measurement errors, possibly including results of attempts to adjust for bias or attenuation in the analysis. Frost and Thompson (2000) discuss methods of correcting for regression dilution.
- Knowledge about how random and systematic measurement errors tend to operate for the particular determinant or outcome at issue, and for the type of measurement setting and study scenario.

In practice, such detailed information is often unavailable, in which case interpretation cannot go beyond a careful expression of suspicion and a call to consider the statistics dubious.

When the determinant-outcome relation is studied with adjustment for other factors, measurement error in these covariates becomes an additional concern. In such scenarios, systematic errors in the covariates can have complex effects on the relative risk estimates that are more difficult to predict than in the bivariate scenarios discussed above.

This sub-section has dealt mainly with the effect of measurement error on outcome parameter estimates in descriptive studies, but most of it is also relevant to analytical studies (next sub-section).

27.3.5 Special Considerations of Internal Validity in Analytical Studies

In analytical studies, including trials, the considerations discussed above regarding errors in study variables apply; however, adjustments for multiple covariates are more often done in analytical studies because (potential) confounders need to be entered in the models as covariates. For the adjusted estimates to be valid and interpretable, the covariates need to be measured with excellent precision and accuracy. After data analysis, the researcher of an analytical study may be faced with several types of situations with respect to internal validity. Some of the common situations are:

1. Situation-1: The result of the study may be that, in full accordance with the study protocol, an 'internally valid' estimate has been produced and/or a null hypothesis endorsed to some extent. The likelihood of the results being due to an unmeasured confounder or another design flaw is considered negligible. The decision that bias is negligible may be partly based on whether, apart from exposure of main interest, all other major determinants of the disease outcome are known with reasonable certainty, measured reliably, and accounted for. In many instances, however, that one cannot think of any major neglected confounder does not mean that there is none.
2. Situation-2: The result of the study may be that an estimate has been produced with an anticipated precision and/or a hypothesis endorsed to some extent but some time during the study the investigators have come to the realization that *study design could have been substantially better* by, for example, inclusion of a neglected confounder or a more precise measurement method for an important variable. This realization may have come through maturation of thought, through recent availability of novel scientific knowledge on the topic or by external advice. The problem may, in the ideal case, have led to additional sensitivity analyses with some estimate of the possible size of the internal bias. In practice, however, sensitivity analyses are seldom performed and tend to be based on assumptions that are themselves surrounded with unquantified uncertainty. Sometimes the forgotten or unmeasured potential confounder can be shown to be unassociated with the determinants. Finally, in research aimed at the detection of the existence of a causal link, it can sometimes be argued that the covariate, albeit not taken into account, would have only increased the observed effect.
3. Situation-3: The result of the study may be that study design was appropriate and an estimate has been produced and/or a hypothesis endorsed to some extent, but increased uncertainty has arisen as to the level of internal validity because of *problems during study implementation*. Many different problems may exist simultaneously. For example, one of the confounders has a large number of missing values (of which the randomness is questionable), poor performance of study personnel at some stage, and losses to follow up for unknown reasons. These problems may, in the ideal case, have led to additional sensitivity analyses with some estimate of the possible size of the internal bias and adjustments.

27.4 Causal Interpretation Criteria

In analytical studies internal validity is the main key to causal interpretation of a statistical association. Some researchers and critical readers find it useful to re-evaluate an analytical study at the interpretation stage using checklists of causal criteria. Historically, the most used are the Henle-Koch postulates and the Hill criteria. Though these criteria have major limitations in the applicability in modern epidemiology, they have been important facilitators of causal interpretation of research findings.

27.4.1 Henle-Koch Postulates

These criteria, known as the Henle-Koch postulates, are listed in Panel 27.2. They were created to demonstrate, via a series of related studies, the effect of a single infectious agent on a disease. The contemporary usefulness of these criteria is limited and depends on whether, in the study at hand, the phenotypic case definition is detailed enough to specifically match the sub-typing of the infectious agent, provided that such a one-to-one relationship between sub-type and disease phenotype actually exists.

27.4.2 Hill Criteria

These criteria were proposed by Austin Bradford Hill as an aid for causal interpretation in observational analytical studies (Hill 1965). An adapted version of these criteria (based on Glynn 1993) is listed in Table 27.4, together with some caveats about their usage and relationship with confounding. The relevance of these criteria tends to increase with decreasing internal validity. If internal validity is excellent, most criteria lose their relevance. For example, when a study is (1) near-perfectly designed, (2) near-perfectly conducted, and (3) completely free from confounding, there is no reason why a strong association would be more likely to be causal than a weaker association.

Panel 27.2 The Henle-Koch Postulates *[Rephrased]* to Demonstrate, Via a Series of Studies, the Causal Effect of a Pathogen on a Disease

- The pathogen must be shown to be present in every individual with the disease
- The pathogen must not be found in cases of other diseases
- The pathogen should reproduce disease in experimental animals
- The pathogen should be recovered from the diseased experimental animals

Table 27.4 Caveats regarding the use of 'Hill criteria' [*rephrased*] for causal interpretation

Hill criterion	Caveats
Strength of association – A strong association is more likely to be causal than a weak association	A strong confounder can cause a strong association. Thus this criterion has value only in situations where one is sure that any forgotten confounders are weak
Dose-response relationship – A dose–response relationship is more likely to be causal than another relationship	A confounder with a graded effect can cause an apparent dose–response relationship
Temporality – Cause must precede outcome; otherwise there cannot be a causal relationship	This is always true, but reverse causality (*See:* Chap. 2) does not preclude causality. Both can exist simultaneously. Temporality of the confounders must also be taken into account
Consistency – If the results are consistent with other findings, a causal interpretation is more often justifiable	A consistent confounder can cause a consistent error. A flawed study can be consistent with other flawed studies
Specificity – When the association is found to be specific to one illness or one defined set of illnesses and not other illnesses, a causal link is more likely	Less relevant to chronic non-communicable diseases, which tend to have multiple causes, each of which becomes a confounder when the effect of one factor needs to be singled out
Plausibility – If one can point to a plausible underlying mechanism, the association is more likely to be causal	Something plausible can be very wrong. Many plausible causal associations were later proven to be due to confounding
Reversibility – Removal of the study factor leads to a decrease in the outcome occurrence	For those affected, reversibility depends on how irreversible effects are. For those not yet affected, reversibility depends on how important the cause is relative to that of other causes

In contrast, if considerable doubt exists regarding internal validity, there may be some value in looking at the strength of association. For example, imagine that one is convinced that all unaccounted confounders are very weak, then those confounders likely cannot explain a strong association, and the strong association will therefore more likely have a causal component. Similarly, if one is convinced that none of the 'forgotten' confounders can have a dose–response relationship with the outcome, then an observed dose–response relationship between study factor and outcome will more likely represent a causal effect. The applicability of most Hill criteria thus reduces to questions like: how sure are we that the 'forgotten' confounders are only weak and cannot have a dose–response effect? Indeed, most Hill criteria have disputable and highly conditional applicability.

27.4.3 Causal Interpretation in Ecological Studies

Drawing false conclusions from ecologic studies about effects of individual-level factors is called the ecological fallacy, also called ecological bias. Causal inference is possible in ecological studies, but the inference is, in principle, about group-level phenomena, not about processes in individuals. Ecologic data can warrant further etiologic research on individual-level factors but can rarely demonstrate causality at that level. Only in very rare circumstances is such individual-level inference convincingly justifiable (Greenland 1992; Morgenstern 1998).

27.4.3.1 Ecological Fallacy

The ecological fallacy is encountered frequently in ecological studies. Perhaps the most frequent misinterpretation is of the type: 'since there were better average health outcomes in communities where a health intervention was introduced, the intervention was effective in the individuals who accepted and underwent the intervention'. A hypothetical example of ecological fallacy is described in Textbox 27.2.

The textbox shows that better average health outcome in communities where a health intervention was conducted may still mean that individuals who used the new services were worse off. Even if the outcome is more frequent or intensive in groups with more frequent exposure, the outcome may still be less frequent in individuals with the exposure.

27.5 External Validity and Study Implications

With the statistics evaluated in the light of internal validity, the critical appraiser can proceed to further interpretation. The additional interpretations may be perceptions about the severity of the health problem studied or the unexpectedness and

Textbox 27.2 An Example of Ecological Fallacy

As an illustration of **ecological fallacy**, consider two areas, area A with a low coverage of the health intervention and a high post-intervention morbidity rate, and a similarly sized area B with high coverage and a much lower morbidity rate. This might wrongly suggest that, on average, the health intervention benefited individuals who underwent it.

In reality, the intervention might have been beneficial, or it might have had no effect at all or an adverse effect. Consider this scenario:

Area A (10,000 inhabitants):

- 30 % took up the intervention, n = 3,000; morbidity rate, 1,300/3,000 (43 %)
- 70 % refused intervention, n = 7,000; morbidity rate, 700/7,000 (10 %)
- Total morbidity rate for area A = (1,300 + 700)/(3,000 + 7,000): **20 %**

Area B (10,000 inhabitants):

- 70 % took up the intervention, n = 7,000; morbidity rate, 700/7,000 (10 %)
- 30 % refused intervention, n = 3,000; morbidity rate, 300/3,000 (10 %)
- Total morbidity rate for area B = (700 + 300)/(7,000 + 3,000): **10 %**

(continued)

Textbox 27.2 (continued)

In this scenario, the total observed morbidity rate (areas pooled) is actually twice as high in individuals who underwent the intervention than in those who refused it! Indeed, the morbidity rate among individuals who took up the intervention was 20 % [2,000 illness cases (1,300 from area A, plus 700 from area B) out of a total of 10,000 who took up the intervention (3,000 from area A, plus 7,000 from area B)]. Among individuals who refused the intervention the morbidity rate was only half of that, 10 % [1,000 illness cases (700 from area A, plus 300 from area B) out of a total of 10,000 who refused the intervention (7,000 from area A, plus 3,000 of area B)].

Thus, although area B had higher uptake and lower total morbidity than area A, a conclusion that this is evidence for effectiveness of the intervention would be a serious case of ecological fallacy in this scenario.

originality of the findings. They may concern appreciations of external validity, namely opinions about contributions made to the overall evidence on the topic and to general knowledge. Finally, ideas may be formulated about how the study evidence or the updated overall evidence supports certain actions to be pursued by investigators, the community, caregivers, or health authorities.

27.5.1 Perceptions About Severity, Unexpectedness, and Originality

Of course, every investigator is allowed to have her/his own individual perceptions and emotions around study findings, but these are in principle of limited interest. Some interest may be gained when the feelings are based on the findings' discrepancy or conformity with reasonable prior expectations: there needs to be a reasonable point of reference as a justification. For example, expressions such as '*strikingly* prevalent,' '*substantially* higher/lower in group A,' or '*remarkably* higher/lower in group B' are interpretations that would better be based on some fact-based prior expectation. It is appropriate to explain what these expectations were and what they were based on. For prevalence and incidence rates, standardization may provide a reference base for expressions such as 'there was *excess morbidity*' or '*excess mortality*'. If researchers wish to interpret their own findings then the anticipated burdens and effects should perhaps be stated in the protocol. Typically, contrasting results with those of similar studies ('*originality* of findings') and, oppositely, comparing results convergent with those of similar studies ('*consolidation*') are both reasons for celebrating the study's importance in the mind of the investigators.

27.5.2 External Validity

According to the official definition, external validity concerns the degree to which results of a study may apply, be relevant, or be generalized to populations or groups that did not participate in the study (Porta et al. 2008). This means that external validity is the strength of the basis for inference beyond the space and time constraints of the study base experience, e.g., about future circumstances and other geographical areas.

Internal validity is a first condition for external validity. Informativeness is a second. Sample size problems may have led to an interval estimate so much wider than anticipated that one considers the results inconclusive or un-interpretable. The same judgment can sometimes be made on the basis of design flaws and serious unadjusted biases. A third condition is that the overall evidence on the issue should be trustworthy. For a discussion of this issue, please see the publication bias section in Chap. 31. Finally, the set of underlying effect modifiers that operated in the study needs to be considered. To the extent that this set is quite unique, generalizing beyond the study population tends to become less justifiable.

27.5.2.1 The Erratic Path from Evidence to Knowledge

The way in which it became known that smoking causes lung cancer is illustrative of the fact that subjectivity plays an important role in the assessment of external validity (Vandenbroucke 1989). Paradigm shifts can resemble wars of thought rather than discoveries, and indeed there were many heated debates about the role of smoking in lung cancer. Thus, no researcher can expect the results of her/his study to provoke a scientific paradigm shift. There are always some colleagues who are more skeptical and critical than others, and some may be hard to convince even in the face of rather hard evidence and near-consensus. Yet, there are elements that make it more likely that a study will have a large impact. In addition to high internal validity, precision, and projected impact on care and research, numerous other factors tend to determine whether study results, if published, will be considered *credible*, *interesting,* or *ground-breaking*, including:

- High prior credibility of the hypothesis
- Unexpectedness of the size of a point estimate (strength of association)
- Novelty of a causal association with very strong supporting evidence
- Large sample size
- Number of zeros in a P-value
- Good reputation of investigators and research institutes

27.5.3 Support for Particular Lines of Action

The usefulness of an estimate tends to heavily depend on the width of the interval estimate. When confidence intervals are narrow, *broad calls* for public health action can be justifiable given the importance of the research question. It is problematic to make conclusions about a strong causative effect of a risk factor (high ‘attributable fraction') in a single observational study as a justification for recommending some *specific action* to combat the risk factor. Causative effects must not be considered as quantitatively identical to effects of removal of the cause, and protective effects demonstrated in observational studies should not be taken to imply guaranteed success of some new intervention policy. The most interesting results from observational studies are often good justifications for calls for further monitoring of the problem and for further research, especially intervention research. Findings of efficacy trials may be a good basis for planning cost-effectiveness studies.

A major theme in the present chapter has been how peer reviewers and other readers critically appraise and interpret one's scientific writings. In the next chapter (Chap. 28: Scientific Writing) we shift the focus to the actual writing about one's study and the evidence it provides.

References

Frost C, Thompson SG (2000) Correcting for regression dilution bias: comparison of methods for a single predictor variable. J R Stat Soc 163:173–189

Glynn JR (1993) A question of attribution. Lancet 342:530–532

Greenland S (1992) Divergent biases in ecologic and individual-level studies. Stat Med 11:1209–1223

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

Morgenstern H (1998) Ecologic studies. In: Rothman KJ, Greenland S (eds) Modern epidemiology, 2nd edn. Lippincott Williams & Wilkins, Philadelphia, pp 459–480. ISBN 0316757802

Porta M, Greenland S, Last JM (2008) A dictionary of epidemiology. A handbook sponsored by the I.E.A, 5th edn. Oxford University Press, New York, pp 1–289. ISBN 9780195314496

Vandenbroucke JP (1989) Those who were wrong. Am J Epidemiol 130:3–5

Scientific Writing

28

Cora Grant and Jan Van den Broeck

The ill and unfit choice of words wonderfully obstructs the understanding.

Francis Bacon

Abstract

The core focus of scientific research surrounds the achievement of new knowledge, a consensus line of thinking that rests on a collective body of evidence. Research-based evidence is considered hidden or incomplete until it is made accessible to the relevant audience. The process of effectively communicating is an art which, when practiced and honed, should increase awareness of and insight into the scientific knowledge-base of public health. Chapter 30 will describe principles and guidelines for communication to a variety of stakeholders. However, communication to scientists specifically has evolved as a separate art with its own principles and practical conventions. The present chapter aims to provide an introduction to the art of scientific writing, the primary form of dissemination to other members of the research community, with practical advice for selected types of it.

C. Grant, MPH (✉)
Department of Epidemiology and Public Health, University College Cork,
Cork, Ireland
e-mail: Cora.Grant@gmail.com

J. Van den Broeck, M.D., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_28,

28.1 Types and Purposes of Scientific Documents and Presentations

Scientific writing can take many forms (Table 28. 1). The ultimate aim of all types of scientific writing is to contribute to better health and wellbeing through the dissemination of health-relevant evidence and/or the development of theory. The scientific writing process can also contribute to a broadening of the writers' own expertise. Indeed, there is an ethical responsibility incumbent upon researchers to publish their results, thereby enhancing their research capacity and sharing their findings with others (Bhopal 2002). Indeed, many researchers have an intrinsic desire to further their own body of published work as a matter of career advancement and personal fulfillment. Additionally, the presentation of research regularly necessitates travelling and networking with fellow professionals, a practice which can benefit the individual and wider research community.

Some important terms and concepts relating to scientific writing are listed in Panel 28.1.

Table 28.1 Common types of scientific documents

Type	Essence and purpose as a method of scientific communication
Research proposal or protocol	Description of study rationale and design. Proposals are submitted for amendment or approval to collaborators, ethics committee and/or stakeholders. Protocols are official version of a study proposal approved by sponsor and ethics committee
Abstract	Provides a succinct overview of object, methods and findings of a research project or develops a flow of expert thoughts on any science-related topic, thereby enticing the reader to read the entire scientific manuscript
Poster	Printed document about a scientific topic, exhibited at a venue where it is meant to catch the attention and be read by ambulant persons with a potential interest in the topic. Often used by young researchers and students as an opportunity for discussion of their research work at scientific meetings
Oral presentation	A scientific presentation that orally addresses the audience with or without the use of supporting audiovisual material, often followed by a discussion with the audience
Original research paper	A scientific document that describes the evidence from a study i.e. the study protocol together with the findings
Review paper	Provides a summary description and evaluation of a specific scientific topic, drawing upon the author's specialized knowledge in the area and/or presenting methods and findings of a meta-analysis
Book	A series of bound printed or assembled electronic papers on the broader topic reflected in the book title
Dissertation or thesis	A presentation and/or document in which a series of arguments are developed and discussed around a scientific topic, in partial fulfillment of the requirements of a university degree

Panel 28.1 Selected Terms and Concepts Relevant to Scientific Writing

Abstract (An -) Written summary of a scientific manuscript or publication

Acknowledgments section Section of a scientific publication for acknowledging contributions made to the research by non-authors

Bibliography List of document identifiers, provided to endorse statements in a scientific manuscript or publication and/or as a source for further reading about the topic at issue

References list List of document identifiers, provided to endorse statements in a scientific manuscript or publication

Journal Newspaper for scientists and/or practitioners

Literature Pool of published scientific documents

Scientific writing Providing a written account of methods used and evidence produced by specific research projects, or, developing written arguments around theoretical epidemiological issues

Structured abstract An abstract whose main structure of content is visualized by section headers

Writer's block Psychological inhibition, often experienced by researchers, to write up and publish the findings of a research study

28.2 General Principles of Scientific Writing

A first fundamental requirement of scientific writing is the use of clear and accurate concepts and terms. Dictionaries and glossaries may be of help. For example, Dorland's Illustrated Medical Dictionary (2011) is widely trusted in the clinical medical community. For epidemiological concepts and terms, there is a dictionary endorsed by the International Epidemiological Association (Porta et al. 2008), which is useful for acquainting oneself with mainstream definitions of concepts. However, like scientific evidence, definitions of concepts are topics for discussion in the scientific community and are amenable to improvement. One epidemiologist who has devoted an important part of his life to thinking about epidemiological concepts and terms is O.S. Miettinen, whose decades of work has led to his recent book, *Epidemiological Research: Terms and Concepts* (Miettinen 2011a). This book is a useful source for any researcher concerned with concepts and the use of terms. We also refer to the terms and concepts of this textbook for further propositions. Note that the sources mentioned are, to date, largely non-overlapping in their entries, meaning that an individual may need to use multiple sources to locate the proper terms and concepts. Indeed, as one consults various sources of terms and concepts, one should carefully weigh the appropriateness of what is proposed and make careful decisions about what is to be propagated in one's own writings.

A second requirement of scientific writing is that any scientific evidence should be presented according to principles of science and epidemiology. This requirement imposes another, to present in one's scientific writing details of the object design, method, and implementation that are relevant to the study's validity. Results are important, but 'the science is in the methods' in the sense that scientific knowledge can emanate only from valid studies. In an attempt to guide writers of scientific papers in this respect, guidelines have been developed for the submission of scientific papers (ICMJE 2011). Mostly as a reaction to common deficiencies in published papers, editors and scientific writing expert groups have issued guidelines and 'statements' that constitute a checklist of necessary or recommended content and style for use by writers of papers on specific types of studies. Sometimes, these recommendations and checklists are also formally used by journals as strict requirements for submitted papers. Examples of such 'statements' are the CONSORT (Consolidated Standards of Reporting Trials) statement, the QUOROM statement for reporting of meta-analyses of clinical trials (Moher et al. 1999), and the STROBE statement (2009) for observational studies.

Scientific writing remains an art. The style of language in scientific writing should be fluid and easily understood. Indeed, it is essential to remember that the objective is to impart information to the reader. The skills include the capacity for using:

- Clear concepts referred to by unambiguous terms
- Familiar lexicon when possible
- Logical sentence structure
- Clarity of expression (keep it short and to-the-point)
- Logical flow of thoughts

A note about tenses

Reference books are available (e.g. Strunk et al. 1999) that describe the correct use of grammar and tenses. Some major points:

- The present tense should be used to state scientific facts and refer to the present study; it is also suitable when referring to tables and figures (e.g., Table 28. 1 indicates that obesity is a major public health concern in countries in all stages of development)
- When referring to findings from previous studies or indeed to those which are still continuing, the present perfect tense (e.g., a previous study has shown…) or past tense (e.g., a previous study showed…) are preferred
- When writing the Methods and Results sections, the past tense should be used. This tense is also used to describe both unpublished findings and those that cannot be generalized
- When referring to observations that occurred before those that you mention in your manuscript, the past perfect tense is suitable (e.g., previous studies had shown that…)

28.3 Preparing to Write a Paper

28.3.1 Organizing a Writing Team

Most manuscripts are multi-author and evolve from collaborations among multiple scientists - sometimes from various disciplines. In this case, it is prudent to negotiate the authorship issues at the project's outset. A key issue for review at the initial stage of the process involves a clear structure regarding the functioning of the writing team, the aim being that potential clashes of team dynamics can be minimized. Commonly, the first author will produce the first draft of the paper, but every co-author should be able and willing to contribute to the manuscript's text. Indeed, each contributor ought to be in a position to defend the paper publicly and answer any questions about it. Given the collaborative effort that may be involved, some members of the writing team are likely to be working on a number of projects concurrently, and frequently in different parts of the world. The practical considerations involved in advancing with manuscript amendments should be discussed collectively. One should avoid engaging in guest authorship, an ethical topic of authorship discussed in Chap. 29.

28.3.2 Selecting a Journal

In the case of multiple authors, a unified decision should be sought about this. When selecting a journal, the first question concerns the target audience: who will be most interested in reading about the research (Neill 2007)? The next questions are often about the following:

- *The impact of the journal* – a high impact journal is commonly preferred when the research question and results surpass a specific, non-general academic interest
- *The scope of the journal* – new research findings with potentially important public health or clinical implications may be best submitted to a general interest journal
- *Level of access provided by the journal* – open access to the paper is increasingly popular, as it tends to offer more readers the chance to access your work
- *Turn-around time of the journal's review process* – particular journals are traditionally regarded as 'slow' or 'fast' in this regard
- *Potential costs charged by the journal* – many journals charge page fees, color printing fees (if applicable), or optional 'open access' fees
- *Journals' directives on prior presentation of the findings* – some journals preclude publication of work that has been previously presented in a format other than a conference poster.

28.3.3 Documents Supporting Scientific Writing

Many journals have developed their own specific instructions for authors and guidelines for manuscript submission. It is recommended to keep at hand any instructions before beginning to write. Moreover, it may be of benefit to select a number of recently published articles from the journal for further orientation. It is also important to note any instructions on the word count of the manuscript, as exceeding this may jeopardize initial approval at the editorial stage of submission. While writing the manuscript, it is useful to keep a number of research-related documents close at hand for reference (Day and Gastel 2008; Toft and Jaeger 1998). Examples include the study protocol or research proposal, questionnaire, data management report, analysis dataset, data dictionary, reports on findings, (preliminary) analysis results, copies of key articles on the subject, summaries of literature review, and epidemiologic dictionaries or glossaries.

28.3.4 When to Start Writing

When considering at what stage to begin working on the manuscript, it may be wise to start the first draft before all the data are collected. Although a literature review should have been completed at the protocol design stage, it may need to be updated regularly. Indeed, the literature review will often provide material for the introduction section and the discussion section of the manuscript. Similarly, the study protocol will include details about the objects and methods and can thus be developed into the methods section of the paper. Finally, the protocol should include the analysis plan which can be helpful to structure the results section. This can also be used to create empty tables and figure axes/legends for the results section.

28.3.5 Suggestions for Dealing with Writer's Block

The phenomenon of writer's block is widespread; particularly amongst young researchers whose mastery of scientific writing may need honing. Scientific writing is a skill that must be learned. Indeed, in an effort to contribute to the scientific knowledge-base, most will seek to develop it. There is also an impetus towards improving writing skills from an ethical perspective, i.e. it is imperative to respect the study participants by seeking to report the results based on the information obtained from them. Thus, combating writers block is an important aspect of epidemiological research practice. The task ahead may appear daunting. Indeed, many experienced writers continuously 'conquer the blank page.' The perceived or actual importance of the writing should not lead to paralysis. There are a number of practical approaches to overcoming writer's block, as further discussed in this sub-section (*See also:* Textbox 28.1).

Textbox 28.1 The Bullet Point Method of Overcoming Writer's Block

A popular approach to writing that often proves successful is writing bullet points which will later be expanded into the full text. Thus, paper writing may usefully start with preparing a traditional slide presentation. This initial flow of ideas should help to create some structure. To begin, consider starting with the section that most appeals, and once a writing flow has been established, do not let issues of spelling and syntax accuracy slow you down. Matters of spelling and syntax can be addressed later in an editing phase.

28.3.5.1 Writing Is a Stepped Process

It is advisable for any writer to create a clear program of priorities, to establish a vision of how to proceed, and to remain steadfast with this vision. Crucial is the setting of a time frame for reaching milestones in the writing. The aim is a stepped approach to progress. However, time scales shouldn't equate to pressured deadlines. They are there only to create gently focused reminders. For first drafts, it can be said that 'the perfect is the enemy of the good.' In other words, seeking perfection can thwart or even halt progress. The aim of the writing is to inform in the clearest, most accurate, and most straightforward manner, but this aim may be achieved only step-by-step. The process of writing may not always flow well once started. It usually entails much editing and re-writing and discussions with other members of the writing group. These steps represent a necessary maturation of thought that adds value and confidence to the developing writer. It is pertinent to remember that 10–20 drafts may be required. Trying to make some progress everyday often proves useful in maintaining a focus and developing a momentum for progress.

28.3.5.2 Optimizing Writing Conditions in Conditions of Writer's Block

Despite these trusted strategies to overcome writer's block, there are times in many writers' careers when getting started continues to be a challenge. In this situation, there are other options. One option is to reconsider the chosen writing environment. It is crucial to work in a space that is conducive to reflection and writing. Creating a work area that is comfortable and quiet is essential. These elements may contribute to times of concentration and productivity. The time dedicated to writing ought to be solely for that purpose, with as few distractions or unnecessary excursions as possible. If writing inertia does set in, it may be wise to take a short break. Allowing oneself time to get away from work at certain points can provide much needed sustenance and relaxation, which in turn will support increased productivity and efficiency later. Writing is a process that lends itself well to the stepped approach of small-goal accomplishment. Given the solitary nature of the work, it is important to create personal incentives when targets are met. Consider promising yourself a reward once certain milestones have been achieved.

Finally, it may be worthwhile to remember the wider community of scientific writers who will, no doubt, have experienced similar set-backs and challenges in their work. It is possible that valuable input offered from a fellow writer may instill a renewed energy for progress. Reaching out and making contact before paralysis ensues is essential.

28.4 The Content, Format, and Style of a Research Paper

Conventionally, a research paper is composed of a number of sections (Hall 2008). This ordered approach has its origins in the publication of the first journals less than 350 years ago (Day and Gastel 2008). It has led to the development in the last century of the IMRAD structure (Introduction, Methods, Results and Discussion) of original research papers (Hall 2008; Day and Gastel 2008). We discuss each section of the paper in terms of commonly stated guidelines for the content, format and style. But preceding the IMRAD sections, there is always a title and abstract section. Below we discuss these in the order in which they appear in a manuscript.

28.4.1 Title Section

When choosing the title of the manuscript a number of points are worth bearing in mind.

- The title must characterize the object and methods of the study in very broad terms and not announce any interpretation of the evidence produced (Miettinen 2011b: Up from Clinical Epidemiology & EBM, p.104)
- If the study was abstract scientific instead of particularistic (*See:* Chap. 1), it is preferable to avoid references to place and time in the title
- The general principle of scientific writing with regards to concepts and terminology applies to titles and the clearest and shortest way to convey the content is, in principle, the best
- The concepts and terms in the title should conform to the major concepts and terms in the body of the text
- Writers are often required to follow a traditional title style of the journal they submit to (Hall 2008)

Most writers produce the main body of the manuscript first and return to the title section as the last task (Toft and Jaeger 1998; Gregg 2008). A title page may need to comply with journal-specific demands for information about authors, affiliations, among others.

28.4.2 Abstract Section

This section is meant to be a stand-alone summary of the entire manuscript, provided to readers for quick orientation about the evidence of the study. As per instructions of the journal editors it is usually required to have 200–300 words at most. This is sometimes the only part of the paper that is freely accessible via electronic search engines, so attention to informativeness is imperative (Hall 2008). It should be considered an independent document that can be understood separately from the main manuscript. A clear and concise structure is advised. References and abbreviations are to be avoided in the abstract section (Neill 2007) (Table 28.2).

As to format, abstracts can be structured or unstructured, the difference being the addition of sub-headings in the structured format. The former is fast becoming the more common type (Hall 2008). Regardless of the format, certain information is nearly always expected by the journal editors: background information on the study rationale; the main aims and objectives; a concise description of the subjects enrolled and methods employed; outcome parameter estimates; and conclusion of the study (Baguma et al. 2010). The conclusions, if required, should ideally be restricted to a very brief repeat of the main evidence. Any interpretation of evidence cannot be considered essential anywhere in an original research paper (Miettinen 2011b) but is conventionally accepted and in some journals even required.

Table 28.2 Conventional editorial key-content requirements for an abstract section

Background	Domain of the study
	What is known about the research topic
Aims and objectives	The key research questions
	The target population
	Study area
Methodology	General study design
	Sampling and enrollment scheme
	Measurements and variables used
	Outcome parameters
Results	Synopsis of the main outcome parameter statistics

28.4.3 Introduction Section

One main purpose of this section is to outline the reasons for undertaking the study; i.e. to describe what new evidence was needed and why (Baguma et al. 2010; Hall 2008; Gregg 2008). This outline may start with a brief orientation about the general domain of interest (e.g., treatment of children with asthma) followed by some information about the state of knowledge (including gaps) in this domain, but relevant to the objective of the study (e.g., the efficacy and safety of a certain class of medications is well known but the relative efficacy of a newly marketed medication within this class remains to be investigated). Provide a thorough, succinct description of the background information available and the research currently accessible (Woolever 1980). Clearly show the dearth of epidemiological research (i.e., the knowledge gap) that provoked your study (Neill 2007). Also, provide the reader with a clear picture of what the rest of the paper will contain (Gregg 2008). Most journals require the introduction section to have a concise format, ideally not exceeding two double-spaced pages in length. No study results or discussion of results ought to be included.

The second main purpose of the introduction section is to formally describe the general objectives and the specific aims of the study (*See:* Chaps. 4 and 5).

28.4.4 Methods Section

Conventionally the methods section is required by journal editors to be approximately 3.5 double-spaced pages long and to have a clear, logical flow. The study protocol forms the basis for the methodology section of the paper, including any protocol amendments implemented during study implementation. The methods section provides a thorough breakdown of the study design and is conventionally required to contain information as listed in Table 28.3. Ideally, it should provide a template for a similar study to be reproduced in another setting (Neill 2007). The inherent benefits of writing this section before undertaking the study are apparent. However, it should also assure the reader that the study was conducted in a manner that satisfies all principles of ethics and validity.

Table 28.3 Conventional key-content requirements of a methods section

Research setting	Study setting; environmental and geographical references and descriptions, where appropriate
General study design	Type of study design, e.g. randomized control trial, observational follow-up study, case report, survey, case control study etc.
Inclusion and exclusion criteria	Target population description
	Planned inclusion and exclusion criteria and reasons behind them
Recruitment, sampling and enrolment	The approach to sampling, e.g. statistical versus non-statistical
	Who carried out both procedures; how and where they were undertaken
	Details relating to the enrolment period
Interventions	Types, intensity, duration of any planned interventions and how they differ among intervention groups
Measurements and variables	Measurements, measurement scales, measurement instruments, measurement sessions, timing in individual follow-up
	Who conducted the measurement procedures; training and supervision methods
	Outcome variables, determinants, modifiers, confounders
	For follow-up studies, mention the possible end-points of individual follow-up
Quality assurance including quality control	Methods to maximize validity and integrity of the data
	For trials: good clinical practice guidelines, standard operating procedures; adherence optimization and measurement plan
	Performance statistics planned; data quality expectations; data quality statistics planned
Data handling	Database and data entry system
	Database management procedures, including the data cleaning process
	Variables in the database
Sample size and/or power calculation	Motivation for chosen target size
	Anticipated refusal and dropout rates and missing value rates
	Anticipation of level of precision of estimates
Analysis plan	Statistical software package used
	The general analysis approach undertaken, i.e. estimation and/or testing, intention-to-treat
	The choice of statistical measures (e.g. causal rate ratios, prevalence ratios and confidence intervals) and summary statistics (e.g. chi-square and p-value)
	In the case of analytical studies state the methods used to control for confounding during analysis
Ethical aspects	Institutions and committees that provided ethical approval, oversight and support
	Relationships with important stakeholders
	Details about the informed consent process
	Dispositions for ensuring participant confidentiality and privacy
	Guidelines in place to ensure participant safety and the provision of medical care during the study

Table 28.4 Conventional practice guidelines for presenting data in tables and figures in journal articles

Tables and figures	The maximum number of tables and figures is usually 5
	Ensure that numbers and percentages add up correctly
	Table titles should be placed above the table
	Figure titles are placed below the figure
	Table rows represent categories, columns statistical results for those categories
Numerical data notation	Use a zero before a decimal point e.g. 0.5 not .5
	Present P-values correctly and with preferably two non-zero digits after the initial zeros e.g. P=0.0036
	Avoid unnecessary precision, e.g. use one decimal place for percentages e.g. 66.7 %, not 66.66 %
	Place a space between a number and its unit of measurement
	Spell out a number at the beginning of a sentence

28.4.5 Results Section

Most contemporary medical journals expect the results section of a submitted manuscript to be 1.5–2.5 double-spaced pages in length. The primary aim of this part of the manuscript is to succinctly and objectively describe study sample characteristics, protocol adherence information, and statistical evidence of the study (Neill 2007; Hall 2008). Generally, the format involves the use of tables, figures, and text containing statistical outcomes. It is important to maintain a logical sequence that directly addresses the aims and objectives posed in the introduction section. Much thought should be given to the presentation of the data, particularly if it is intricate and detailed. A moot point also pertains to the difference between data and results. The data refers to information, including statistics, based on the measurements undertaken. The results are the interpretation of this information (Hall 2008). This subtle difference becomes an important detail when writing this section. A common approach involves outlining the descriptive statistics of the study, thereby providing the reader with a basis from which to understand the analytical results presented later (Gregg 2008).

Consider the use of tables and figures for displaying the main findings (Table 28.4). These can then be interpreted in the text. Ordinarily, figures show trends in the data or differences in the distributions.

28.4.6 Discussion Section

This section of the manuscript should emphasise the most pertinent findings of the study by outlining the most important evidence clearly for the reader (Gregg 2008). Most journals expect a maximum of 2.5 double-spaced pages in the submitted manuscript and the following content.

Table 28.5 Discussion section content commonly required by journal editors

Principle result	Highlight the most significant result from the study
Appraisal of the study	Mention the strengths and weaknesses of the study
	Detail potential sources of bias, confounding and random error
Major challenges encountered	Outline any problems encountered during the implementations of the study
Interpretation of results	Evaluate the results of the study, allowing for clear interpretation by the reader
Recommendations	Describe the potential public health implications of the study results
	Explain any suggestions for improvements to potential future study
Conclusion	Summarize the main scientific results of the study
	Outline again the main recommendation(s)

One of the first paragraphs of the discussion should outline the primary results of the study, stating clearly and unambiguously what the direct interpretations of the main statistical results are. The next paragraph could refer to the main strengths and weaknesses of the study. An honest and objective approach here is most suitable and will lend merit and value to the overall impression of the study and the investigators (Lilleyman 1995; Gregg 2008). If any noteworthy challenges were encountered at any stage of the study it is important to draw attention to them again. The most favorable outcome would entail an honest outline of the limitations without placing too much emphasis on them. The same is true of the strengths of the study; it is important to mention strengths yet not focus on them excessively. It is wise to remain succinct and modest in one's interpretations. Reference can be made to the possible wider implications of the study results within the scientific community. It may be possible to foresee how the findings will augment the current knowledge-base on the topic. Furthermore, potential improvements that might add value to a future study can be mentioned. Writers should defer to the individual journal's instructions with regard to the inclusion of a paragraph dedicated to summarizing conclusions, as it is sometimes included and sometimes not (Baguma et al. 2010) (Table 28.5).

28.4.7 Acknowledgements

The journal instructions should be consulted for guidance about this section of the manuscript though it is customary to mention the source of funding for the project and any conflict of interests for the writing team (Hall 2008). It is also good practice to recognize the efforts of those who assisted with the project. *See:* Chap. 31 for a discussion of ethical aspects of this section.

28.4.8 References List

The references section of a manuscript requires strict adherence to the journal's specifications (Woolever 1980). It is convenient to have established the chosen format at the beginning of the writing process. The references are an on-going element of the work that should be rigorously managed throughout the writing process (Hall 2008). In essence, they serve as a scientific basis for the undertaking. Many writers file hard copies of their referenced papers while also taking advantage of reference management software, some of the most popular ones being EndNote (Reuters 2011) and Reference Manager (2012).

28.5 The Oral Scientific Presentation

Part of the remit of adding to the existing scientific knowledge-base involves oral presentations of the results to the target audience. While the skills necessary to successfully deliver the results of a study orally are quite different to those of writing the manuscript, the objective is the same: to impart new information to the audience in a clear, concise, and informative manner. In this section we offer some suggestions for young researchers who are unfamiliar with oral scientific presentations (e.g., at medical conferences).

28.5.1 Preparations for Oral Presentations

One of the first considerations involves knowing your audience (Gregg 2008). This will have a significant effect on the preparation and style of the presentation. Communicating the results to a lay audience will involve a very different approach and technique compared to delivering it to an audience of one's peers in the scientific conference format. In the case of delivery to a professional audience, it is wise to be aware of the research backgrounds and domains of the majority of those attending. Regardless of the composition of the audience, a large degree of flexibility may be required to adapt to the needs of those present.

Another similarity between the preparation of the presentation and writing the manuscript concerns the preparatory phase. It is likely that each will take significantly more time and re-drafting than is first anticipated. For this reason, it is wise to allow sufficient time to prepare the content, including time to rehearse the delivery (Thompson et al. 1987; Alley 1996). It is possible that older presentations could be altered and updated to merge with the new results being presented. However, if this is the case, it is vital to invigorate the presentation with a new and relevant approach; it is never worth tarnishing one's reputation in a misguided effort to save time. This is especially relevant in the case of collaboration. Co-authors should have the opportunity to review the presentation and provide feedback (Gregg 2008). Take note of possible questions that colleagues ask. Preparing carefully at this point may help to

Textbox 28.2 Potential Advantages of Using Slides During an Oral Presentation

The vast majority of scientific presentations today include the use of PowerPoint slides or the equivalent from another presentation software program. This method of delivery has become the norm today with many obvious advantages. The inclusion of relevant charts, tables, or graphs can enliven the talk and keep the audience engaged and interested (Gregg 2008). Indeed, the use of slides may even be necessary if the presentation relies heavily on data and statistics. After all, the purpose of the talk is to impart the findings to the audience in a clear and interesting way. Furthermore, the use of slides can provide a focal point and guide for both the presenter and the audience.

create a more confident and poised approach during the presentation and particularly at question time. At this point, it is worthwhile preparing for expected questions. It appears that certain questions are commonly asked. For example, audience members are often interested to discover if you have conducted a broad, inclusive analysis. They may also enquire if you have examined the data in terms of subgroup analysis or if you have considered other outcomes or exposures. As a result, it may be wise to prepare a response in advance.

When choosing to use slides for your presentation (Textbox 28.2), it is important to be discerning about the number to include. This will be determined by the time assigned to present. As a rough rule, one slide will take approximately 1 min to deliver. Time should also be allocated for questions from the audience.

When designing the slides, most presenters will adhere to the format of the scientific paper, i.e., Title, Background, Objectives, Methodology, Results, and Conclusions. It is wise to maintain a consistent typesetting style, i.e., retain the same bullets and numbering and sub-level fonts throughout. Another common approach followed in the design of the presentation involves the 'Rule of 7.' This refers to limiting the number of lines on each slide to 7 (excluding the title), with the aim of maintaining an orderly and easily understood presentation. When preparing the slides it is best to avoid the use of red or pale-colored fonts, as these may be difficult to read.

The process of being selected to present your work begins with producing an impressive and comprehensive abstract. It should follow a structured format that includes a clear hypothesis and new and complete results. Statistical estimation data should be presented using confidence intervals.

28.5.2 Avoiding Presentation Pitfalls

During the actual delivery of the presentation there are a number of pointers to be aware of in order to perform to your best standard. Clearly, a degree of uncertainty exists but it will pay dividends to be well prepared. In terms of the presentation software, ensure that you take the most up-to-date version to the venue. Being organized in this regard is essential. It is also worthwhile doing the following:

- To download the file containing your presentation
- To arrive at the venue ahead of schedule to become familiar with the room allocated for the presentation
- To introduce yourself to the chairperson of your session, as this may be an opportunity to develop a rapport with him or her, and it may offer a moment to discretely suggest an interesting question after your presentation
- To become familiar with the stage and podium, and to locate the microphone, pointer, USB plug, and screen
- To practice using the microphone if one is available
- To ensure that a glass of water will be available during the lecture, should you need it
- When possible, to have a practice session at this time and become comfortable on stage, taking note of the correct distance to maintain from the microphone in order to project your voice most successfully throughout the room

For the actual presentation, it is important to take time at the beginning to introduce yourself and the lecture. It is essential not to rush over the slides but to clearly explain the material presented. This can be a challenge for those inexperienced in presenting and will be greatly improved by performing many practice sessions of the lecture in advance. Be mindful of the time assigned to your presentation and be careful to adhere to this schedule during the practice sessions. Also remember that written material (handouts) can be used as an aid for delivery. Allow yourself to breathe and try to maintain as calm an approach as possible. After all, the focus of the presentation ought to be on the material being presented. It may allay nerves to imagine oneself merely as the conductor of the information to the audience. A steady and centered style should help to achieve this. It is worthwhile incorporating a joke occasionally if this is an approach that works for you. However, try to maintain a presence that is natural to you. Remember to smile. This will create an appearance that you are comfortable. On this point, remember to dress comfortably for the presentation; remain true to your own sense of style while dressing appropriately for the occasion (Gregg 2008; Alley 1996).

Throughout the lecture, remember to address the audience regularly, paying attention to the whole room (Gregg 2008). It may help to use a pointer to concentrate attention on specific detailed information. If you intend to include any controversial statements it is important to be in a strong position to defend them should you be challenged by the audience. On this point, it will be of benefit to attempt to sense the reactions of the audience during the talk and adjust the style and speed of delivery accordingly. It is invaluable to develop an affinity for your audience and try to remain aware of the atmosphere in the room. Once you have concluded the lecture, take time to thank your audience for their attention (Gregg 2008). Take a moment to compose yourself with a drink of water if necessary. Remain poised and centered for the questions that will follow. Before alighting from the podium, remember to take off the microphone.

28.5.3 Dealing with Questions

If you are a novice presenter and very uncomfortable with the idea of having to answer questions, consider extending the lecture by 1 or 2 min beyond the time limit in order to restrict the available time for questions. At this point, the preparation undertaken earlier will hopefully benefit you now. You should largely be in a position of composure, safe in the knowledge that you are totally familiar with the content of your lecture and may have already examined potential questions. If, however, you don't quite understand a question or don't know the answer at all, firstly ask for the question to be repeated. This allows you a moment to gather your thoughts. If you are still unclear about the answer, it is best to admit this and thank the audience member for their question (Gregg 2008). It is also important to commit to investigating the answer at a later date. Should it happen that the person persists in engaging in further discussion, suggest that you meet with them individually later and discuss the issue in more detail. If this fails to quell their enthusiasm, consider referring them to another member of your team.

In this chapter we discussed aspects of content and style of delivery in scientific papers and presentations. An aspect of content that we wish to expand upon further is reporting of data quality because it is crucial for the demonstration of internal validity. This topic is the subject of the next chapter.

References

Alley M (1996) The craft of scientific writing, 3rd edn. Springer Science+Business Media, New York, pp 1–273. ISBN 0387947663

Baguma S, Anandajayasekeram P, Puskur R (2010) Writing convincing research proposals and effective scientific reports. Part B: scientific writing. International Livestock Research Institute, Nairobi

Bhopal RS (2002) Concepts of epidemiology: integrating the ideas, theories, principles and methods of epidemiology. Oxford University Press, Oxford, pp 1–472. ISBN 0199543143

CONSORT statement. http://www.consort-statement.org/home. Accessed Sept 2012

Day R, Gastel B (2008) How to write and publish a scientific paper, 6th edn. Cambridge University Press, Cambridge, pp 1–72. ISBN 9780521671675

Dorland's illustrated medical dictionary (2011) 32nd edn. WB Saunders Company, Philadelphia, pp 1–2176. ISBN 9781416062578

Gregg MB (2008) Field epidemiology, 3rd edn. Oxford University Press, Oxford, pp 1–572. ISBN 9780195313802

Hall GM (2008) How to write a paper, 4th edn. Blackwell, Oxford, pp 1–122. ISBN 9781405167734

ICMJE (2011) Uniform requirements for manuscripts submitted to biomedical journals: ethical considerations in the conduct and reporting of research: authorship and contributorship. http://www.icmje.org/ethical_1author.html. Accessed Sept 2012

Lilleyman JS (1995) How to write a scientific paper–a rough guide to getting published. Arch Dis Child 72(3):268

Miettinen OS (2011a) Epidemiological research: terms and concepts. Springer, Dordrecht, pp 1–175. ISBN 9789400711709

Miettinen OS (2011b) Up from clinical epidemiology & EBM. Springer, Dordrecht, pp 1–175. ISBN 9789048195008

Moher D et al (1999) Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Lancet 354(9193):1896–1900

Neill US (2007) How to write a scientific masterpiece. J Clin Invest 117:3599–3602

Porta M, Greenland S, Last JM (2008) A dictionary of epidemiology. A handbook sponsored by the I.E.A., 5th edn. Oxford University Press, New York, pp 1–289. ISBN 9780195314496

Reference Manager (2012) www.refman.com/. Accessed Sept 2012

Reuters, ENDNOTE Thomson (2011) Endnote X4. Thomson Reuters, New York. http://www.endnote.com. Accessed Sept 2012

STROBE statement (2009) http://www.strobe-statement.org/. Accessed Sept 2012

Strunk W, White EB, Angell R (1999) The elements of style, 4th edn. Longman, New York, pp 1–105. ISBN 0205313426

Thompson WM et al (1987) Scientific presentation: what to do and what not to do. Invest Radiol 22(3):244–245

Toft CA, Jaeger RG (1998) Writing for scientific journals I: the manuscript. Herpetologica 54:42–54

Woolever P (1980) How to write a scientific paper. Bios 51:12–16

Reporting Data Quality

29

Jonathan R. Brestoff and Jan Van den Broeck

It is the mark of an instructed mind to rest assured with that degree of precision that the nature of the subject admits.

Aristotle

Abstract

This chapter offers practical advice for investigators on how to report the quality of their own data in scientific papers. The proposed guidelines are based on an analysis of the concept of aggregate data quality. We first clarify the multidimensional concept of aggregate data quality and then proceed by deriving principles and practical recommendations for reporting data quality. When describing data quality, one may need to consider study-specific and variable-specific factors that influence data quality requirements. In this chapter we argue that reporting on data quality should be more comprehensive than currently accepted practices. Among the array of useful data quality parameters, we selected digit preference and intra- and inter-observer reliability statistics for more in depth discussion. Finally, we discuss the quality of laboratory data, an issue that deserves separate reporting.

J.R. Brestoff, MPH. (✉)
Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck, M.D., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_29,

29.1 Aggregate Data Quality

'*Data quality*' is commonly understood to be a characteristic of a single data point, variable, or dataset used for analysis. It is a multidimensional concept. The multiple quality aspects include the data's accuracy, completeness, and ethics of collection, confidentiality, and assignment to particular measurements of particular observation units.

Data quality can also be considered at a further level of aggregation of the data. Data points and variables undergo processing, including summarization into a final statistical estimate or P value. These statistics are the most condensed and aggregated forms of the data. With '*aggregate data quality*' (Panel 29.1) we mean the quality of these final statistical estimates and P values. Aggregate data quality is also a multidimensional concept. At this level, the main quality aspects are unbiasedness, ethicality, and verifiability of all underlying data collection, handling and analysis, and precision.

Serious flaws in design or violations of ethical principles *before, during, or after data collection* tend to reduce aggregate data quality towards zero. Minor design weaknesses or occasional sub-optimal behavior may or may not affect data quality. For example, if the informed consent process was carried out a bit hastily with one or two out of a hundred subjects, one may consider this to be a regrettable violation of ethical principles but not necessarily a serious enough infraction to completely devalue the study's data quality. Indeed, data quality can never be perfect (DeMets 1997). All aspects of aggregate data must reach a level of quality that is *acceptable for the particular study*. Aggregate data quality cannot be better than the worst-scoring dimension in a given study or analysis.

The most efficient and often the only way to ascertain aggregate data quality is to assess protocol compliance and the quality of all the data-related processes that have led up to the final aggregate data. Appropriate study design is a pre-condition.

The aggregate quality of a study's data is deeply connected to the methods used to collect the data; a complex network of interactions among investigators, technicians, and subjects; and the numerous data management activities that lead to study analysis. Thus, the assessment and reporting of aggregate data quality heavily rests on assessments of data integrity and performance indicators during data collection, management, and analysis. At every step, the study is susceptible to factors that could potentially reduce the quality of the data, and it is therefore critical to monitor for indicators of poor data quality.

Panel 29.1 Selected Terms and Concepts Relevant to Reporting of Data Quality

Data quality A characteristic of (the set of) data points used for analysis that takes account of the data points' accuracy, completeness, ethicality of collection, confidentiality, and their assignment to particular measurements of particular observation units

(continued)

Panel 29.1 (continued)

Aggregate data quality A characteristic of statistical estimates and P values that takes account of their unbiasedness, ethicality of underlying data collection/handling, verifiability and precision

Data quality report Written account of which quality assurance and quality control activities were carried out as well as of values for data quality statistics

Data quality statistics Measures summarizing quantifiable aspects of data quality

Digit preference Tendency to record numerical values containing specific digits at an unexpectedly high frequency

Efficiency The reciprocal of the resources spent to achieve a defined goal

Inter-observer reliability statistics Data quality statistics expressing the degree to which the observers in a study, when replicating *each other's* measurements with an accurate instrument, tend to obtain values that are close to each other

Intra-observer reliability statistics Data quality statistics expressing the degree to which the observers in a study, when replicating *their own* measurements with an accurate instrument, tend to obtain values that are close to each other

Unbiasedness Characteristic of an empirical estimate of an outcome statistic, namely its closeness to the true population value

29.2 Principles of Reporting Data Quality

29.2.1 Specification of Data Quality Expectations

A first important principle is that one needs to report on the extent to which data quality expectations and requirements were met. These expectations are study-specific and variable-specific to an important extent but there are also some *general requirements* for quality of data on study variables, which will be discussed separately for outcome variables and for determinant variables and covariates.

For any given study, it is useful to specify in the study proposal and protocol one's expectations regarding data quality and the methods that will be used to assess it. These expectations may have been formulated as quantitative thresholds, e.g., "an error rate in variable *x*, as assessed by database-to-source document comparison, of less than 1 % will be deemed acceptable." Alternatively, prior expectations may be described qualitatively, e.g., "Technical Errors of Measurement and Kappa statistics of observers should approach those of their trainers and supervisors."

Other expectations may concern the number and gravity of protocol violations and estimations of the effects of misclassification and other sources of bias. Typically, prior expectations need to be adapted to the general study design and the type of study variables being considered.

29.2.1.1 Quality Requirements of Data on Outcomes

When a study aims to estimate *the size of an exposure effect* on a binary outcome event (rather than just the existence of an effect), specificity and sensitivity of outcome detection are important considerations. Perfect specificity of outcome detection is desired (no false positive cases) (Miettinen 1985) because any decreases in specificity will tend to invalidate effect estimates. Moreover, the sensitivity of outcome detection must be similar across levels of exposure and modifiers of interest because unequal sensitivity across these levels may lead to bias. If the sensitivity of detection across levels are *equal*, then the relative risk estimate will be *unbiased*. However, if the purpose of the study is not only to estimate a rate ratio or rate difference but also to estimate *absolute rates* of the outcome, then both specificity and sensitivity of outcome detection must be very high. In other words, there must be very few false positives and very few false negatives.

When the outcome is not a binary event but a continuous characteristic such as systolic blood pressure, the average bias in its measurement must be the same over levels of the exposure of interest in order to obtain valid mean differences. For example, the estimated mean difference in systolic blood pressure between exposed and unexposed remains unbiased if systolic blood pressure was measured with an average bias of −5 mmHg in both groups.

When the study aims only to detect the *existence* of an exposure effect (but not the size of it), the requirement is that any misclassification of a binary outcome or, alternatively, any average bias in a continuous outcome measurement must be similar over levels of the exposure and the modifiers. Improved measurement precision of a continuous outcome measurement increases the efficiency of detecting the exposure effect, but perfect measurement precision is not needed.

29.2.1.2 Quality Requirements of Determinants and Covariates

When a case–control approach is used, exposure histories need to be assessed with similar quality and in the same way among cases and controls. To ensure valid inference, confounders should be measured without error. Imprecision in continuous exposure measurements attenuates rate ratio estimates (regression dilution, *See:* Chap. 27). For corrections, the degree of imprecision should be documented.

The exact way of reporting quality of study variables depends on which of the above scenarios and requirements are applicable, but often there will be a need to:

- Provide evidence on misclassification or average bias of the outcome measurement overall and by levels of the determinant and modifiers
- Calculate reproducibility statistics for the exposure of interest (this is done separately for cases and controls in a case–control approach)
- Document high accuracy and precision of measurements of all variables included in the statistical model (including confounders)

29.2.2 Reporting of Data Quality in Research Papers

When reporting on data quality in a research paper, it is good to consider:

- Reporting the a priori quality expectations of key variables
- Describing minor design flaws (and adjustments made to counteract said flaw) and minor violations of ethical principles before data collection
- Describing the frequency and severity of deviances in the execution of data-related study processes (at all stages and regarding all quality assessments)
- Reporting oddities in the end results of data-related processes (at all stages)
- Reporting whether encountered problems were successfully amended
- Providing arguments about whether any remaining problems are compatible with what was considered to be an acceptable level of quality (defined a priori)

Unfortunately, in many scientific publications there is insufficient information regarding protocol compliance and data-related processes. This precludes readers from interpreting the quality of the data being reported in scientific papers. Moreover, the widespread practice of scarce reporting on data quality precludes any judgment as to whether the non-reporting was in any way influenced by a desire to hide the poor quality of data. A practice of comprehensive and honest reporting allows one to avoid violating ethical principles and safeguards against accusations of misconduct. Reputations have been ruined over data quality concerns. If none of these or other issues arose during the study (a rare scenario), then the authors should include a statement to that effect. It is worth noting that the most impressive research papers do not shy away from reporting data quality metrics and one may earn the reputation as an astute epidemiologist should data quality issues be flagged appropriately. Online publications nowadays may allow for the attachment of appendices that report data quality in greater detail.

29.3 Practical Advice for Describing Total Data Quality

When there are serious ethical or scientific flaws, the study will not usually come to the publication stage, unless in instances when investigators and reviewers/editors are unaware of the problem. Serious lack of data quality is thus never reported in research papers by the investigators themselves. Concerning early stages of a study (before data is collected), investigators must report what they see as minor quality problems, such as minor design flaws, slight validity problems of chosen measures, or indications of selection bias. The challenges in reporting are (1) to be exhaustive about those weaknesses and (2) to provide fair arguments in support of the (implied) view that their effect on overall data quality is indeed minor. The latter may involve arguments that appropriate and effective measures were taken to turn an expected medium-to-large negative effect on quality into a minor one, for example by adjustments during analysis.

Concerning data-related study stages, the challenge in reporting is to depict process quality at all stages, paying equal attention to ethical and scientific dimensions of quality. Maximal use can be made of performance statistics produced during quality

Table 29.1 Describing data quality of a well-designed study

Category	Data quality parameters
Execution of data-related **processes**	Omissions in quality assurance and quality control
	Attention paid to data quality by clinical monitors, data and safety monitoring board, and other oversight bodies
	Intra- and inter-observer reliability statistics (*See:* text)
	Percentages of re-measuring using the method of maximum allowable differences between independent replicate measurement values
	Qualitative evaluations of the execution of procedures (outcomes of supervision activities; instances of deviances)
	Query rates (frequency with which data managers come across errors or matters that need clarification)
	Minor instances of data manipulation or falsification that were corrected
	Changes made to initial data plans (with reasons)
	Number of a *posteriori* exclusions from analysis (with reasons)
	Results of validation studies
	Existence of data dictionary and audit trail
Oddities in the end results of data-related processes	Item non-response rates and estimations of their biasing potential
	Rates and types of unresolved or true outliers (*See:* Chap. 20)
	Terminal digit preference statistics (*See:* text)
	Preferences for questionnaire answers located at the top or bottom of the list
	Error rates in database-to-source document comparisons

control and data cleaning (*See:* Chaps. 11, 19, 20 and 21). Table 29. 1 is a checklist of reportable process features and metadata, not all of which will be available in all studies.

Data quality reports should focus as much as possible on the processes that concern important study variables. When the study has a long data collection period attention should be paid to temporal comparability of data and data quality (Cull et al. 1997).

Data quality can be monitored and assessed during or after data collection using internal and external validation studies (*See:* Textbox 29.1). Internal validation studies may focus on the completeness of data collection, the presence of data errors, the precision of laboratory tests, or other factors. External validation studies usually include the determination of sensitivity and specificity in comparison to a gold-standard or the comparison of descriptive statistics against a known reference population. When planning a study, it is recommended to incorporate validation tools where possible and perhaps even to include sub-studies aimed at addressing potential concerns about validity.

Investigators reporting on the quality of their own data should be mindful of the fact that objectivity may be endangered and that one may consciously or subconsciously

Textbox 29.1 Confidence Codes

When accessing a database, one should identify whether confidence codes exist for the values being studied. Some questionnaires or structured interviews permit the subject or interviewer to indicate their level of confidence in a value. If designing a new study, including confidence codes can greatly facilitate improving the overall data quality. There is no universally accepted confidence code system; however, the following system proposed by Bhagwat et al. (2009) is reasonable for most purposes. One might, for example, exclude from analysis any measurements with a confidence code of C or D.

Confidence code	Meaning
A or 4	The user can have considerable confidence in this value
B or 3	The user can have confidence in this value; however, some problems exist regarding the data on which the value is based
C or 2	The user can have less confidence in this value due to limited quantity and/or quality of data
D or 1	There are significant problems with this value related to limited quantity and/or quality of data

This table is a modified and reproduced, with permission of the authors, from Bhagwat et al., 2009

attempt to distort information on data quality in service of obtaining a perceived higher impact article (Cope and Allison 2010). One might also have other personal motivations, perhaps enhanced by external forces that interfere with objectivity and good data quality reporting practices. The opposite may be bias created by a drive to take on a crusader role. Maintaining an acceptable degree of objectivity during a research career occasionally requires examination of one's own motivations and their potential impact on objectivity. It is most useful to do this right before starting to self-report data quality.

29.4 Digit Preference and its Reporting

Digit preference – a common phenomenon in epidemiological research – is the tendency to record numerical values containing specific digits at an unexpectedly high frequency. Most commonly involved are terminal digits (also called 'last digits' or 'end-digits'), especially 0, 5, and even numbers (Altman 1991). Less commonly, there is an excess of terminal digits adjacent to 0 or 5 or to a combination of two or more numbers. Particular combinations of pre-terminal and terminal digits may also be preferentially reported, as is the case with reported birth weight (Edouard and Senthilselvan 1997).

Table 29.2 The four major processes leading to digit preference and examples of common measurements susceptible to these processes

Process	Common measurements susceptible to digit preference
Approximation	Height
	Circumferences (arm, head, waist, etc.)
	Skin-fold thickness
	Blood pressure
	Pulse
	Tuberculin skin test
Range preference	Diagnostic cut-offs (includes any measurement that the measurer could use to infer the diagnostic category for the subject)
Value preference	Any direct numerical measurement
Retrieval error	Dates
	Times
	Biological measurements

29.4.1 Processes Leading to Digit Preference

There are at least four common processes leading to digit preference (Table 29.2):

1. *Approximation: Approximate reading of instrument scales*
   The characteristics of the scale, especially its graduations and readability, may influence digit preference. The measurer may round to the nearest big graduation mark or use a similar strategy to approximate a measurement value rather than obtaining a careful measurement. This issue is more likely to be problematic if the measurement workload is so high that staff must rush in order to meet specified deadlines. Digital measurement scales tend to produce less digit preference than analog scales and may be a useful element employed to increase the chances that collected data will be of high quality (*See:* Panel 29.2 for additional strategies to reduce digit preference).
2. *Range preference: Conscious or unconscious preferences for or against certain ranges of measurement values on a numerical scale*
   This may lead to measurement bias in the immediate zone of a threshold value (e.g., whether or not the subject has hypertensive blood pressure) and tends to do so *in a specific direction*. The effect is to avoid or promote measurement values that reach the threshold level.
3. *Value preference: Conscious or unconscious preferences for or against certain specific values on a numerical scale*
   When the conscious or unconscious concern is not an undesired range/zone on the numerical scale but specific values only, this may translate into a simple trend to avoid specific values and to select preferentially an adjacent value. Conversely, specific values may seem preferable, leading to the avoidance of adjacent values. Ironically, measurers may deliberately choose to avoid zero end-digits out of a concern to avoid digit preference.
4. *Retrieval error: Approximation of a numerical value when the exact value cannot be remembered or would be cumbersome to retrieve*
   In this scenario, the measurer tends to select a value at the mid-point between major scale marks.

Panel 29.2 What Can Be Done to Prevent or Amend Digit Preference?

- Selection of attentive and motivated observers
- Choice of measurement instruments with digital display
- Intensive training of observers
- Measurement standardization (*See:* Chap. 10)
- Intensive quality control (*See:* Chap. 11)
- If digit preference is detected in a study, then adjustments may be possible to partially correct the problem during data analysis (Eilers and Borgdorff 2004)

29.4.2 Digit Preference and Data Quality

Digit preference can cause bias and imprecision, both of which adversely affect data quality. Digit preference has led to erroneous estimations of prevalence, determinants of disease, and outcomes, as was shown by Eilers and Borgdorff (2004); Burnier and Gasser (2008); and Hessel (1986). Due to the relatedness of digit preference and observer reliability, digit preference is sometimes reported in descriptions of data quality. Some researchers have suggested using digit preference as a measure of quality control (Li and Wang 2005) or a marker of observer reliability.

29.4.2.1 The Relationship Between Digit Preference and Observer Reliability

Digit preference often involves settling for a less than optimal accuracy (Tourangeau et al. 2000), a problem that tends to occur when the accuracy of recorded measurement values requires mental effort by the observer (e.g., reading from an analog scale) or by the measured subject (e.g., answering a difficult question). The mental effort required to perform a measurement sometimes depends on the subject's willingness to collaborate during the measurement process. With 'difficult subjects,' such as an infant who is fearful of the measurer, more effort is needed to get an accurate reading, and digit preference may become more likely. An experienced observer should be able to minimize the impact of difficult subjects on the measurement quality; however, even the most experienced, disciplined observers are liable to digit preference if fatigued, over-worked, and pressured for time, inattentive, or distracted. These factors are known to contribute to observer reliability in general. Digit preference tends to be more common in secondary than in primary data.

29.4.3 Analysis of Digit Preference

Digit preference can be analyzed using a Chi-square goodness of fit test (Snedecor and Cochran 1980). In its simplest form, a Chi-square test has one Degree of Freedom (*See:* Table 29.3), but if a researcher is testing for terminal digit preference for each integer from 0 to 9, there will of course be nine Degrees of Freedom.

Table 29.3 Example of a cross-tabulation for a simple analysis of terminal digit preference (degrees of freedom = 1)

Terminal digit:	Expected (%)	Observed (%)
0 or 5	20	__
Other digits (1, 2, 3, 4, 6, 7, 8 0r 9)	80	__

The level of significance is typically set at 5 %, although one should always anticipate whether multiple testing calls for lowering the level of significance. Care should be taken in determining the expected terminal digit frequencies (Crawford et al. 2002).

A significant Chi-square test would raise suspicion of imprecise measurements and poor observer reproducibility, but concluding one way or the other requires confirmation from other performance statistics, such as Technical Errors of Measurement of the observers involved. Although a non-significant Chi-square test supports a conclusion that there is an apparent absence of digit preference, it is uninformative about observer reproducibility (*See:* Sect. 29.5, Intra- and Inter-observer Reliability Statistics).

29.4.4 Reporting Digit Preference

A study protocol may explicitly state a priori that observers measuring on a numerical scale will be trained to a level such that no digit preference is apparent. Another explicit expectation may be that implementing quality control protocols will minimize digit preference during data collection. These expectations should and digit preference analyses should be reported along with an interpretation about the successes of these strategies for individual observers (e.g., digit preference was not apparent in 8 of 9 observers) and for the entire database, with all observers pooled. In reporting the analyses of digit preference, one should report the observed and expected frequencies, the P-value, and the Chi-square statistic. The latter can serve as a rough measure of the degree of digit preference. One should further report about possible biases resulting from apparent digit preference and how this issue was dealt with during data analysis (Edouard and Senthilselvan 1997; Eilers and Borgdorff 2004).

29.5 Intra- and Inter-observer Reliability Statistics as Measures of Data Quality

In reporting on data quality, frequent use is made of intra- and inter-observer reliability statistics. *Intra-observer reliability* is high when an individual observer is able to reproduce measurement values of the same object over time (e.g., an observer's three independent measurements of a subject's waist circumference are nearly identical). *Inter-observer reliability* is high when there is agreement among multiple observers at any given time (e.g., five observers independently measure the waist circumference of one subject and obtain similar results). For both intra- and

Table 29.4 Kappa statistics and the strength of agreement, as proposed by Landis and Koch (1977)

Kappa statistic	Strength of agreement
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Very strong

inter-observer reliability, it is assumed that the same measurement technique is used for each independent measurement and by each observer.

Two of the simplest and most commonly used statistical estimates of intra- and inter-observer reliability are:

- The kappa statistic (κ) for categorical variables, such as whether or not a subject is hypertensive; and
- The intra- or inter-class correlation coefficient (ICC) for continuous variables, such as weight or height.

Many alternatives to κ and ICC exist; however, space constraints preclude their inclusion here.

29.5.1 The Kappa Coefficient (κ)

For comparing the reliability of two observers or two measurements performed by one observer, the kappa coefficient is calculated using the simple equation described in Chap. 11. If the categorical variable is not binary but ordered, then partial κ can be calculated using the method developed by Cohen (1968). If there are more than two observers or more than two measurements by one observer, then the Fleiss κ can be used instead of the equation found in Chap. 11 (Fleiss 1971).

As discussed in Chap. 11, perfect agreement or no agreement (beyond what is expected purely by chance) is indicated if $\kappa = 1$ or $\kappa \leq 0$, respectively. Although there is no universally accepted rule, Landis and Koch (1977) suggested the following interpretation (Table 29.4):

Many statistical packages will report P-values for κ; however, in the context of testing for observer reliability, these P-values should be ignored because the null hypothesis of ‘no agreement’ is nonsensical (Kirkwood and Sterne 2003). If a valid measurement technique is used in a study, then the possibility of ‘no agreement’ within or among observers is, by definition, illogical; there must be some level of agreement.

Although κ may provide a useful indication of intra- and inter-observer reliability, this statistic must be evaluated carefully. It is possible to achieve very desirable values for κ but have systematic misclassification that may lead to substantial bias in a study. Even if a study achieves a value for κ of 0.9, it is best practice to produce contingency tables to verify that the proportion of disagreement is within an acceptable range, to determine whether disagreement is driven by a particular observer, and to evaluate whether cases of disagreement disproportionately represent one or more comparison groups (a sign of bias).

29.5.2 Intra-class Correlation Coefficient (ICC)

To estimate the intra- or inter-observer reliability of a continuous variable, one may calculate an ICC using the following equation:

$$ICC = \frac{\sigma^2_{Actual}}{\sigma^2_{Actual} + \sigma^2_{Error}}$$

where σ^2_{Actual} is the variance of the true (underlying) values as estimated from the sample and where σ^2_{Error} is the variance of the measurement error. The ICC can be any number from 0 to 1. If ICC = 1, then the variance of measurement error is zero, implying that the measurement is perfectly precise. This scenario is often referred to as *complete reliability*, although it is important to note that perfect precision provides no information about accuracy. If ICC = 0, then the variance of the true value is zero (i.e., the most unlikely scenario in which all subjects have exactly the same value) and the observed variance is due purely to measurement error. Since the value of an ICC depends on the actual or true variance, ICC is a relative indicator of intra- and inter-observer reliability only when the assumption of equal true variance is met. The ICC can be calculated using the above equation or derived from analysis of variance (ANOVA) or random effects models.

Alternatives to the ICC are the concordance correlation coefficient (CCC) described by Lin (1989) and Cronbach's α (described in Bland and Altman 1997).

29.6 Reporting on the Quality of Laboratory Data

Data collected by laboratory tests can be very expensive, and there are multiple opportunities during the analysis procedure that can lead to uncertainty in the data and reduced data quality. To ensure the high quality of laboratory data, the first step one should take is to calibrate the equipment being used. The calibration protocol and standards should be identical for all laboratory sites for a given study. It may be necessary to obtain technical support from the manufacturer, although usually this step can be avoided if the calibration curves from each machine are strongly correlated (Pearson $r \geq 0.98$) and if the calibration curves at different sites are themselves very similar. The second step involves the optimization of the laboratory protocol (if not already done as part of an internal validation study) and perhaps the establishment of a quality control checklist for optimized protocols.

29.6.1 Standard Curves

In many cases, laboratory techniques are amenable to the inclusion of standard curves that describe a laboratory value across a broad range of known concentrations of the factor being analyzed. For example, if serum insulin levels are being analyzed by an enzyme-linked immunosorbent assay (ELISA), known concentrations of insulin (i.e., the standards) are run alongside the specimens (i.e., the unknowns). The known concentrations are plotted against their respective ELISA values to generate a function that describes the insulin concentration for a given ELISA value. Using this function, the insulin concentrations of the unknowns can be determined. Standard curves are usually plotted on a linear scale and should ideally have a correlation coefficient of $r \geq 0.98$, although the degree of reliability required in a study may cause one to adjust the acceptable level of correlation. Since unknown values are based on the standard curve, it is best practice to build the standard curve on measurements made at least in triplicate and to repeat routinely the standard curve (and calibration curve) to ensure its consistency over time. In addition, each independent analysis usually must have its own standard curve to control for possible variations between one analysis and another and to improve comparability of analyses.

Standard curves are useful not only for determining unknown values in a given sample but also for identifying suspicious values that should be flagged for repeat analysis in an independent analysis. When planning a study, it is recommended to build into one's budget room for repeat testing of a reasonable number of samples. The number of samples that require unplanned repeat testing should be reported. If planned repeat testing is performed on a random selection of the collected specimens, then one should report the coefficient of variation for repeat measurements (see below) and, if possible, intra- and inter-observer statistics. If multiple laboratories are performing measurements for a study, then the different labs should compare their results for the same specimens.

29.6.2 Coefficient of Variation (CV)

In many laboratory measurements, the variation of a measurement is related to the mean value obtained. For example, if a non-diabetic subject's blood glucose levels are measured while being fasted (usually for 12–16 h) and again 60 min after consuming a 75 g bolus of glucose, that subject will have a low value for the fasted measurement and a high value for the second measurement. If blood glucose concentrations are measured in triplicate for both sampling times, one can calculate the mean and standard deviation for each. The standard deviation of the fasted measurement will likely be lower than the standard deviation of the fed measurement. The coefficient of variation (CV) is calculated by expressing the standard deviation as a percentage of the mean. Although the means and standard deviations differ in the example above, the CV should be similar for the two sampling times. Similarity of the CV across a range of measurement values signifies that the laboratory measurement technique is reliable. A high CV for some or all measurements on a scale may signify bias and raise suspicion of data quality issues that require further attention.

In this chapter we discussed scientific reporting on data quality and argued that it should be more comprehensive than in currently accepted practice. In the next chapter (Chap. 30: Dissemination to Stakeholders), we move the emphasis from scientific reporting to the challenging task of dissemination of research findings to other stakeholders, such as funding bodies and the public.

References

Altman DG (1991) Practical statistics in medical research. Chapman and Hall, London, pp 1–611. ISBN 0412276305

Bhagwat SA, Patterson KY, Holden JM (2009) Validation study of the USDA's data quality evaluation system. J Food Compos Anal 22:366–372

Bland JM, Altman DG (1997) Statistics notes: Cronbach's alpha. BMJ 314:572

Burnier M, Gasser UE (2008) End-digit preference in general practice: a comparison of the conventional auscultatory and electronic oscillometric methods. Blood Press 17:104–109

Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70:213–220

Cope MB, Allison DB (2010) White hat bias: a threat to the integrity of scientific reporting. Acta Paediatr 99:1615–1617

Crawford SL, Johannes CB, Stellato RK (2002) Assessment of digit preference in self-reported year at menopause: choice of an appropriate reference distribution. Am J Epidemiol 156:676–683

Cull CA et al (1997) Approach to maintaining comparability of biochemical data during long-term clinical trials. Clin Chem 43:1913–1918

DeMets DL (1997) Distinctions between fraud, bias, errors, misunderstanding, and incompetence. Contr Clin Trials 18:637–650

Edouard L, Senthilselvan A (1997) Observer error and birthweight: digit preference in recording. Public Health 111:77–79

Eilers PHC, Borgdorff MW (2004) Modeling and correction of digit preference in tuberculin surveys. Int J Tuberc Lung Dis 8:232–239

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76:378–382

Hessel PA (1986) Terminal digit preference in blood pressure measurements: effects on epidemiological associations. Int J Epidemiol 15(1):122–125

Kirkwood BR, Sterne JAC (2003) Essential medical statistics, 2nd edn. Blackwell, Malden, pp 1–501. ISBN 9780865428713

Landis JR, Koch GC (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Li L-J, Wang P-S (2005) Should we use digit preference as an indicator of quality control? Med Hypotheses 65:192–204

Lin LIK (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255–268

Miettinen OS (1985) Theoretical epidemiology. Delmar, New York, pp 1–359. ISBN 0827343132

Snedecor GW, Cochran WG (1980) Statistical methods, 7th edn. The Iowa State University Press, Ames, pp 1–507. ISBN 0813815606

Tourangeau R, Rips LJ, Rasinski K (2000) The psychology of survey response. Cambridge University Press, Cambridge, pp 1–401. ISBN 9780521576291

Dissemination to Stakeholders

30

Jonathan R. Brestoff, Jan Van den Broeck,
Michael C. Hoaglin, and Nora Becker

> *Too often, the products of research are not disseminated or translated into community settings where the information is likely to have positive effects.*
>
> R.C. Brownson, et al.

Abstract

Dissemination of scientific work to others is an essential component of the research process. The most common form of dissemination is to publish articles in academic journals (*See:* Chaps. 28 and 31); however, a researcher may wish to disseminate work directly to the public, policymakers, or other non-academic stakeholders to achieve desirablwe effects on public health and to enhance the profile of the research team. A general discussion around engaging with stakeholders is found in Chap. 8. The current chapter extends our discussion of engaging with stakeholders from the perspective of disseminating scientific work in forms other than academic journal articles. First, we will introduce principles of dissemination and diffusion of information. We will then provide practical advice on developing dissemination strategies and on communicating with selected types of stakeholders, such as news media reporters. Finally, we discuss some ethical aspects of influencing public health policy and summarize practical advice in this respect.

J.R. Brestoff, MPH (✉) • M.C. Hoaglin • N. Becker
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck, M.D., Ph.D.
Faculty of Medicine and Dentistry, Centre for International Health,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_30,

30.1 Theories and Principles of the Dissemination and Diffusion

Information is transmitted within any given population through two general phenomena: *dissemination* and *diffusion*. The former refers to any purposeful action in which new information is introduced to members of that population, whereas the latter refers to the non-deliberate process by which members of a population accept or reject disseminated information (Buller 2006). In this section, diffusion is described first as a foundation for understanding dissemination, and in this context the social marketing approach is borrowed and modified as a suggested strategy for maximally efficient dissemination of scientific work (Dearing et al. 2006). Following this discussion, we describe practical approaches to enhance the likelihood of successful dissemination.

30.1.1 Diffusion of Innovations Theory and the Social Marketing Approach

Perhaps the most important advance in the understanding of human behavior change on a societal level – as opposed to behavioral change in an individual – occurred in 1962 when Everett Rogers published his groundbreaking book *Diffusion of Innovations*. This theory of social change is presented only in a brief summary format here, but those interested in this topic are referred to Rogers' book (Rogers 2003) and to Malcolm Gladwell's *The Tipping Point* for formal and popularized discussions, respectively. Terms and concepts relevant to this chapter are defined in Panel 30.1.

Panel 30.1 Selected Terms and Concepts Relevant to Dissemination to Stakeholders

Diffusion the non-deliberate process by which members of a population accept or reject information

Dissemination any purposeful action in which new information is introduced to members of that population

Innovation any novel development, such as information, ideas, behaviors, or rules

Social marketing a dissemination practice in which opinion leaders in a society are identified and leveraged as the points of dissemination

Social network a social structure defined by members and the dyadic ties between those members (a dyad is a two-person relationship, the smallest possible social group)

SOCO (Single Overriding Communication Objective) the single most important message to deliver about an innovation

Fundamental to the Diffusion of Innovations theory is the idea that society has inherent albeit dynamic channels of communication through which a message is transmitted from one entity in society to another. When an individual, group, or organization adopts an innovation – which may take the form of a new idea, behavior, practice, or object – others in society are directly or indirectly exposed to that innovation via one or more communication channels. These individuals may or may not be persuaded to adopt the innovation. The decision of whether to adopt an innovation can be optional, like trying the latest diet; collective, such as a professional association's approval of clinical guidelines by consensus; or authority-driven, such as laws that ban dumping waste in public water supplies. Early adopters interface with those who have remained undecided about whether to adopt the innovation themselves, and some will perpetuate this cycle of adoption in what's known as the multiplier or snowball effect (Fig. 30.1). After a party decides to adopt an innovation, the individual makes further decisions about whether to continue or discontinue use of the innovation and how. Ultimately, the steps in diffusion depend on social, ecologic, and systematic factors in society; therefore, dissemination strategies will differ at the individual, organizational, and network levels (Estabrooks and Glasgow 2006; *See:* RE-AIM Framework section).

Insight into diffusion dynamics allows for an understanding of how new information, once introduced to society, succeeds or fails to spread amongst its constituents. Social marketing is the practice in which opinion leaders – who may be socially influential people, peer-educators, celebrities, authoritative organizations, etc. – are identified and made the point of dissemination. These opinion leaders then leverage the diffusion process by influencing many individuals, some of whom will be particularly socially influential among peers, to adopt an innovation (Fig. 30.1). Dearing et al. (2006) suggests that social marketing principles can be employed for the implementation of campaigns to increase physical activity, the implication of which is that epidemiologists in general can use a social marketing-like approach to disseminate research findings. In this context, the epidemiologist's innovation is typically new information about the benefits or harms of a particular health-related behavior or phenomenon.

30.2 Practical Considerations for Dissemination Strategies

It is important to consider the dissemination strategy in detail because the entity that or person who disseminates the innovation may influence the success of diffusion. The dissemination of the Behavioral Risk Factor Surveillance System (BRFSS), a low-cost health survey system originally developed in the United States in the 1980's, is an example of a well-designed, well-controlled dissemination process (Bauman et al. 2006). The BRFSS was launched in 15 states in 1984 and, over the course of a decade, was adopted by the other 35 states and the District of Colombia; by 1998, some US territories in the Caribbean also adopted the BRFSS (Nelson et al. 1998). After successful implementation in the United States, other countries – including Mexico, China, and Russia – made culturally appropriate revisions before deploying the system themselves (McQueen and Puska 2003). However, one

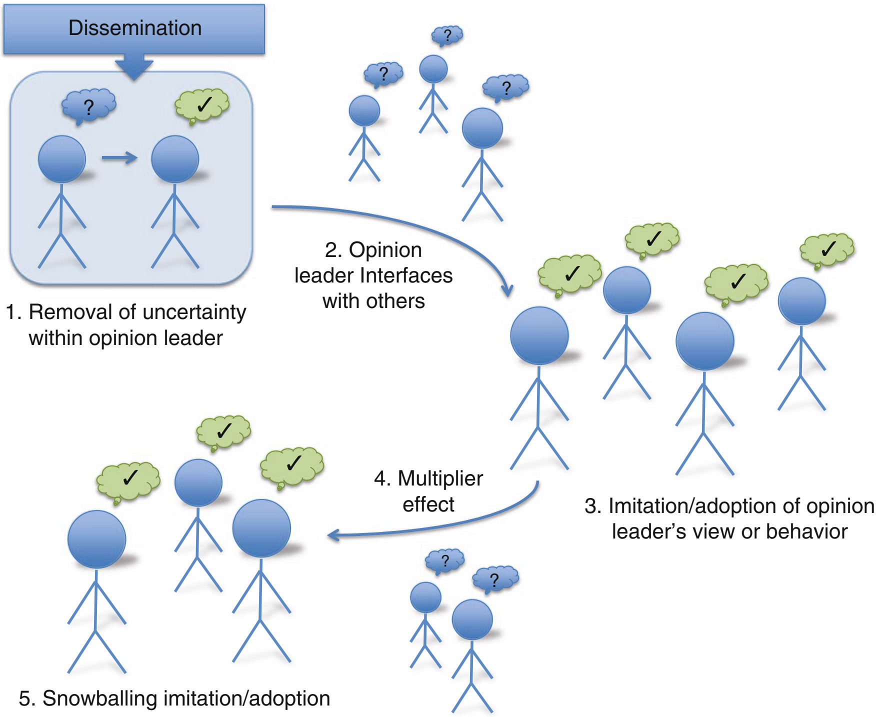

Dissemination
?
✓
1. Removal of uncertainty within opinion leader
?
?
?
2. Opinion leader Interfaces with others
✓
✓
✓
✓
3. Imitation/adoption of opinion leader's view or behavior
4. Multiplier effect
✓
✓
✓
?
?
5. Snowballing imitation/adoption

Fig. 30.1 *Dissemination and diffusion of information.* Dissemination refers to any purposeful action intended to introduce new information to society. The dissemination process may be highly variable and is represented by the *blue square* (*Step 1*). All individuals in society are inherently uncertain regarding whether to adopt never-before encountered information, a state of mind represented by a *blue cloud* with a question mark. Exposure to new information prompts a decision of whether or not to adopt that information (represented by the conversion of a *blue* with a question mark to a *green cloud* with a check mark), a process that depends on the removal of uncertainty (e.g., by logic or incentive). Among the early adopters are "opinion leaders," who have the capability of interfacing with (*Step 2*) and influencing others in society to imitate the opinion leader or adopt the information (*Step 3*). Some of these individuals will be opinion leaders themselves and propagate the pattern just described, a scenario known as the *multiplier effect* (*Step 4*). A consequence of the multiplier effect is snowballing imitation and adoption (*Step 5*), a phase of diffusion that represents the greatest change in the number of imitators and adopters. Note that, although this figure depicts the opinion leader and others as people, this representation is not intended to be exclusive of other entities in society, such as organizations

should be aware that dissemination and diffusion are highly variable processes and may or may not go according to plan.

The experience of the International Physical Activity Questionnaires (IPAQ), perhaps the most widely used survey tools for the surveillance of physical activity, shows us that dissemination processes are not always so structured, even when dissemination is successful. The IPAQ investigators had planned further research and development of IPAQ before promoting widespread use of the tool; however,

Table 30.1 Examples of dissemination-based intervention strategies reported for five cancer-related topics between 1980 and 2004[a]

Level of intervention	Strategy
Healthcare provider	Physician/nurse training
	Office systems (e.g., prompts and reminders)
	Audit and feedback
Individual	Reminders and invitations/postal delivery
	Telephone counselling
	Healthcare advice
	Self-help or patient education
	Financial incentives and competitions
	Counselling
	Role modelling
	Peer educators
Enhanced access	Removal of financial barriers
	Removal of access barriers
	Media-based education campaigns
	Policy-level interventions
Multi-component	One each from provider- and individual-level
	Treatment algorithms/clinical guidelines
Groups/organizations	Recruitment of professional organizations
	Workshops/conferences
	Peer educators
	Influencing a social network
	Radio broadcasts

[a]This table is derived and modified from Ellis et al. (2005) in accordance with American Psychological Association *Permission Policies*

the need for the instrument was so strong that research groups throughout the world adopted IPAQ rapidly and early (a more detailed history is found in Bauman et al. 2006). The IPAQ story suggests that some innovations are so needed or desired that the simplest form of dissemination (e.g., publication of a single journal article) is sufficient to trigger widespread diffusion, but such instances are rare and cannot be expected.

30.2.1 Six-Step Model to Enhance Dissemination of Information

Although there is not just one approach to planning a dissemination strategy, Bauman et al. (2006) proposed a six-step model to enhance dissemination. This model, shown in italics below, is annotated here with questions that may be useful to consider when planning a dissemination strategy. Some of these questions are derived from the RE-AIM Planning Tool (discussed below), whereas others are suggested by the authors. Table 30.1 also provides a framework for thinking about possible dissemination strategies. Collectively, the six-step model, RE-AIM Framework, and Table 30.1 should be helpful in selecting the processes that balance ideality, practicality, and cultural appropriateness.

30.2.1.1 Step-1: Describe the Innovation and its Rationale, Evidence Base, and International Contexts

- In full detail and also in just 1–2 sentences, what is the innovation?
- Why is the innovation important?
- What evidence suggests that this innovation will succeed?
- Under what sociodemographic contexts were efficacy and effectiveness studied, and how do these compare to those of target audiences?

For dissemination, having information about both **efficacy and effectiveness of the dissemination strategy** is ideal, indeed sometimes necessary, but information on effectiveness is usually unavailable. Lack of effectiveness data is especially common when attempting to disseminate innovations to international or underserved populations; useful advice for these situations is available in Cuijpers et al. 2005 and Yancey et al. 2006, respectively.

30.2.1.2 Step-2: Characterize the Dissemination Strategy

- Identify the target audience for dissemination and its size.
  - Who should adopt the innovation and why?
  - Who/what are the relevant opinion leaders?
  - Do you hope to reach all members of the target population?
  - What is the sociodemographic breakdown of the target population?
- Anticipate and describe the sequence, timing, and format of the dissemination strategy. (This may require simultaneous or prior consideration of Step 3.)
  - How might characteristics of relevant opinion leaders influence dissemination?
  - How might characteristics of the target population influence the dissemination strategy?

30.2.1.3 Step-3: Define the Current Communication Channels Through Which Diffusion Might Take Place

These channels may occur within or across multiple levels of society:

- Who do the relevant opinion leaders interact or affiliate with?
- How does the target audience perceive the identified opinion leaders?
- Are any relevant changes in the communication channels foreseen?

30.2.1.4 Step-4: Determine the Role of Decision-Makers and Partners that will be Necessary for Dissemination at Various Levels of Society (e.g., Local, National, International, etc.)

- Which societal sectors, industries, organizations, or government bodies are particularly important for the planned dissemination strategy?
- What are the political structures in the relevant areas?
- Who in your social network has direct or indirect connections with identified decision-makers and partners?

30.2.1.5 Step-5: Identify Factors that Might Impair or Facilitate the Dissemination Strategy

- Which cross-cultural or political factors might influence the route of dissemination in different locations and how?
- What barriers might limit your reach to the target population, and how do you plan to overcome them?

30.2.1.6 Step-6: Create an *a priori* Plan for Evaluating the Dissemination Process (May Not Be Necessary Depending on the Goals of Dissemination)

- How will dissemination and subsequent diffusion be assessed?
- What resources will be required to conduct evaluations?
- Are the stakeholders in agreement with the evaluation plan?

30.2.2 The RE-AIM Dissemination Framework

The RE-AIM (Reach, Effectiveness, Adoption, Implementation, Maintenance) framework was proposed by Glasgow et al. (1999) as a practical tool to enhance dissemination. Although championed by the National Cancer Institute (NCI) of the U.S. National Institute's of Health (NIH), RE-AIM can be used to plan dissemination strategies targeting a range of health issues, not just cancer. Various helpful RE-AIM resources can be downloaded from the NCI's Cancer Control website (*See:* Sect. 30.6). Perhaps most notable are the RE-AIM Planning Tool and the RE-AIM Checklist for Study or Intervention Planning, both of which we recommend. To exemplify the utility of the RE-AIM framework, complex dissemination strategies are presented in a generalized form at three levels of society – an individual, an organization/specific location, and a network/population – in Table 30.2.

30.3 Communicating with the Public and News Media

In order to get the attention of a target constituency – those whom the researcher intends to influence, such as the public, policymakers, or other stakeholders – one must inform the members of that constituency about the innovation. Various routes are available, of which we explore two especially prominent ones: conventional media interviews and the emerging venues provided by online social networks. Reaching out to journalists, media broadcasters (e.g., radio and television), and media relations specialists is uniquely challenging and requires summarizing findings in ways that reflect the current state of the science, not just an isolated conclusion. Advice on interacting with journalists and preparing for interviews is therefore provided below. Much of this information is readily translatable to communicating with the public via social networks, a venue that has enabled companies to interact with and learn about their customers and that should, by extension, be useful for epidemiologists as well.

Table 30.2 Examples of general dissemination-related interventions aimed at various levels of society using the RE-AIM framework[a]

RE-AIM dimension	Individual level	Single delivery site	Systems or network level
Reach the target population	Compare characteristics of participants to non-participants or population	Information mailed, in exam room, waiting rooms	Incentives for screening
	Use multiple channels	Use all staff to recommend	Pair advice with usual contacts
Effectiveness or efficacy	Tailor to individual	Report at each contact	Public feedback
	Stepped care	Proactive planning	Adopt goals as part of organizational mission
	Outreach components	Record goals	Continuous quality improvement
	Evolve over time	Elicit patient barriers and concerns	Collaborate/network
Adoption by target settings or institutions	Elicit consumer demand	Flexible program that can be customized	Make top priority for system
	Publicize status of various organizations	Optional components or modules	Release or remove other competing demands
		Present data on cost effectiveness	Trial adoption
Implementation and consistency of delivery of intervention	Publicize consumer guidelines	Feedback on implementation	Public recognition, reinforcement, and rewards for staff
		Involve many staff	Fine tune, experiment, revise
		Prompt staff behaviours	
		Automate for consistency	
Maintenance of intervention effects in individuals and settings over time	Take-home messages	Ensure follow-up	Automate and institutionalize reports
	Publicize importance of re-screening	Outreach for those not involved	Publicize results
	Schedule follow-up	Feedback on re-screening rates	Commit to national or state organizations
	Provide advocacy opportunities for patients		

[a] This table's contents are in the public domain, but they have nonetheless been modified and reproduced with the permission of its creator, Russell Glasgow, National Cancer Institute

Panel 30.2 Routes of Communication with the Wider Community

- Participate in press conferences and submit press releases
- Publish a newspaper or magazine article or assist a journalist to do so
- Contribute to a television or radio program
- Produce and distribute a DVD or information brochure
- Engage the study's Community Advisory Board into dissemination
- Engage patient advocacy groups or health care professionals (e.g., community health workers, peer-educators, counselors) in dissemination
- Give information on and promote a study website or webpage
- Contribute to other web sources for health information (e.g., blogs)
- Organize public dissemination meetings
- Broadcast information via the radio or other media outlets

30.3.1 Routes of Communication and Getting a Message Across

Communicating scientific health information to the public is especially difficult, as most members will not be well-versed in the subject at hand. When communicating with the public, a researcher can select one or several routes, the best of which will depend on the researcher's goals of dissemination and characteristics of the target audience. Usually, there is at least one intermediary (e.g., a journalist) but direct communication is increasingly possible with technological innovation (e.g., through blogs and social networking websites). Panel 30.2 lists several common routes of communication, but this list should not be considered all-inclusive, nor should discussion of selected routes be interpreted as an indication of their importance.

How one delivers a message can influence whether or not the public adopts that message (most often through changes in understanding and behavior). Since that message is often relayed through an intermediary, the importance of proper delivery is heightened. Central tenets of effective communication are elucidated below through a discussion of preparing for an interview with a reporter, one of the most common intermediaries that epidemiologists encounter. These tenets may be extrapolated to other scenarios in which a researcher might need to deliver a message to a lay audience.

30.3.2 Preparing for an Interview with a Journalist

Perhaps the most critical exercise in preparing for an interview with the press is to identify the Single Overriding Communication Objective (SOCO), i.e., the single most important message to deliver about a study or health issue (Dan 2008; *See also* Panel 30.3). Typically, a research project will produce several interesting findings, and although discussing them all is tempting, doing so is almost always counterproductive.

Panel 30.3 Suggested Activities in Preparation for an Interview with the Press

- Learn about who reads the outlet and tailor responses to that audience
- Identify the Single Overriding Communication Objective (SOCO) and practice saying it to people with different backgrounds and interests
- Prepare how to explain in everyday language the meanings of relevant jargon
- Prepare and practice saying a single phrase or sentence that conveys or supports the SOCO
- Write down the SOCO, sound bytes, and most important statistics; bring this "cheat sheet" to the interview
- Carefully plan how to answer the six questions included in most science interviews: Why did you do the study? What is the one main result of the study? What challenges did you encounter and/or overcome? What mechanism explains the association? What is the public health message? Where should future research focus?
- Gather descriptive data on relevant determinants and outcomes
- Consult with public affairs personnel if that resource is available

For example, if the researcher provides two main findings, then constraints of space and time may force the reporter to select one of the two and thereby miss the main thrust of the work. Simplicity is not often a hallmark of research findings, but it is an imperative for successful communication of one's findings and interpretations to the public. Consequently, the SOCO must also be rid of jargon wherever possible. Should jargon be necessary, it is useful to be aware of the jargon in advance of an interview and to be prepared to define terms clearly. These two difficult tasks – to find simplicity within complexity and to translate scientific terms into commonly understood language, all without introducing inaccuracies – are critical and therefore deserve pause and consideration before beginning an interview.

One should realize that the job of the reporter is to interpret a story and then to convey it to a broader audience, not to be up-to-date on the scientific field relevant to the interview topic. It is therefore necessary for the researcher to help the reporter understand the essential background information, the purpose of the research, the *one* main finding, and any major ambiguities. Some researchers insist that the reporter read the related scientific paper before the interview, although the utility of that strategy is dubious to us because it probably has low yield, blurs the line of who is the epidemiologist, and risks alienating the reporter (reading a research paper can be difficult and time consuming even for an epidemiologist!). Alternatively, we suggest providing the reporter with a copy of the paper without insistence and to offer to serve as a resource for them should they have any questions in preparation for the interview.

30.3.3 During the Media Interview

During an interview – which may occur via e-mail, in online discussion forums, over the phone, or in person – the interviewee's goal is to turn the interview into an opportunity to relay the SOCO (Dan 2008). An imperative is to communicate the SOCO early and repetitively. If the reporter attempts to discuss a point that deviates too far from the SOCO, one commonly employed approach is to transition back to the SOCO using phrases such as 'the most important thing to say about that is…,' 'the bottom line in those situations is…,' or 'the take-home message is….' This technique is called *bridging*. Importantly, the phrases used in bridging are also useful for flagging the SOCO and thereby signaling to the interviewer the importance of what will be said next.

To help the reporter understand the context and importance of one's findings, the researcher will need to convey specific epidemiological terms and statistics that might have multiple or alternative definitions in common parlance or be difficult to understand. For example, the word *risk* to an epidemiologist signifies a concrete concept defined by an objective numerical value, but to the public *risk* may be an abstract concept implying a subjective degree of danger or hazard (Loukissas 2011). Given the potential for the reporter or reader to misinterpret the specific meaning of a term or statistic, it is imperative for the researcher not only to select carefully which statistics to discuss but also to phrase the terms associated with those statistics in the clearest way possible.

Typically, an interview will contain the following six basic questions (Loukissas 2011):

1. Why did you do the study?
2. What is the one main result of the study?
3. What challenges did you encounter and/or overcome?
4. What mechanism explains the association?
5. What is the public health message?
6. Where should future research focus?

If these questions are not asked directly, the researcher should attempt as naturally as possible to provide the answers during conversation, or give verbal clues that cause the reporter to ask these questions. It is acceptable for the researcher or reporter to politely redirect the interview to the most important content. Indeed, both parties control the content of the interview. Reporters do so mainly by asking questions, and interviewees do mainly by regulating how questions are answered and for how long.

Other common questions that the researcher should be prepared to answer are:

7. Did anything about the study surprise you?
8. Is there anything you would like to add?
9. Is there anything I did not ask you?

Other useful advice for participating in an interview can be found in summary format in Panel 30.4 (Dan 2008; Loukissas 2011). We refer readers to Dan (2008) for useful advice on how to deal specifically with press conferences and interviews over the radio/podcast or television.

Panel 30.4 Tenets of Communication During Interviews

- Be prepared
- Be concise
- Use plain language (avoid jargon); use the local language
- Avoid speaking too quickly; speak deliberately and carefully
- Employ a single overriding communication objective (SOCO)
- Lead with the SOCO and repeat it during the interview
- Avoid providing excessive levels of detail. Keep things simple
- Avoid giving seemingly different or contradictory information about the same study
- Avoid superfluous information
- Avoid all speculation
- Be a resource
- Throughout the encounter, develop a relationship with the reporter

Hint

Either at the beginning or end of the interview, it is common for the interviewee to request the opportunity to check quotes or even read the media article before publication. Some reporters will be accommodating, but this practice should not be expected, nor is it always possible.

30.3.4 Online Social Networks

Increasingly, social networks have become an outlet for disseminating health policy ideas and objectives. Social networks are a loosely defined group of websites through which users share and disseminate information. They include websites where users set up social profiles (e.g., Facebook), post public updates, follow other users' updates (e.g., Twitter), review local businesses (Yelp), post their current location and activity (e.g., Foursquare), rate content that they enjoy (e.g., Digg), or post videos and photos (e.g., Youtube and Flickr). Many of these websites overlap in functionality, and the available services are constantly evolving.

In 2006, Facebook began allowing organizations to create profiles. The creation of a profile on a social networking site can achieve many goals. Some of the uses that social network profiles can achieve include:

- Identifying and gathering contact information for people interested in the organization's product or policy goals
- Providing information, including links to pertinent news articles and websites
- Publicizing upcoming events
- Recruiting volunteers
- Providing a forum for discussion and feedback
- Soliciting donations

A content analysis of 275 nonprofit organizations' Facebook profiles found that the profiles of nonprofits tend to provide transparent explanations of their purpose. But many of them failed to take advantage of the social aspect of Facebook, rarely offering users opportunities to get involved (Waters et. al. 2009).

Furthermore, social networking can afford an opportunity to identify topics of discussion and possible questions, concerns, or points of intervention for organizations that wish to disseminate particular information. Many for-profit companies regularly monitor mentions of their products on Twitter in order to identify dissatisfied customers and address their concerns. This opportunity exists for more public service-oriented organizations as well. A recent study on the use of the word "antibiotics" on Twitter found hundreds of tweets that contained misinformation or discussed inappropriate uses of antibiotics (Scanfield et. al. 2010). These users could serve as intervention points for organizations searching to publicize a public health message, for instance.

30.3.5 Making Change in the Real World

It is helpful for the researcher to realize that information, in the traditional sense, is not the only essential prerequisite for action. Health information is often difficult to understand without substantial background. Epidemiologists are trained to think about public health issues from a highly technical, methodological, and intellectual perspective. It is therefore most natural for many epidemiologists to communicate about their work in such terms. However, most people rely on emotions, feelings, instincts, and heuristics to guide their actions. We often form judgments by subconsciously asking: "How do I feel about this?" (Oz 2010). When communicating with the press or public, one should keep this in mind and cater to the strengths of one's target audience, not to those of one's colleagues.

30.4 Communicating with Study Participants and Their Healthcare Providers

The most important stakeholders of any study are its participants. Study findings should be reported to them, perhaps employing the strategies discussed above regarding interviews with reporters. But sometimes the researcher must communicate with particular patients about new personal health information. This task requires a great deal of care and sensitivity.

30.4.1 New Personal Health Information

Researchers often acquire information on participants' personal health that was previously unknown to the participants and their health care providers. The new information may be of a diagnostic or prognostic nature. It is standard practice to provide such information to the participant's indicated primary health care provider, although participants should be given the option to bar this action (unless laws require otherwise). The new information can be quite sensitive, such as the risk of having, acquiring, or transmitting a strongly heritable illness or a potentially stigmatizing illness. Indeed, not all participants want to know if they have certain diseases or disorders. One option is to let the participants decide, as part of the informed consent process, whether they wish to be informed about new personal health information and in which way (White et al. 2008). However, there are rare instances in which laws mandate reporting of some diagnoses to individuals and/or agencies; therefore, it is necessary to identify and follow such laws and to explain these laws to participants during the informed consent process.

Moreover, new personal health information can also be of an acute nature that requires immediate intervention. Each study should have a system of referral and/or treatment of new health issues that may be discovered during the study. These provisions relate mainly to anticipated problems such as known side-effects of treatments or measurements, or possible values of health measurements (e.g., a positive HIV test).

30.4.2 Communicating Overall or Interim Study Results to Participants

To facilitate the communication of overall or interim study results to participants, one may use the strategies for communicating with the public described previously in this chapter. Common strategies for reporting study results are to send a newsletter or to host a special post-study gathering. In doing so, one should be careful to avoid creating anxieties that are out of proportion to the size of the risk (White et al. 2008), and to never reveal personal health information (a breach of confidentiality).

30.5 Striving for a Desirable Influence on Public Health Policy

Policymakers consist of administrators and elected officials who make decisions that affect laws, rules, and regulations. Ideally, they respond to health issues that have the greatest importance for public health, but in reality, politics often determine the salience of the issues at hand. Epidemiologists have the opportunity to interface with policymakers in order to increase the likelihood that they will effect change on the most important health issues and to intervene in a manner that is supported by empirical evidence. Indeed, epidemiologists can raise awareness of an issue and potentially change policymaker's priorities.

This discussion raises an important question: to what extent are epidemiologists responsible for ensuring that their research results are 'translated' into policy? The answer to this question is not clear-cut and comes with many caveats.

One obvious element of the answer is that any such responsibility should be limited by the internal and external validity of the research. ‘Translation’ should ideally not be attempted unless research evidence has turned into scientific knowledge. Scientific knowledge is a majority consensus phenomenon, so health policy decisions need to be informed by expert consensus on scientific issues. In reality, however, the experts often disagree on whether a sufficient basis for action has been reached, on what more evidence is needed, and minority views – whether they are right or wrong – can be influential.

Of course, scientists should be concerned about the consequences of their activities and findings, and therefore communication with policy makers is unavoidable. Viewed from the other side, health policy decisions need to be based on a rational system for prioritizing competing health intervention needs, and policy makers are heavily reliant on experts when it comes to understanding a body of research. Creating and maintaining active channels of communication between policy makers and the scientific community is therefore a dual responsibility of the involved parties. Scientific experts should be involved in health policy decision-making, but such involvement is contingent on relationships. Dissemination represents an important approach by which researchers can form and build relationships with policy makers. In creating communication channels, researchers should avoid the temptation to increase the impact of their work by becoming policy makers themselves, or by becoming activists who try to pressure public health authorities and politicians in a somewhat desperate way.

Efficient structures of communication between policy makers and the scientific community are needed for giving policies an appropriate evidence base and to avoid that influencing policy becomes merely a matter of who shouts the loudest, talks the smoothest, or has the most political friends. *Bounded rationality* – a term coined by Herb Simon – is the idea that, in decision making, the rationality of individuals is limited by the information they have at hand, the cognitive limitations of their minds, and the finite amount of time they have to make decisions. Given the complexities of health policy, it could be said that a consulting epidemiologist’s role is to reduce bounded rationality by efficiently communicating the most important information.

30.5.1 The Right Messages at the Right Time

It is relatively uncommon for a single research study to have any direct positive impact on public health. Conclusions based on just one study are the most susceptible to bias – although a conclusion based on no evidence at all is markedly less reliable unless that conclusion is agnostic – and should be promoted with caution. Without training in epidemiology, those who are exposed to a researcher’s SOCO are much less likely to understand the nuances of object design and statistics than a researcher’s fellow epidemiologists and are, therefore, usually unable to fully appraise a single study’s internal and external validity.

This problem is especially apparent when multiple studies on a similar topic yield seemingly different conclusions. For instance, if Study A and Study B represent the total available evidence on whether exposure to a substance is associated with heart failure, where Study A shows a 20 % increase and Study B shows a 20 % decrease in the risk of developing heart failure, then a group of people might reasonably conclude that there is, on balance, no association between the substance and heart failure. Yet, the two studies would likely suffer from differing degrees of bias, confounding, and other flaws that might make one study less valid than another. These complex features are usually omitted in media coverage of scientific research, thereby making the public vulnerable to misinterpretation of potentially useful health information. If Study A is more valid than Study B, then the public may not recognize the substance's potential harm.

Even if the internal validity of the study is optimal, there are known instances of authors incorrectly placing and interpreting the evidence in the light of external evidence on the same topic. The external evidence may only be partially mentioned and the author's own interpretation of the importance of their study may be exaggerated. Various stakeholders may then 'jump on' the most categorical and striking statements in a discussion section or in the conclusions of a scientific paper. Take away messages (Single Overriding Communication Objectives) can be dangerous if they are distorting simplifications. The result may then well be another wave of unfounded health-related anxieties and another contribution to over-medicalization of society. It is our view that the dissemination of research findings charges the researcher with the responsibility to attempt to protect the public from misinterpretation of potentially useful health information. This may occasionally imply restraining from dissemination of findings of single studies.

The same logic just described also applies to the dissemination of intervention campaigns. In such a scenario, however, there is usually a body of evidence supporting the efficacy or effectiveness of the intervention and sometimes even of the dissemination strategy. Selecting the best estimates to convey to the target audience is a challenging task, especially when multiple studies on a similar topic provide interpretations as dissimilar as Studies A and B above. Meta-analyses may be useful for selecting a best estimate but must be interpreted very carefully and always in the context of possible publication bias (*See:* Chap. 25).

30.5.2 Advice for Dissemination Aiming at Public Health Impact

Taking into account the above considerations, the appropriate combination of scientific and ethical concerns translates into the advice list for dissemination of research results in Panel 30.5.

Panel 30.5 Advice for Optimizing Public Health Impact of Research Through Dissemination of Research Findings

- Prime stakeholders for the receipt of new information before and during the data collection period
- Only publish valid study findings, in scientific journals, independently of size or direction of estimates, and with due attention to precision; publish in widely visible and easily accessible journals
- Discuss strengths and weaknesses, point out the need for more research if relevant, and avoid over-interpretations; fairly assess external validity
- Participate seriously with the peer review process in order to help other authors with refining the assessment of the internal and external validity of their work
- Only engage into advocacy when the overall evidence on the topic is convincing; work with other scientists towards a consensus
- Use appropriate intelligible language adapted to each stakeholder but do not simplify too much when communicating with stakeholders
- React appropriately to misinterpretations, undue simplifications and ill-founded advocacy from other scientists or other stakeholders
- Provide or contribute to complete information on a topic i.e., on burdens as well as on efficacy, safety, cost and acceptability of alternative intervention strategies
- Be active within the existing systems of communication between public health authorities and researchers or lobby for setting up or improving such structures

30.6 Additional Dissemination Resources

This chapter provides several different frameworks to be used as tools for the development of dissemination strategies. Though the body of literature on dissemination strategies is relatively sparse and usually unavailable for or irrelevant to a specific dissemination project, some of that body of literature and a wealth of experience have been embedded in many helpful resources. We recommend the following resources for those who wish to extend and deepen their knowledge of dissemination strategies:

- Cancer Control PLANET (Plan, Link, Act, Network with Evidence-based Tools): "Links to comprehensive cancer control resources for public health professionals." http://cancercontrolplanet.cancer.gov/index.html
- RE-AIM: A planning tool to facilitate translation of research into action. http://cancercontrol.cancer.gov/IS/reaim/whatisre-aim.html
- Research to Reality: "An online community…that links cancer control practitioners and researchers and provides opportunities for discussion, learning, and

enhanced collaboration on moving research into practice." https://researchtoreality.cancer.gov/

- PRIMER: "A toolkit for health research in partnership with practices and communities." *See:* the "Disseminating and Measuring Impact" page in the "Disseminating and Closing Research" section. http://www.researchtoolkit.org

In this chapter we discussed theories about how information spreads among groups of people and society, and we leveraged those theories to advise on the development of a dissemination strategy. Dissemination to stakeholders (Chaps. 29 and 30) raises some ethical issues with study reporting. Some of these have already been mentioned, but several major issues around publication/authorship policy and publication bias remain to be discussed in the final chapter.

References

Bauman A et al (2006) Physical activity measurement – a primer for health promotion. Int Un Hlth Promot Educ 13:92–103

Buller DB (2006) Diffusion and dissemination of physical activity recommendations and programs to world populations. Am J Prev Med 31:S1–S4

Cuijpers P, de Graaf I, Bohlmeijer E (2005) Adapting and disseminating effective public health interventions in another country: towards a systematic approach. Eur J Public Health 15:166–169

Dan BB (2008) Dealing with the public and the media. In: Gregg MB (ed) Field epidemiology, 3rd edn. Oxford University Press, Oxford, pp 1–572. ISBN 9780195313802

Dearing JW et al (2006) A convergent diffusion and social marketing approach for disseminating proven approaches to physical acivity promotion. Am J Prev Med 31:S11–S23

Ellis P et al (2005) A systematic review of studies evaluating diffusion and dissemination of selected cancer control interventions. Health Psychol 24(5):488–500

Estabrooks PA, Glasgow RE (2006) Translating effective clinic-based physical activity interventions into practice. Am J Prev Med 31:S45–S56

Glasgow RE, Vogt TM, Boles SM (1999) Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 89:1322–1327

Loukissas J (2011) When epidemiologists talk to the press and public. Paper presented at the 3rd North American Congress of epidemiology, Montreal, 24 June 2011 (Oral presentation)

McQueen D, Puska P (2003) Global behavioral risk factor surveillance. Kluwer Academic/Plenum, New York, pp 1–262. ISBN 9780306477775

Nelson DE et al (1998) Objectives and design of the behavioral risk factor surveillance system. In: Proceedings of the American statistical association section on survey research methods, Dallas TX, 9–13 August 1998

Oz MC (2010) How surgery taught me about media. World J Surg 34:635–636

RE-AIM Checklist for Study or Intervention Planning. National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. http://www.crn-ccrc.org/design-and-methods/tools/checklists.aspx. Accessed Sept 2012

RE-AIM Planning Tool. National Cancer Institute, National Institutes of Health, Bethesda MD. http://www.crn-ccrc.org/design-and-methods/tools/checklists.aspx. Accessed Sept 2012

Rogers EM (2003) Diffusion of innovations, 5th edn. The Free Press, New York, pp 1–512. ISBN 9780743222099

Scanfield D et al (2010) Dissemination of health information through social networks: Twitter and antibiotics. Am J Infect Control 38:182–188

Waters R et al (2009) Engaging stakeholders through social networking: how non-profit organizations are using Facebook. Public Relat Rev 35:102–106

White E, Armstrong BK, Saracci R (2008) Principles of exposure measurement in epidemiology. Collecting, evaluating, and improving measures of disease risk factors, 2nd edn. Oxford University Press, Oxford, pp 1–428. ISBN 9780198509851

Yancey A, Ory MG, Davis SM (2006) Dissemination of physical activity promotion interventions in underserved populations. Am J Prev Med 31:S82

The Ethics of Study Reporting

31

Eimear Keane, Jan Van den Broeck,
and Jonathan R. Brestoff

Evidence described in medical journals tends to constitute, in the aggregate, a biased base for learning about any given issue.

O.S. Miettinen

Abstract

The reporting of research evidence is vital for the achievement and distribution of knowledge and the advancement of science. This chapter discusses the ethical aspects of reporting, with a prime focus on ethical issues associated with publishing scientific reports. Misconduct in study reporting can occur in a number of forms. This includes misleading reporting, plagiarism, and misrepresenting authorship. These practices as well as publication bias undermine the general principles of epidemiology. Furthermore, authors need to respect the right to confidentiality of research participants in their publications and avoid causing stigma to participants, communities, and themselves. They are also responsible for disclosing all potential and real conflicts of interest, of which there are many types, such as intellectual property.

E. Keane, MPH (✉)
Department of Epidemiology and Public Health, University College Cork,
Cork, Ireland
e-mail: Eimear.Keane@ucc.ie

J. Van den Broeck, M.D., Ph.D.
Centre for International Health, Faculty of Medicine and Dentistry,
University of Bergen, Bergen, Norway
e-mail: Jan.Broeck@cih.uib.no

J.R. Brestoff, MPH
Perelman School of Medicine, University of Pennsylvania,
Philadelphia, PA, USA
e-mail: brestoff@mail.med.upenn.edu

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3_31,

31.1 General Advice on Publication Policy and Data Sharing

Selected terms and concepts around the ethics of publication are defined in Panel 31.1.

Each new publication should add additional evidence or insight to the already available pool of scientific information. Some important general responsibilities of authors seeking to have their study findings published in scientific papers are listed in Panel 31.2. The first four points of the panel are discussed succinctly in this section; each of the others is covered more extensively in other sections.

In epidemiological research multiple outcomes can be addressed during the same study. It is good to make a clear distinction between reports of the primary outcome and reports of secondary or tertiary outcomes. This is especially advisable for

Panel 31.1 Selected Terms and Concepts Around the Ethics of Publication

Author Researcher taking the responsibility of writing and defending the content of a scientific publication

Authorship rules List of minimum criteria regarding who should be invited to take up the responsibility of (co-)authorship

Co-author *See:* author

Copyright Law-enforceable exclusive right to publish, copy, adapt and distribute

Peer review Pre-publication check, by scientists knowledgeable of the type of content at issue, of the validity of the presented manuscript and its acceptability for publication in the journal or other publication medium

Publication bias Bias in the overall pool of evidence around a research question due to selective publication based on the magnitude or direction of study findings

Publishing Making a document publically available

Panel 31.2 General Ethical Guidelines Regarding Publication Policy

- Avoid redundant or (partially) duplicate publication
- Choose peer-reviewed journals for publication
- Publish in a timely manner in a peer reviewed journal
- Consider making anonymised raw data publicly available
- Avoid any form of misleading reporting and plagiarism
- Fairly manage authorship
- Publish regardless of how non-spectacular the results are
- Avoid breaches of confidentiality and stigma
- Avoid or declare potential and real conflicts of interest

large-scale multi-center trials that may have a number of specific endpoints. Secondary outcomes should not be presented as if they were primary. Likewise, reports of preliminary data and re-analyses should be clearly identified as such and justified. This will help clarify any issues of (partial) duplication or 'salami slicing' (*See:* below) for editors and other researchers working in the same field.

Redundant or (partially) duplicate publication occurs when numerous papers with overlapping results are published using data from the same study. This form of publication may not always be apparent, especially if papers are (1) without cross-reference, (2) are without acknowledgement of the original study, or (3) list different authors with each new publication. This form of publishing can be very misleading as excessive weight can be given to observations that have been repeatedly reported. This is especially true when such papers are all included in systematic reviews or meta-analyses (Huston and Moher 1996).

The term '*salami slicing*' refers to the practice of unnecessarily pursuing multiple publications from a single study. Large quantities of published literature can be produced using a limited amount of data, each bringing fragments of evidence, or each with only a slightly different angle or methodological adaptation. Unless there is a convincing rationale, these publications are likely to be repetitive, misleading (particularly in terms of results produced subsequently by systematic reviews and meta-analyses), and wasteful of other researchers' time (Huston and Moher 1996).

It is wise to give preference to *peer-reviewed journals* when choosing which journal to publish a research paper in. Prior to publication, the editor of a peer-reviewed journal will send the manuscript to other experts in the same field of research. These 'experts' critically assess the manuscript in order to ensure that scientifically valid research methods were used. They also evaluate the external validity and make judgments about study implications. Publishing literature in peer-reviewed journals is often associated with greater levels of future citation. This should be viewed as an incentive to publish work in such journals.

Research findings, especially those from trials, should always be made available and in a timely fashion, as failing to pursue publication is a form of misconduct, unless it is apparent that internal validity is severely compromised. *Timely publication* can have a positive impact on both clinical practice and community medicine. Results of trials are of little benefit to patients, practitioners, policy makers, or the general population unless they are reported timely and clearly, without any spin or data framing (Kramer et al. 2006). Sometimes sponsors and scientific journals require authors of an article to be prepared to share their raw data with other researchers. There are many advantages to *data sharing* (Hrynaszkiewicz et al. 2010; Campbell and Blumenthal 2002). These include:

- Increased transparency in research
- Better possibility to replicate research findings
- Avoidance of unnecessary data collection
- Increased opportunities to carry out meta-analyses
- Generating and testing other hypotheses related to that topic
- Use of the data for teaching purposes

The data should be made publicly available along with the relevant metadata, such as questionnaires, a data dictionary, and other data-related tools. It is essential, however, that data sharing occurs in a way that respects the rights of the participants. Confidentiality and stigma issues are discussed later in this chapter. Any negative impact of an intervention needs to be discussed with the participants before the results are made widely available (Partridge and Winer 2002).

Data withholding is still common among researchers. One survey of geneticists found that 47 % of respondents had been denied a request for data in the past 3 years (Campbell et al. 2002). Of the respondents who had themselves denied data requests, 80 % reported that fulfilling the request took too much effort. Other reasons given included the cost of transmitting the data, and protecting the right to publish using the data first before sharing it with others. Seventy-three percent of the respondents felt that data hoarding had been detrimental to the progress of science in their field.

Many types of data sharing arrangements exist, and they vary in their origin and purpose. On a peer-to-peer level, investigators often respond personally to data requests by extracting and sending the datasets personally. More formally, so-called *data enclaves* have been created. These are controlled, secure environments in which eligible researchers can use data resources to perform analyses. Similarly, a "data archive" is a place where machine-readable data are acquired, manipulated, documented, and distributed. Mixed modes of data sharing have also been used, with more than one version of a dataset made available, each providing a different level of access. Despite these arrangements, much data are proprietary and unavailable to most researchers. Issues of restricted access to data are an ongoing discussion in the research community. There is a need for a global electronic data archive of accessible research data and their metadata. But such a resource is not currently a reality.

In general, a *data request* should contain the following information:

- Research team and institution
- General aims and specific research questions that will be addressed
- Specification of variables and records
- A timeline and agreement on how long data access will last. If after a certain period of time no research has been produced, the data should be made available to other researchers with similar research questions

Once access to data is granted, a *data sharing agreement* should be put in place to make expectations and responsibilities explicit. These agreements should address:

- What the data will be used for
- Agreements about sharing of data with third parties
- A confidentiality statement that addresses any proprietary, patent, or privacy issues
- A description of the dataset and information on how the data were gathered
- An agreement about co-authorship, if appropriate
- A timeline on the length of data access

In addition to the data, proper documentation is needed to ensure that others can use the dataset and to prevent misuse, misinterpretation, or confusion. The dataset should be provided without personal identifiers or any information that could violate confidentiality for the research subjects. One open question is whether only “raw” untransformed data should be provided, or if derived variables should be included. In general, this decision is left to the discretion of the provider of the dataset. An anonymized dataset needs to be accompanied with a maximum of metadata, including the questionnaires used to collect the data, the study protocol, data collection Standard Operating Procedures, value code lists of variables, etc. Providing this ancillary information is likely to maximize the understanding of the data and avoid misinterpretation and wasted time on the part of the recipient researcher.

Once the main findings of a study have been published, data sharing should start as quickly as possible. Sometimes data from a large study can be publically released in waves, as publications using different aspects of the data are released. For example, large questionnaire surveys may have sections that more or less represent the different topics for publication. Once a topic is published, data can be released for sharing.

Many scientific publications require or request the publication of source data, citing the importance of transparency and reproducibility. Despite these requirements, there is not universal compliance, and the enforcement of these rules varies from journal to journal. Often, it is not clear who should enforce these good practices. Should it be the journal, the funding agency, or the academic institution? This question remains an important – and undecided – issue in the field of data sharing.

31.2 Types of Misconduct in Scientific Reporting

Research misconduct can occur during study reporting. Forms of ethical breaches that may occur can be categorized as follows:

- *Data fabrication*: This involves presenting fictitious data.
- *Misrepresenting evidence*: This includes manipulating data, such as removing unexpected results with the intention of achieving desired results
- *Plagiarism*: This involves using the thoughts and words of others without proper recognition by referencing or quotation
- *Misrepresentation of authorship*: This includes ghost authorship and coerced authorship (these terms are explained in greater detail below)
- *Delaying publication* for personal gain or to satisfy sponsor expectations: Personal gain for research investigators may include prestige or financial gain. Sponsors can delay publication to obtain patents (though pursuing patents may be reasonable if there is therapeutic potential that is very unlikely to be realized without the protection conferred by a patent)
- *Failure to publish* results: This creates publication bias. Failure to publish results is commonly associated with the direction or magnitude of the study results. Inconclusive or negative results are less likely to be published when compared to positive results, but this should not dissuade the pursuit of publication. However, failure to publish is not misconduct if the researchers determine that there is very poor internal validity (in which case the reasons for poor internal validity may serve as an alternative subject on which to publish so others can avoid preventable errors)

Misleading reporting is a broad term that encompasses many of the aforementioned ethical breaches in study reporting (e.g., data fabrication, misrepresenting evidence, and plagiarism), but there are other ways in which reports can be misleading. One example is reporting point estimates of measures of effect *without associated confidence intervals*. This practice can be very misleading because it may make inconclusive data appear to be significant. Failing to show interval estimates around a point estimate should raise suspicion that the authors are hiding wide confidence intervals to avoid having their results considered inconclusive or due to a major methodological weakness. Another unacceptable approach that others have taken is *manipulation of graphs*, e.g., stretching or shrinking the abscissa and changing the ordinate. This may not be entirely evident when one is reading a manipulated graph. The reader can be misled into thinking that the results are more significant than they are, that a relationship is stronger than it is, or that an outcome appears sooner than it does in reality.

Furthermore, lack of protocol adherence information in the results section and lack of discussion of limitations can disguise any problems with design, data collection, and analysis. *Hiding design weaknesses and implementation difficulties*, such as an unrepresentative survey sample, non-blinding in a randomized controlled trial, and important random and systematic errors of measurement can make the statistical results appear more unbiased than warranted (*See also:* Chaps. 11 and 27). One must therefore be forthcoming in design weaknesses and implementation difficulties. In fact, shedding light on these issues tends to make the researchers appear honest, which can in turn be useful information when interpreting the results of the study (as honest investigators tend to be forthcoming).

Hint

Statements regarding the safety of an intervention past the total follow-up time cannot be made. Investigators cannot be certain that an intervention will continue to be safe in the long-term. Therefore, claims regarding the long-term safety of an intervention should not be made. Even within the trial follow-up time, a trial may not be large enough to make usefully precise estimates about occurrence frequency of safety issues.

Research misconduct can have a number of negative consequences which include:

- Fraudulent work can reduce the integrity of epidemiological research. In turn, the public's confidence in scientific research may be reduced as a result of research misconduct (Benos et al. 2005).
- Dishonesty of researchers can waste the time of other researchers. Unnecessary studies have been conducted as a result of unreliable data from fabricated results being published.
- Dishonest and invalid evidence (that is obscured to appear valid) may lead to ineffective or harmful interventions being put in place or, conversely, it can lead to effective interventions not being put in place

31.3 Ethical Issues of Authorship

31.3.1 Writing Groups

Writing groups have a responsibility to ensure that study reporting and publication occurs ethically along the lines of Panel 31.2. Writing groups should complete work fairly, honestly, and objectively (Benos et al. 2005), and working as part of a writing group should be considered a means of making it more difficult for any one individual to act fraudulently. In general, writing groups can be considered as a safeguard against misleading reporting. All members should be trained and supervised to ensure that they are completing their tasks both effectively and ethically (MRC 2005).

Differences in publication expectations between the sponsor and the investigators are an issue for the writing group. Any conflicts of interest arising for any member of the research team can negatively impact the quality of the results. For example, fraudulent study reporting can occur as a result of a financial incentive offered by the sponsor to the investigators of a study. Company-sponsored trials are more likely to produce favorable results compared to trials sponsored by bodies with no vested financial interest in the result of that study (Perlis et al. 2005).

31.3.2 Authorship Rules

Authorship is both a responsibility and a privilege. Therefore, only those who are willing to accept responsibility for at least one crucial aspect of a study, such as the study design, and in addition are capable of publicly defending the content of the scientific paper can be authors (Benos et al. 2005). However, the number of individuals who are included as authors on scientific papers is increasing. This is making it more difficult to differentiate between major from minor contributors. There are three requirements that need to be fulfilled for one to be considered a co-author, according to the International Committee of Medical Journal Editors (ICMJE 2008):

- Significant contribution to study design, conduct, data acquisition, or data analysis
- Significant contribution to revising or drafting the paper
- Read and approve the final draft before submission for publication

It has been argued that the criteria outlined by the ICMJE are too vague, which results in inconsistent enforcement (Laflin et al. 2005). Authorship is often considered unwarranted if the person's role and contribution solely concerns:

- Interviewing
- Acquisition of study funding/financial support
- Having been a member of the study personnel
- Scientific advising
- Technical support, e.g., laboratory technicians, IT specialists

Eligible authors should not be deliberately excluded from the authors list. The term 'ghost authorship' is used to describe such an occurrence. Once authorship has been established, those who accept the responsibility must sign an authorship statement. Other forms to be completed by authors include those signifying responsibility for the work and disclosure of interest forms (Flanagin et al. 2002). Some scientific journals require a description of each author's contributions, in order to address authorship issues. The Journal of the American Medical Association is an example of such a journal (JAMA 2006). The contribution that each member of a writing group made is commonly placed in the footnote section (Flanagin et al. 2002).

An *acknowledgements section* is commonly included at the end of a paper. The acknowledgements section may include the names of those who contributed to a study in some way but who do not fulfill all the necessary authorship criteria (listed in the authorship rules section above) (Laflin et al. 2005). Conventionally, the authors of an article are responsible for obtaining written permission from all persons to be acknowledged by name.

31.3.3 Author Listing

With the number of people collaborating on a single paper increasing, it is becoming increasingly difficult to establish authorship order. Traditionally those whose names are placed first and last on a paper are considered the most significant authors. The author who produces the original draft of a paper is typically placed as the first author (Benos et al. 2005). In turn, the last author tends to be the individual who is at the most senior level. Some scientific journals allow for several members of the writing group to share a joint first authorship status.

Group authorship occurs when the name of the group (e.g., the name of a multicenter randomized controlled trial group) is listed rather than listing each author's name (Dickersin et al. 2002). This allows for equal credit to be allocated to each individual investigator. Modified group authorship can occur when the name of each investigator is placed before or after the name of the group. Some universities and sponsors judge scientific output of professionals using a system that gives more credits for those listed as the first and last author when compared to other positions on the paper. Thus, group authorship can be particularly useful for large-scale and multi-center studies where it can be difficult to give everybody fair credit. This in turn may avoid issues in authorship that can lead to disputes and slow down publication (Horner and Minifie 2011b).

Other ethical issues associated with author listings include:

- *Honorary authorship:* This occurs when the name of a well-known person or expert in a certain area is placed on a paper. This individual may not have worked on the paper or may not meet the three necessary authorship criteria. Honorary authorship commonly occurs out of obligation (Laflin et al. 2005) or in order to increase the possibility of publication (Feeser and Simon 2008)
- *Coerced authorship:* This arises when senior members of a study group use their status to get their names placed on the authorship list of a paper. These senior members of staff may not meet the authorship criteria

- *Ghost authorship:* This involves deliberately excluding a person's name from the authorship list, though they meet the authorship criteria. This can occur intentionally. This has occurred in industry-sponsored trials in order to conceal connections between potential authors and industry (Feeser and Simon 2008)

31.4 Publication Bias

Publication bias is a phenomenon that receives a lot of attention, as it has a number of ethical and scientific consequences. Montori et al. (2000) published an introduction to this topic.

31.4.1 What Is Publication Bias?

Publication bias is a bias in the overall pool of evidence surrounding a research question, and is the result of the selective publication of manuscripts based on the magnitude or direction of the study results (Montori et al. 2000). One study concluded that positive studies are up to three times as likely to be published when compared to inconclusive studies (Egger and Smith 1998). Positive results are also more likely to get published in higher profile journals (Easterbrook et al. 1991). Along with this, positive results also get published sooner and are cited more regularly when compared to negative or inconclusive studies (Dickersin and Rennie 2003). It can be argued that all these perceived benefits of creating positive results contribute to publication bias.

Publication bias tends to be greater for observational studies than for randomized trials. Small-scale studies are more likely to produce inconclusive results. Therefore, they are more likely to remain unpublished (Begg and Berlin 1989; Vickers et al. 1998).

31.4.2 How Does Publication Bias Occur?

Publication bias is an important problem both from a scientific and an ethical perspective. Therefore, understanding the roots of this phenomenon and finding ways to prevent and amend it are important. In fact, there are several stages of the research process where there are forces that can act towards publication bias. They are listed in Panel 31.3.

It is commonly misperceived that publication bias occurs predominantly because journals reject manuscripts with non-significant results (Scholey and Harrison 2003). On the contrary, the primary explanation for publication bias is the failure of investigators to complete or submit negative or inconclusive results for publication (Easterbrook et al. 1991). Investigators may decide that their non-significant or inconclusive findings do not add anything to the pool of evidence already available, that they are undeserving of publication, or that publication would be harmful for

Panel 31.3 Stages of the Research Process in Which Publication Bias Can Arise

- Completion of data collection
- Decision to write an article and submit for publication
- Initial evaluation by editors
- Results of peer review
- Decision to re-submit after initial rejection

their reputation. They may also think that their findings are unlikely to be published. However, non-publication can be viewed as disrespectful towards study participants and sponsors (Scholey and Harrison 2003).

In the context of evidence-based medicine, systematic reviews and meta-analyses are commonly conducted in order to assess existing evidence on the effectiveness of a certain treatment or intervention. The reliability of such overall evidence depends on the completeness and quality of the available literature (Tumber and Dickersin 2004). Selective publication of positive results creates a biased estimate for the overall effects of a treatment or intervention. Non-reporting of negative or inconclusive results is an act of scientific misconduct as it contributes to building this biased pool of evidence for decision-making (Chalmers 1990), though deciding not to report a study that has poor internal validity is not a form of misconduct. The decision of an investigator not to submit for publication can result in redundant trials being conducted, which may unnecessarily expose study participants to a potentially harmful intervention. Harmful or ineffective interventions may also remain in practice due to non-publication (Tumber and Dickersin 2004).

31.4.2.1 Publication Bias by Interruption of Data Collection

Some investigators or sponsors may decide to stop data collection for reasons that are associated with an interim picture of significance or magnitude of the effect size. Investigators may feel that the interim results are not spectacular enough to warrant publication. It may also be that the results are not in the originally predicted direction (positive or negative). Sponsors may have a view that the research is no longer a worthwhile investment and therefore ‘pull the plug’ on the funding.

31.4.2.2 Selective Decisions to Write an Article and Submit for Publication

Some investigators or sponsors may decide not to report study results during the analysis stage. This again may be associated with significance or magnitude of effect. Another possible explanation regarding decisions not to publish results may relate to a publication veto from sponsors. This occurs when sponsors prohibit investigators from publishing results without their approval. Sponsors may put this publication veto in writing in the form of a contract to be signed by investigators prior to the study commencing.

31.4.2.3 The Responsibility of Editors and Peer Reviewers in Publication Bias

Journal editors or conference organizers may refuse to consider a publication. Their reason for rejection may be influenced by the significance/magnitude of the study results. Journal editors tend to have a preference for results that are positive and spectacular. The editor's knowledge of how a study was funded can also influence their decision. For example, government agency funded studies are more likely to get published than those funded by pharmaceutical companies. Editors may perceive the results of studies produced by government agencies as more important or reliable than studies funded by other means (Easterbrook et al. 1991). This potential source of publication bias has probably received more attention than any other explanation.

It is possible that some peer reviewers are more critical of studies with less spectacular results. The reviewer could also potentially decide that a study without any spectacular results does not add anything to the overall pool of evidence already available on that certain topic. Editors commonly ask reviewers to judge an article based on its 'appropriateness' for the journal. This can turn out to play a role in publication bias.

31.4.2.4 Decision to Re-submit After Initial Rejection

A number of factors may influence an author's decision to re-submit for publication. These may include the amount of time a person is willing to spend on one paper or perhaps the benefits that a person expects as a result of publication. For example, the prospect of career advancement may be felt to be stronger when the study findings are positive. Authors who view their results as 'fascinating' or of clinical significance may be more likely to re-submit their paper for publication.

31.4.3 How Can Publication Bias Be Detected?

There are two main approaches to study the existence of publication bias in a certain domain. The first approach is to show that one or more of the selection processes described earlier have played out in the studies about the topic. A potential method to do this is to follow registered trials and studies submitted to ethics committees. Of the studies that remain unpublished, each could be assessed for explanations regarding failure to publish.

The second approach is to detect deficiencies in the pool of published evidence on a particular topic. This approach involves assessing systematic reviews or meta-analyses. Assessing asymmetry in funnel plots can sometimes show the existence of publication bias. A funnel plot is a graph that depicts, for all studies included after the literature search, the sample size or the inverse of the standard error as a function of the outcome estimate's magnitude (*See:* Chap. 25). When a large number of studies are included, a plot without publication bias will be symmetrical in shape and will resemble an inverted funnel. Asymmetry of the funnel can, under certain conditions, point to the existence of publication bias. Statistical methods to test for asymmetry have been developed; however, the validity of these methods has been questioned (Sterne et al. 2001).

Textbox 31.1 A Case of Publication Bias: The Effects of Antiarrhythmic Drugs on Mortality Rates in Patients with Myocardial Infarction

Antiarrhythmic drugs were administered to patients following an acute myocardial infarction, as there were biologically plausible reasons for administering the drug. Furberg published a systematic review in 1983 which consisted of 14 trials assessing the relationship between class 1 antiarrhythmic drugs and myocardial infarction. This meta-analysis did not detect a beneficial effect on the primary outcomes. However, the results demonstrated an increase in sudden death occurring in patients with ventricular arrhythmias. Antiarrhythmic drugs continued to be used in practice, as the evidence from this review did not convince clinicians to change their behaviors with regard to this drug. Many additional trials were conducted assessing this relationship between antiarrhythmic drugs and myocardial infarction. In 1993, a study that was conducted in the 1980s was published. This study demonstrated the harmful effects of administering the drug. The use of class 1 antiarrhythmic drugs was then halted. If the results of this study were published in a timely manner, the use of this drug would have been stopped earlier and many lives would have been saved. Estimates suggest that in the US alone between 20,000 and 75,000 people died each year during the 1980s as a result of the inappropriate administration of the drug (Dickersin and Rennie 2003).

Ethical issues arising from this example include:

1. The importance of timely publication of research
2. The effects that non-publication of results can have on meta-analyses
3. The effects that non-publication and late publication can have on clinical practice along with serious consequences in terms of morbidity and mortality
4. Redundant research being conducted as a result of non-publication and late publication. This redundant research wastes the time researchers, wastes resources, and can harm participants

Additional reading on this topic: Teo, Yusuf and Furberg (1993)

31.4.4 Publication Bias in the Literature: An Example

An example of publication bias and its consequences is described in Textbox 31.1. This example highlights the possible serious consequences of failing to publish the results of a study. A number of ethical considerations are highlighted.

31.4.5 What Can Be Done About the Problem of Publication Bias?

Solving the problem of publication bias will require a huge shift in the scientific tradition within medicine. Several strategies to tackle the problem have been proposed

Panel 31.4 Possible Strategies to Combat Publication Bias

- Peer review and editorial decisions based on papers submitted without outcome parameter estimates and P-values
- Promotion of group authorship
- Clinical trials registries and observational research registries
- Websites for posting study protocols
- Journals focusing on publication of negative and inconclusive results
- Incentives to promote publication of negative results
- Incentives for publishing unpublished studies

or attempted (Panel 31.4 lists some). Some measures have been put in place, such as clinical trial registries, but overall the problem remains acute, especially for observational research. Debate about the best strategy is ongoing.

Websites containing study protocols that are published prior to study commencement can make people aware of research that is currently being conducted. Such websites would also make it difficult for investigators to change research methods during the trial.

Other proposed strategies, such as journals focusing on the publication of negative results or incentives to promote publication of negative results, have not yet been met with any enthusiasm. Increasing the recognition of those who work on studies who produce negative results can also potentially improve the problem. Investigators, institutions, and those who fund such studies should be increasingly acknowledged for their work. Studies that produce negative results can be just as informative as those that produce spectacular results. They can, for example, demonstrate interventions and treatments that are not effective. Such studies can also demonstrate an investigator's ability to do good quality research (Tumber and Dickersin 2004).

31.5 Confidentiality and Stigma Issues in Publication and Data Sharing

31.5.1 Confidentiality

Study participants have the right to confidentiality. When preparing a manuscript for publication, it is essential that all participants of a study remain anonymous. Any written identifying information must be removed from tables, graphs, and associated text. For example, any initials, dates of birth, hospital record numbers, etc. must be removed from the study results prior to publication. The only exception is when a participant gives written consent to publish some potentially identifying information. For example, participants could be identified by placing photographs or other identifying information in the paper. In such circumstances, participants should be made aware of this in the informed consent process and give explicit written permission to allow potentially identifying information to be published. Under this consent, they are allowed to view the manuscript prior to publication (ICMJE 2008). In any picture, the participant's identity should not be immediately recognizable (e.g., by masking a person's eyes), and potentially distinguishing bodily features (e.g., tattoos and other body art) should be covered. All of these measures must be outlined in the informed consent process.

Breaches of anonymity can occur in data sharing when the shared datasets contain direct personal identifiers, extreme or very rare true values of variables (e.g. a woman who birthed a very large number of children in a small, tightly knit community), or responses to open-ended questions (e.g., unique phraseology). It is therefore necessary to trace all 'dangerous' data values and reduce the amount of information on the participants concerned in any publication or in any dataset that is to be publically archived.

Textbox 31.2 Stigma and Study Reporting: An Example for Discussion

Podoconiosis (non-filarial elephantiasis) is an endemic condition in many parts of Africa and is most prevalent in barefoot communities. It results in progressive swelling of the lower limbs due to long-term exposure to red clay soils of volcanic origin. Recent evidence has suggested that there is a genetic basis to the disease, implying that it is possible for entire families to face stigmatization as a result of an individual's condition (Tekola et al. 2009). Those who suffer from podoconiosis are stigmatized in a number of ways. For instance, they have been excluded from schools and local events and banned from marrying those who are unaffected by the condition. In the study by Tekola et al., those already suffering from the condition were afraid that genetic research would only further increase stigma and social isolation experienced by families and communities suffering from the condition.

Discussion Points Consider that you are preparing to conduct and publish a study assessing family and genetic factors associated with podoconiosis in an endemic region.

1. What ethical issues do you need to take into account?
2. How would you ensure that participants in the study remain anonymous?
3. What ethical issues must you consider while preparing a manuscript for publication?
4. What forms of study reporting would you use to increase awareness and reduce stigmatization associated with the condition in endemic areas?

31.5.2 Stigma

Individuals with a certain disease or specific groups of people (such as those in a specific geographic location) can be stigmatized as a result of publication. Stigma is especially prone to occur if an illness has adverse outcomes, is acquired as a result of risky behavior, or results in an altered appearance. Textbox 31.2 outlines an example of a stigmatizing condition and provides some discussion points.

Authors can also be stigmatized and even *ostracized* as a result of the work that they publish. For example, their published findings may indicate that a health service policy systematically under-serves a certain segment of the population. There should be no punishment for these so-called 'whistle blowers.' Such honesty is necessary to ensure that interventions and policies do not negatively impact individuals or populations and do not waste public resources.

31.6 Conflicts of Interest

Conflicts of interest can occur at any stage of the research process (MRC 2005). Thompson (1993) described a conflict of interest as "a set of conditions in which professional judgment concerning a primary interest (such as a patient's welfare or the validity of research) tends to be unduly influenced by a secondary interest (such as financial gain)." Financial gain is one of many factors that can result in such a conflict. A financial conflict of interest can take many forms, such as fees for consulting or speaking, employment, or stock ownership (Perlis et al. 2005). Other factors that can generate conflict of interest include personal beliefs, relationships, political factors, religious considerations, and academic competition (Benos et al. 2005; Campbell et al. 2007; Krimsky and Rothenberg 1998).

The level of interaction occurring between for-profit companies and medical researchers has increased significantly over recent years (Morin et al. 2002). Many researchers are dependant on for-profit companies in order to assist with funding of projects. These academic-industry collaborations are a common source of conflicts of interest. Non-profit organizations may also have special interests, so their involvement in any stage of a research project may also introduce concern over conflict of interest.

The existence of a secondary interest does not necessarily imply a conflict or any wrongdoing on the part of the researcher (Haines and Olver 2008; Krimsky and Rothenberg 1998).

However, conflicts of interest can result in poor decision-making, introduce bias into a study (e.g., modifying design to favor one result over another), or even lead to criminal offences (e.g., altering or falsifying results to achieve the sponsor's desired outcome). Conflicts of interest can also lead to a perception that the researchers were 'bribed' for-profit companies (Haines and Olver 2008).

31.6.1 Disclosure of Conflict of Interest

Fundamentally, it is not unethical for a researcher to have a conflict of interest. But if a conflict of interest exists, it must be recognized and dealt with in an appropriate manner. A typical solution is to fully disclose all secondary interests, especially if the investigator feels uncomfortable at the thought of others becoming aware of any secondary interests. Indeed, the credibility of the relevant research and of epidemiologic research in general can be improved by disclosing potential conflicts of interest (MRC 2005).

When submitting a manuscript for publication, authors are fully responsible for disclosing all potential conflicts of interest. To prevent any uncertainty, the authors must state specifically where any secondary interests exist. Conflict of interest notification pages must be placed in manuscripts sent to journal editors for review. It is then the editor's decision whether to publish the information provided by authors regarding any conflicts (Davidoff et al. 2001). Additionally, conflicts of interest may be disclosed in the acknowledgements section, an approach that may afford to the authors more control over disclosure. Current practice is to disclose only potential major conflicts of interest, such as holding more than $10,000 (USD) equity in the sponsoring company. Minor potential conflicts of interest (e.g., holding less than $10,000 in equity) are not typically disclosed but perhaps should be. Failing to disclose potential conflicts of interest, especially financial ones, is unadvisable and highly risky.

If a manuscript was produced using data from industry-sponsored research, editors of journals may ask authors to sign statements to ensure full disclosure of any conflicts of interest. Editors can assess the role of the sponsor in data collection, analysis, and study reporting and may ask to review the study protocol and any contracts signed between the sponsor and the investigator (Davidoff et al. 2001). This is important because contracts developed by sponsors can:

- Limit the amount and types of data that the author has available for publication. For instance, extreme values, outliers, or confounding variables may be excluded. This may disguise any adverse events associated with a particular drug or intervention or produce misleading associations
- Reduce the amount of power that an author has to ethically and honestly report study results. Investigators may have to present results to the sponsor prior to publication. Sponsors can then decide what parts of the manuscript are suitable/unsuitable for publication
- Potentially allow for publication bias. Unfavorable results may remain unpublished if the sponsor has power over whether the study results are to be published (Davidoff et al. 2001)

Possible reasons for non-disclosure of conflicts of interest include:

- Payment that is indirectly related to work conducted by the investigator
- Disclosure requirements that are not fully understood by the investigator
- Lack of communication between co-authors when preparing a manuscript for publication
- Absence of the conflict of interest at the time of publication

The current means of determining conflicts of interest are, to some extent, subjective and open to interpretation. This is likely to have contributed to observed

Textbox 31.3 An Ethical Dilemma and Research Sponsorship (Discussion Theme)

You are conducting a randomized controlled trial to assess the effectiveness of a new drug for the treatment of hypertension. The experimental group will receive the new drug while the intervention group will receive the current goal standard drug. The pharmaceutical company who developed the new drug is going to fund the research, and this sponsor presents you with a contract to sign. These contracts state that you must present all results to the company prior to publication. This contract also states that they may alter any results you find.

Discussion Points

1. What ethical issues does this contract present in terms of your own personal responsibilities as a scientific researcher?
2. How would you deal with this situation?
3. How could this contract affect your results?
4. What are the possible adverse consequences of signing such a contract?

discrepancies in reporting conflicts of interest (Okike et al. 2009). Reducing any uncertainty arising around conflicts of interest is crucial in order to reduce non-disclosure. Developing a deeper understanding of how researchers perceive conflicts and collaborations with industry can inform future efforts to establish a gold standard of ethical behavior (Ross et al. 2009). Such standards would, in turn, remove any uncertainty surrounding conflicts of interest.

31.7 Intellectual Property

Intellectual property (IP) is newly achieved knowledge that has been given specific property rights. The Universal Declaration of Human Rights defines an intellectual property right (IPR) as "the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author" (United Nations 1948; Barton et al. 2002). Put alternatively, IPR give ownership or temporary exclusivity of an idea or innovation in exchange for public disclosure of newly achieved knowledge. This section describes the various forms of IP, raises ethical issues created by IPR, and suggests approaches to dealing with those issues when publishing articles on related material.

31.7.1 Forms of Intellectual Property

There are several forms of IP that differ in their respective rights, entitlements, and embodiments. These include copyrights, patents, database protection, trade secrets, and trademarks. The latter two have little relevance to epidemiology and are therefore not discussed here.

31.7.1.1 Copyrights

A copyright is a form of legal protection that gives exclusive rights to the author(s) of an original piece of written work. The copyright holder therefore controls the reproduction and distribution of their work. Copyrights allow authors to further reproduce their own work, create further derivative work, and to transfer their work to others (Horner and Minifie 2011a). An article is the IP of the author(s) or the author(s) assignees (e.g., a journal that holds an article's copyright) and not of those who sponsored the study, unless contractual obligations specify otherwise.

31.7.1.2 Patents

Patents are a commonly used form of IP and the centrepiece of much controversy. In some academic circles, patents are cast in a negative light and, from one perspective, are viewed as mechanisms that delay or preclude dissemination of important knowledge and that reduce access to healthcare innovations, especially in developing countries. From another perspective, patents are viewed as engines of innovation, the absence of which might interfere with the advancement of technology and the public dissemination of knowledge enabling that technology. Regardless of which perspective one might have, patents raise important ethical concerns.

Patents allow a technology's inventors exclusive rights over the production and use of a described invention for a period of time. Patent laws and application processes vary considerably in different countries or international entities (e.g., the European Union), and patents issued in one country may not be enforceable in others. A general description of a patent application process is described below to delineate the point at which IP disclosure is necessary when publishing one's work; this discussion should not be construed as legal advice.

If an investigator is contemplating pursuit of IP, it is advisable first to inform and to consult with technology transfer officials at each of the institutions at which an invention was discovered, as each may have unique policies in place and can provide further guidance and support. This process is known as *disclosure*. The inventors may then file a *provisional* patent application to secure provisional IP protection until the appropriate authorities render a decision on the final application or, if a final application is not filed, until the deadline for submission of the final application passes. Disclosure of an invention to an institution does not necessitate revealing that action in publications. However, having a provisional or final patent application open with one or more regulatory authorities generally confers provisional IP protection; in either scenario, at minimum all relevant inventors should be identified, and the status of the invention should be described as *patent pending* to enable the reader to assess the author's objectivity and conflict of interest (e.g., due to potential financial gain). Neither the patent application itself nor its reference number needs to be disclosed. Having an issued patent, if it is germane to the subject of a publication, requires disclosure of that fact at least through the life of the patent. Sometimes sponsors of an investigation hold IP on the topic of an original article (e.g., a pharmaceutical company that sponsors a clinical trial on its drug), a situation that is usually not reported but inferred by way of disclosing sponsors with a financial conflict of interest.

31.7.1.3 Database Protections

In some countries databases are protected by copyright, but in others separate legislation provides special IPR to database owners. These IPRs are referred to as *database protection*. Laws that establish database protections have been the subjects of controversy in part because they enable database owners from making scientific databases inaccessible to other researchers indefinitely or for a specified period of time. Temporary database restriction allows the original investigators the opportunity to produce the earliest publications, but some databases remain proprietary indefinitely (i.e., closed-access), such as clinical trial databases produced by pharmaceutical companies. Completely closed-access databases have dubious ethics because they allow investigators to conceal any flaws in their statistics or methodology, and they preclude maximal utilization of data. For these reasons, we strongly favor an open access system, wherein databases are made available for use by other epidemiologists for academic purposes.

References

Barton J et al (2002) Integrating intellectual property rights and development policy: commission on intellectual property rights. http://www.iprcommission.org/papers/pdfs/final_report/ciprfullfinal.pdf. Accessed Feb 2013

Begg CB, Berlin JA (1989) Publication bias and dissemination of clinical research. J Natl Cancer Inst 81:107–115

Benos DJ et al (2005) Ethics and scientific publication. Adv Physiol Educ 29:59–74

Campbell E, Blumenthal D (2002) The selfish gene: data sharing and withholding in academic genetics. Science Career Magazine, 31st May 2002

Campbell EG et al (2002) Data withholding in academic genetics: evidence from a national survey. JAMA 287:473–480

Campbell EG et al (2007) A national survey of physician-industry relationships. N Engl J Med 356:1742–1750

Chalmers I (1990) Underreporting research is scientific misconduct. JAMA 263:1405–1408

Davidoff F et al (2001) (Commentary) sponsorship, authorship, and accountability. Lancet 358:854–856

Dickersin K, Rennie D (2003) Registering clinical trials. JAMA 290:516–523

Dickersin K et al (2002) Problems with indexing and citation of articles with group authorship. JAMA 287:2772–2774

Easterbrook PJ et al (1991) Publication bias in clinical research. Lancet 337:867–872

Egger M, Smith DG (1998) Meta-analysis bias in location and selection of studies. BMJ 316:61–66

Feeser VR, Simon JR (2008) The ethical assignment of authorship in scientific publications: issues and guidelines. Acad Emerg Med 15:963–969

Flanagin A, Fontanarosa PB, DeAngelis CD (2002) Authorship for research groups. JAMA 288:3166–3168

Haines IE, Olver IN (2008) Are self-regulation and declaration of conflict of interest still the benchmark for relationships between physicians and industry? Med J Aust 89:263–266

Horner J, Minifie FD (2011a) Research ethics II: mentoring, collaboration, peer review, and data management and ownership. J Speech Lang Hear R 54:S330–S345

Horner J, Minifie FD (2011b) Research ethics III: publication practices and authorship, conflicts of interest, and research misconduct. J Speech Lang Hear R 54:S346–S362

Hrynaszkiewicz I et al (2010) Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. Trials 11:9

Huston P, Moher D (1996) Redundancy, disaggregation, and the integrity of medical research. Lancet 347:1024–1026

ICMJE (2008) Uniform requirements for manuscripts submitted to biomedical journals: ethical considerations in the conduct and reporting of research: authorship and contributorship. http://www.icmje.org/ethical_1author.html. Accessed Feb 2013

JAMA (2006) Instructions for authors. JAMA 295:103–111

Kramer BS et al (2006) Getting it right: being smarter about clinical trials. PLoS Med 3(6):e144

Krimsky S, Rothenberg LS (1998) Financial interest and its disclosure in scientific publications. JAMA 280:225–226

Laflin LT, Glover ED, McDermott RJ (2005) Publication ethics: an examination of authorship practices. Am J Health Behav 29:579–587

Montori VM et al (2000) Publication bias: a brief review for clinicians. Mayo Clin Proc 75:1284–1288

Morin K et al (2002) Managing conflicts of interest in the conduct of clinical trials. JAMA 287:78–84

MRC (2005) MRC ethics series: good research practice. http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC002415. Accessed Feb 2013

Okike K et al (2009) Accuracy of conflict-of-interest disclosures reported by physicians. N Engl J Med 361:1466–1474

Partridge AH, Winer EP (2002) Informing clinical trial participants about study results. JAMA 288:363–365

Perlis RH et al (2005) Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry. Am J Psychiatry 162:1957–1960

Ross JS, Keyhani S, Korenstein D (2009) Appropriateness of collaborations between industry and the medical profession: physicians' perceptions. Am J Med 122:955–960

Scholey JM, Harrison J (2003) Publication bias: raising awareness of a potential problem in dental research. Br Dent J 194:235–237

Sterne JAC, Egger M, Davey-Smith G (2001) Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. BMJ 323:101–105

Tekola F et al (2009) Impact of social stigma on the process of obtaining informed consent for genetic research on podoconiosis: a qualitative study. BMC Med Ethics 10:13–22

Teo KK, Yusuf S, Furberg CD (1993) Effects of prophylactic antiarrhythmic drug therapy in acute myocardial infarction: an overview of results from randomized controlled trials. JAMA 270:1589–1595

Thompson DF (1993) Understanding financial conflicts of interest. N Engl J Med 329:573–576

Tumber MB, Dickersin K (2004) Publication of clinical trials: accountability and accessibility. J Int Med 256:271–283

United Nations (1948) The universal declaration of human rights. http://www.un.org/en/documents/udhr/index.shtml. Accessed Feb 2013

Vickers A et al (1998) Do certain countries produce only positive results? A systematic review of controlled trials. Control Clin Trials 19:159–166

Index

J. Van den Broeck and J.R. Brestoff (eds.), *Epidemiology: Principles and Practical Guidelines*, DOI 10.1007/978-94-007-5989-3,