Designing Clinical Research

FOURTH EDITION

Designing Clinical Research

FOURTH EDITION

Stephen B. Hulley, MD, MPH
Professor and Chair, Emeritus
Department of Epidemiology & Biostatistics
School of Medicine, University of California, San Francisco

Steven R. Cummings, MD
Founding Director, San Francisco Coordinating Center
Senior Scientist, California Pacific Medical Center Research Institute
Professor Emeritus, Department of Medicine, and of Epidemiology & Biostatistics
School of Medicine, University of California, San Francisco

Warren S. Browner, MD, MPH
Chief Executive Officer, California Pacific Medical Center
Adjunct Professor, Department of Epidemiology & Biostatistics
School of Medicine, University of California, San Francisco

Deborah G. Grady, MD, MPH
Professor of Medicine
Associate Dean for Clinical and Translational Research
School of Medicine, University of California, San Francisco

Thomas B. Newman, MD, MPH
Professor of Epidemiology & Biostatistics, and of Pediatrics
Chief, Division of Clinical Epidemiology
Attending Physician, Department of Pediatrics
School of Medicine, University of California, San Francisco

Wolters Kluwer Health | Lippincott Williams & Wilkins
Philadelphia • Baltimore • New York • London
Buenos Aires • Hong Kong • Sydney • Tokyo

Acquisitions Editor: Rebecca Gaertner
Product Manager: Tom Gibbons
Production Project Manager: David Orzechowski
Senior Manufacturing Coordinator: Beth Welsh
Marketing Manager: Kimberly Schonberger
Design Coordinator: Teresa Mallon
Production Service: S4Carlisle

Two Commerce Square
2001 Market Street
Philadelphia, PA 19103 USA
LWW.com

Printed in China

Library of Congress Cataloging-in-Publication Data

Designing clinical research / Stephen B Hulley . . . [et al.]. — 4th ed.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-1-60831-804-9 (pbk.)
I. Hulley, Stephen B.
[DNLM: 1. Epidemiologic Methods. 2. Research Design. WA 950]
R853.C55
610.72—dc23

2013009915

DISCLAIMER

Care has been taken to confirm the accuracy of the information presented and to describe generally accepted practices. However, the authors, editors, and publisher are not responsible for errors or omissions or for any consequences from application of the information in this book and make no warranty, expressed or implied, with respect to the currency, completeness, or accuracy of the contents of the publication. Application of the information in a particular situation remains the professional responsibility of the practitioner.

The authors, editors, and publisher have exerted every effort to ensure that drug selection and dosage set forth in this text are in accordance with current recommendations and practice at the time of publication. However, in view of ongoing research, changes in government regulations, and the constant flow of information relating to drug therapy and drug reactions, the reader is urged to check the package insert for each drug for any change in indications and dosage and for added warnings and precautions. This is particularly important when the recommended agent is a new or infrequently employed drug.

Some drugs and medical devices presented in the publication have Food and Drug Administration (FDA) clearance for limited use in restricted research settings. It is the responsibility of the health care provider to ascertain the FDA status of each drug or device planned for use in their clinical practice.

To purchase additional copies of this book, call our customer service department at (800) 638-3030 or fax orders to (301) 223-2320. International customers should call (301) 223-2300.

Visit Lippincott Williams & Wilkins on the Internet at LWW.com. Lippincott Williams & Wilkins customer service representatives are available from 8:30 AM to 6 PM, EST.

10 9 8 7 6 5 4 3 2 1

To our families and our students

Contents

Contributing Authors

Norman Hearst, MD, MPH
Professor of Family and Community Medicine
School of Medicine, University of California, San Francisco
Attending Physician, University of California Medical Center
San Francisco, California

Michael A. Kohn, MD, MPP
Associate Professor of Epidemiology and Biostatistics
School of Medicine, University of California, San Francisco
Attending Physician, Emergency Department
Mills-Peninsula Medical Center, Burlingame, California

Bernard Lo, MD
President, The Greenwall Foundation
Professor of Medicine, Emeritus
Director of Program in Medical Ethics, Emeritus
University of California, San Francisco

Thomas Edward Novotny, MD, MPH
Professor and Associate Director for Border and Global Health
Graduate School of Public Health
San Diego State University, San Diego, California

Introduction

This fourth edition of *Designing Clinical Research* (DCR) marks the 25th anniversary of the publication of our first edition. It has become the most widely used textbook of its kind, with more than 130,000 copies sold and foreign language editions produced in Spanish, Portuguese, Arabic, Chinese, Korean, and Japanese. We designed it as a manual for clinical research in all its flavors: clinical trials, observational epidemiology, translational science, patient-oriented research, behavioral science, and health services research. We used epidemiologic terms and principles, presented advanced conceptual material in a practical and reader-friendly way, and suggested common sense approaches to the many judgments involved in designing a study.

Many of our readers are physicians, nurses, pharmacists, and other health scientists who, as trainees and junior faculty, are developing careers in clinical research and use this book as a guide in designing and carrying out their studies. Many others are clinicians in residency programs and pre-doctoral students in professional schools—medicine, nursing, pharmacy, and public health among others—who use DCR to help them become discerning readers with a grasp of the strengths and limitations of the research studies that inform evidence-based clinical practice. A third audience consists of undergraduate students preparing to apply to these schools who are interested in looking ahead at the world of clinical research.

What's new in the fourth edition? The most visible innovation is color, which, in addition to improving the esthetics, will speed comprehension of the color-coded components. A larger innovation that accompanies each purchase of the paperback text is an interactive digital experience powered by Inkling®, viewable through a browser or as a download to tablet or smartphone. Its features include rapid index-based search options that link to a newly created glossary; bookmarking, highlighting, and annotating capability; cross-linking of relevant content; the ability to cut-and-paste figures or text into PowerPoint presentations; and live Internet links to jump instantly from citations to articles on PubMed, and to Google topics.

The substantive revisions to the fourth edition include updated and tightened text, figures, and tables in every chapter; many new examples and references; and new sections covering recent advances in the field. For example:

- The chapters on observational studies have been reorganized with an entire chapter now devoted to various case–control designs, including the incidence-density approach for addressing changes in risk factor levels and differences in follow-up time.
- The chapters on clinical trials have an expanded section on the non-inferiority trials that have become popular in comparative effectiveness research, and they address subgroup analysis and effect modification more fully.
- The chapter on studying medical tests has a new section on the growing practice of developing clinical prediction rules.
- The chapter on utilizing existing data sets emphasizes attractive options for beginning investigators to publish rapidly and inexpensively.
- The chapter on research ethics is updated to reflect current policy on whole genome sequencing and other topics, with new cases that illustrate the resolution of ethical dilemmas in clinical research.

- The chapter on data management has been extensively updated with the latest Web-based approaches.
- The chapter on getting funded has strategies for addressing the new NIH grant-writing requirements, as well as updates on funding by foundation and corporate sponsors.

The fourth edition is accompanied by an upgraded DCR website at www.epibiostat.ucsf.edu/dcr/ that contains materials for teaching DCR, including links to a detailed syllabus for the 4- and 7-week DCR workshops that we present to 300 trainees each year at UCSF. There are also instructor's notes for the workshops that faculty who teach this material will find useful, and links to our Training In Clinical Research (TICR) master's degree program at UCSF, with more than 30 other courses and their materials. In addition, there are useful tools for investigators, including an excellent interactive sample size calculator.

Many things have *not* changed in the fourth edition. It is still a simple book that leaves out unnecessary technicalities and invites the investigator to focus on the important things: how to find a good research question and how to plan an efficient, effective, ethical design. The chapters on estimating sample size continue to demystify the process and enable readers with minimal training in statistics to make these calculations themselves, thoughtfully, and without needing to wrestle with formulas. The book still works best when combined with the essential ingredient of one or more long-term mentors. It still *does not* address the important areas of how to analyze, present, and publish the findings of clinical research—topics that our readers can pursue with other books (e.g., 1–4). And we still *do* use the feminine pronoun in the first half of the book, masculine in the second, the goal (besides avoiding the passive tense) being to symbolically empower clinical investigators of both genders.

The process of becoming an independent clinical scientist can be challenging, especially getting over the hump of acquiring a substantial grant for the first time. But it is gratifying that many of our former trainees who used this book have achieved this goal, discovered that they *like* doing research, and settled into a great career. For those with inquiring minds, the pursuit of truth can become a lifelong fascination. For perfectionists and craftsmen, there are endless challenges in creating elegant studies that conclusively answer questions, large and small, at an affordable cost in time and money. Investigators who enjoy teamwork will develop rewarding relationships with colleagues, staff, and students, as well as friendships with collaborators working in the same field in distant places. And for those with the ambition to make a lasting contribution to society, there is the prospect that with skill and tenacity they will participate in the incremental advances in clinical and public health practice that is the natural order of our science.

REFERENCES

1. Vittinghoff E, Glidden DV, Shiboski SC, et al. *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*, 2nd ed. New York: Springer-Verlag, 2011.
2. Katz MH. *Multivariable analysis: a practical guide for clinicians and public health researchers*, 3rd ed. New York: Cambridge University Press, 2011.
3. Newman TB, Kohn MA. *Evidence-based diagnosis*. Cambridge, MA: Cambridge University Press, 2009.
4. Browner WS. *Publishing and presenting clinical research*, 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2012.

Acknowledgments

We are grateful to the Andrew W. Mellon Foundation for bringing the five of us together 30 years ago to begin the five-year journey of developing the teaching materials that became the first edition; to our publisher for steadily inviting a fourth edition until resistance became futile, and for providing exceptionally talented and supportive professionals to help us put it together; to our families for their patient support as we labored over this opus; to many colleagues at UCSF and beyond, whose ideas have influenced ours; to our students over the years, whose accomplishments have been fun to watch and stimulating to our thinking; and to our readers who have put this book to use.

SECTION I

Basic Ingredients

CHAPTER 1

Getting Started: The Anatomy and Physiology of Clinical Research

Stephen B. Hulley, Thomas B. Newman, and Steven R. Cummings

This chapter introduces clinical research from two viewpoints, setting up themes that run together throughout the book. One is the **anatomy** of research—what it's made of. This includes the tangible elements of the study plan: research question, design, subjects, measurements, sample size calculation, and so forth. An investigator's goal is to design these components in a fashion that will make the project feasible and efficient.

The other theme is the **physiology** of research—how it works. Studies are useful to the extent that they yield valid inferences, first about what happened in the study sample and then about how these findings generalize to people outside the study. The goal is to minimize the errors, random and systematic, that threaten conclusions based on these inferences.

Separating the two themes is artificial in the same way that the anatomy of the human body doesn't make much sense without some understanding of its physiology. But the separation has the same advantage: It clarifies our thinking about a complex topic.

ANATOMY OF RESEARCH: WHAT IT'S MADE OF

The structure of a research project is set out in its **protocol**, the written plan of the study. Protocols are well known as devices for seeking grant funds and Institutional Review Board (IRB) approval, but they also have a vital scientific function: helping the investigator organize her research in a logical, focused, and efficient way. Table 1.1 outlines the components of a protocol. We introduce the whole set here, expand on each component in the ensuing chapters of the book, and return to put the completed pieces together in Chapter 19.

Research Question

The **research question** is the objective of the study, the uncertainty the investigator wants to resolve. Research questions often begin with a general concern that must be narrowed down to a concrete, researchable issue. Consider, for example, the general question:

- Should people eat more fish?

This is a good place to start, but the question must be focused before planning efforts can begin. Often this involves breaking the question into more specific components, and singling out one or two of these to build the protocol around:

- How often do Americans eat fish?
- Does eating more fish lower the risk of cardiovascular disease?
- Is there a risk of mercury toxicity from increasing fish intake in older adults?
- Do fish oil supplements have the same effects on cardiovascular disease as dietary fish?
- Which fish oil supplements don't make your breath smell like fish?

TABLE 1.1 ANATOMY OF RESEARCH: THE STUDY PLAN

DESIGN COMPONENTS	PURPOSE
Research questions	What questions will the study address?
Background and significance	Why are these questions important?
Design	How is the study structured?
Time frame	
Epidemiologic design	
Subjects	Who are the subjects and how will they be selected?
Selection criteria	
Sampling design	
Variables	What measurements will be made?
Predictor variables	
Confounding variables	
Outcome variables	
Statistical issues	How large is the study and how will it be analyzed?
Hypotheses	
Sample size	
Analytic approach	

A good research question should pass the "So what?" test. Getting the answer should contribute usefully to our state of knowledge. The acronym **FINER** denotes five essential characteristics of a good research question: It should be **feasible**, **interesting**, **novel**, **ethical**, and **relevant** (Chapter 2).

Background and Significance

A brief **background** and **significance** section in a protocol sets the proposed study in context and gives its rationale: What is known about the topic at hand? Why is the research question important? What kind of answers will the study provide? This section cites relevant previous research (including the investigator's own work) and indicates the problems with the prior research and what uncertainties remain. It specifies how the findings of the proposed study will help resolve these uncertainties, lead to new scientific knowledge, or influence practice guidelines or public health policy. Often, the literature review and synthesis done for the significance section will lead the investigator to modify the research question.

Design

The **design** of a study is a complex issue. A fundamental decision is whether to take a passive role in making measurements on the study subjects in an **observational study** or to apply an intervention and examine its effects in a **clinical trial** (Table 1.2). Among observational studies, two common designs are **cohort studies**, in which observations are made in a group of subjects that is followed over time, and **cross-sectional studies**, in which observations are made on a single occasion. Cohort studies can be further divided into **prospective** studies that begin in the present and follow subjects into the future, and **retrospective** studies that examine information collected over a period of time in the past. A third common option is the **case–control** design, in which the investigator compares a group of people who have a disease or other outcome with another group who do not. Among clinical trial options, the **randomized blinded trial** is

TABLE 1.2 EXAMPLES OF CLINICAL RESEARCH DESIGNS TO FIND OUT WHETHER FISH INTAKE REDUCES CORONARY HEART DISEASE RISK

EPIDEMIOLOGIC DESIGN	KEY FEATURE	EXAMPLE
	Observational Designs	
Cohort study	A group of subjects identified at the beginning and followed over time	The investigator measures fish intake in a group of subjects at baseline and periodically examines them at follow-up visits to see if those who eat more fish have fewer coronary heart disease (CHD) events.
Cross-sectional study	A group examined at one point in time	She interviews a group of subjects about current and past history of fish intake and correlates results with history of CHD and current coronary calcium score.
Case–control study	Two groups selected based on the presence or absence of an outcome	She examines a group of patients with CHD (the "cases") and compares them with a group who do not have CHD (the "controls"), asking about past fish intake.
	Clinical Trial Design	
Randomized blinded trial	Two groups created by a random process, and a blinded intervention	She randomly assigns subjects to receive fish oil supplements or a placebo that is identical in appearance, then follows both treatment groups for several years to observe the incidence of CHD.

usually the best design but nonrandomized or unblinded designs may be all that are feasible for some research questions.

No one approach is always better than the others, and each research question requires a judgment about which design is the most efficient way to get a satisfactory answer. The randomized blinded trial is often held up as the best design for establishing causality and the effectiveness of interventions, but there are many situations for which an observational study is a better choice or the only feasible option. The relatively low cost of case–control studies and their suitability for rare outcomes makes them attractive for some questions. Special considerations apply to choosing designs for studying diagnostic tests. These issues are discussed in Chapters 7 through 12, each dealing with a particular set of designs.

A typical sequence for studying a topic begins with observational studies of a type that is often called **descriptive**. These studies explore the lay of the land—for example, describing distributions of health-related characteristics and diseases in the population:

- What is the average number of servings of fish per week in the diet of Americans with a history of coronary heart disease (CHD)?

Descriptive studies are usually followed or accompanied by **analytic** studies that evaluate associations to permit inferences about cause-and-effect relationships:

- Do people with a CHD who eat a lot of fish have a lower risk of recurrent myocardial infarction than people with a history of CHD who rarely eat fish?

The final step is often a **clinical trial** to establish the effects of an intervention:

- Does treatment with fish oil capsules reduce total mortality in people with CHD?

Clinical trials usually occur relatively late in a series of research studies about a given question, because they tend to be more difficult and expensive, and to answer more definitively the narrowly focused questions that arise from the findings of observational studies.

It is useful to characterize a study in a *single sentence that summarizes the design and research question*. If the study has two major phases, the design for each should be mentioned.

- This is a cross-sectional study of dietary habits in 50- to 69-year-old people with a history of CHD, followed by a prospective cohort study of whether fish intake is associated with lower risk of subsequent coronary events.

This sentence is the research analog to the opening sentence of a medical resident's report on a new hospital admission: "This 62-year-old white policewoman was well until 2 hours before admission, when she developed crushing chest pain radiating to the left shoulder."

Some designs do not easily fit into the categories listed above, and classifying them with a single sentence can be surprisingly difficult. It is worth the effort—a concise description of the design and research question clarifies the investigator's thoughts and is useful for orienting colleagues and consultants.

Study Subjects

Two major decisions must be made in choosing the study subjects (Chapter 3). The first is to specify **inclusion** and **exclusion criteria** that define the target population: the *kinds* of people best suited to the research question. The second decision concerns how to **recruit** an appropriate *number* of people from an accessible subset of this population to be the subjects of the study. For example, the study of fish intake in people with CHD might identify subjects seen in the clinic with diagnostic codes for myocardial infarction, angioplasty, or coronary artery bypass grafting in their electronic medical record. Decisions about which patients to study often represent trade-offs; studying a random sample of people with CHD from the entire country (or at least several different states and medical care settings) would enhance **generalizability** but be much more difficult and costly.

Variables

Another major set of decisions in designing any study concerns the choice of which variables to measure (Chapter 4). A study of fish intake in the diet, for example, might ask about different types of fish that contain different levels of omega-3 fatty acids, and include questions about portion size, whether the fish was fried or baked, and use of fish oil supplements.

In an analytic study the investigator studies the associations among variables to predict outcomes and to draw inferences about cause and effect. In considering the association between two variables, the one that occurs first or is more likely on biologic grounds to be causal is called the **predictor variable**; the other is called the **outcome variable**.[1] Most observational studies have many predictor variables (age, race, sex, smoking history, fish and fish oil supplement intake) and several outcome variables (heart attacks, strokes, quality of life, unpleasant odor).

Clinical trials examine the effects of an **intervention**—a special kind of predictor variable that the investigator manipulates, such as treatment with fish oil capsules. This design allows her to observe the effects on the outcome variable using **randomization** to minimize the influence of **confounding variables**—other predictors of the outcome such as smoking or income level that could be associated with dietary fish and confuse the interpretation of the findings.

[1]Predictors are sometimes termed **independent variables** and outcomes **dependent variables**, but the meaning of these terms is less self-evident and we prefer to avoid their use.

Statistical Issues

The investigator must develop plans for estimating sample size and for managing and analyzing the study data. This generally involves specifying a **hypothesis** (Chapter 5).

> **Hypothesis:** 50- to 69-year-old women with CHD who take fish oil supplements will have a lower risk of recurrent myocardial infarction than those who do not.

This is a version of the research question that provides the basis for testing the **statistical significance** of the findings. The hypothesis also allows the investigator to calculate the **sample size**—the number of subjects needed to observe the expected difference in outcome between study groups with reasonable probability (an attribute known as **power**) (Chapter 6). Purely descriptive studies (what proportion of people with CHD use fish oil supplements?) do not involve tests of statistical significance, and thus do not require a hypothesis; instead, the number of subjects needed to produce acceptably narrow **confidence intervals** for means, proportions, or other descriptive statistics can be calculated.

PHYSIOLOGY OF RESEARCH: HOW IT WORKS

The goal of clinical research is to draw **inferences** from findings in the study about the nature of the universe around it. Two major sets of inferences are involved in interpreting a study (illustrated from right to left in Figure 1.1). Inference #1 concerns **internal validity**, the degree to which the investigator draws the correct conclusions about what actually happened in the study. Inference #2 concerns **external validity** (also called **generalizability**), the degree to which these conclusions can be appropriately applied to people and events outside the study.

When an investigator plans a study, she reverses the process, working from left to right in the lower half of Figure 1.1 with the goal of maximizing the validity of these inferences at the end of the study. She **designs a study plan** in which the choice of research question, subjects, and measurements enhances the external validity of the study and is conducive to **implementation** with a high degree of internal validity. In the next sections we address design and then implementation before turning to the errors that threaten the validity of these inferences.

Designing the Study

Consider this simple descriptive question:

> What is the prevalence of daily ingestion of fish oil supplements among people with CHD?

This question cannot be answered with perfect accuracy because it would be impossible to study all patients with CHD and our approaches to discovering whether a person has CHD

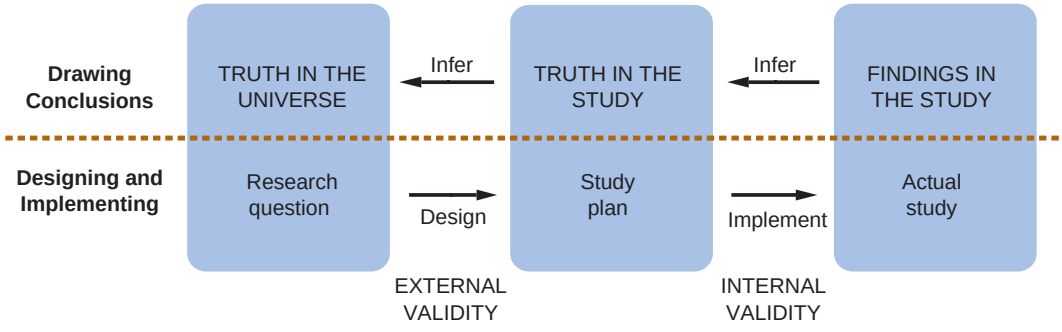


FIGURE 1.1 The process of designing and implementing a research project sets the stage for drawing conclusions based on inferences from the findings.

and is taking fish oil are imperfect. So the investigator settles for a related question that *can* be answered by the study:

> Among a sample of patients seen in the investigator's clinic who have a previous CHD diagnosis and respond to a mailed questionnaire, what proportion report taking daily fish oil supplements?

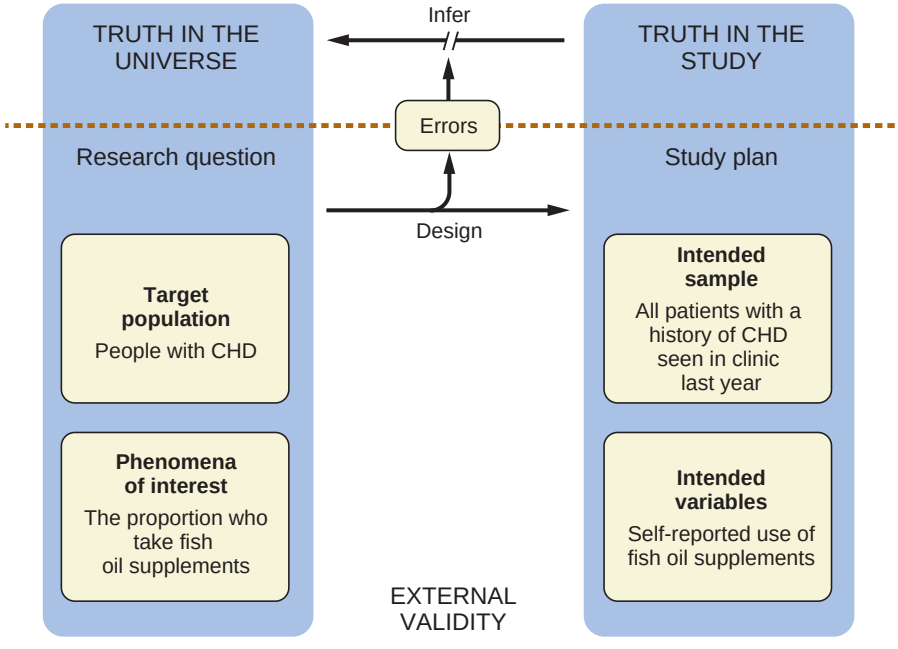
The transformation from research question to study plan is illustrated in Figure 1.2. One major component of this transformation is the choice of a **sample** of subjects that will represent the **population**. The group of subjects specified in the protocol can only be a sample of the population of interest because there are practical barriers to studying the entire population. The decision to study patients in the investigator's clinic identified through the electronic medical record system is a compromise. This is a sample that is feasible to study but has the disadvantage that it may produce a different prevalence of fish oil use than that found in all people with CHD.

The other major component of the transformation is the choice of **variables** that will represent the **phenomena of interest**. The variables specified in the study plan are usually proxies for these phenomena. The decision to use a self-report questionnaire to assess fish oil use is a fast and inexpensive way to collect information, but unlikely to be perfectly accurate because people usually do not accurately remember or record how much they take in a typical week.

In short, each of the differences in Figure 1.2 between the research question and the study plan has the purpose of making the study more practical. The cost of this increase in practicality, however, is the risk that design choices may cause the study to produce a wrong or misleading conclusion because it is designed to answer a somewhat different question from the research question of interest.

Implementing the Study

Returning to Figure 1.1, the right-hand side is concerned with **implementation** and the degree to which the actual study matches the study plan. At issue here is the problem of a wrong answer



■ **FIGURE 1.2** Design errors and external validity: If the intended sample and variables do not sufficiently represent the target population and phenomena of interest, these errors may distort inferences about what actually happens in the population.

to the research question because the way the sample was actually drawn, or the measurements made, differed in important ways from the way they were designed (Figure 1.3).

The actual sample of study subjects is almost always different from the intended sample. The plans to study all eligible clinic patients with CHD, for example, could be disrupted by incomplete diagnoses in the electronic medical record, wrong addresses for the mailed questionnaire, and refusal to participate. Those subjects who are reached and agree to participate may have a different prevalence of fish oil use than those not reached or not interested. In addition to these problems with the subjects, the actual measurements can differ from the intended measurements. If the format of the questionnaire is unclear subjects may get confused and check the wrong box, or they may simply omit the question by mistake.

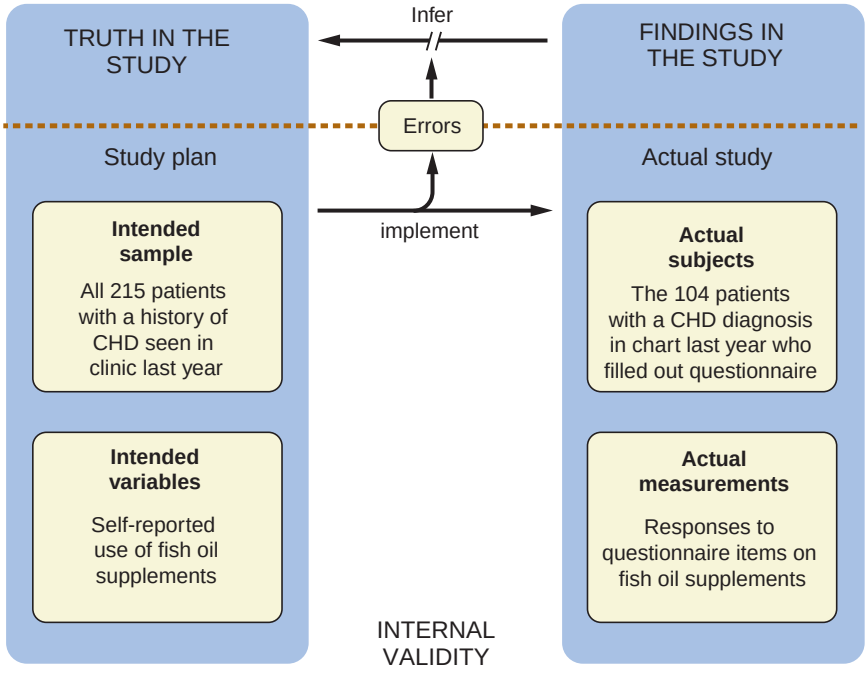
These differences between the study plan and the actual study can alter the answer to the research question. Figure 1.3 illustrates that errors in implementing the study join errors of design in leading to a misleading or wrong answer to the research question.

Causal Inference

A special kind of validity problem arises in studies that examine the **association** between a predictor and an outcome variable in order to draw causal inference. If a cohort study finds an association between fish intake and CHD events, does this represent cause and effect, or is fish intake just an innocent bystander in a web of causation that involves other variables? Reducing the likelihood of **confounding** and other rival explanations is one of the major challenges in designing an observational study (Chapter 9).

The Errors of Research

Recognizing that no study is entirely free of errors, the goal is to maximize the validity of inferences from what was observed in the study sample to what is happening in the population.



■ **FIGURE 1.3** Implementation errors and internal validity: If the actual subjects and measurements do not sufficiently represent the intended sample and variables, these errors may distort inferences about what happened in the study.

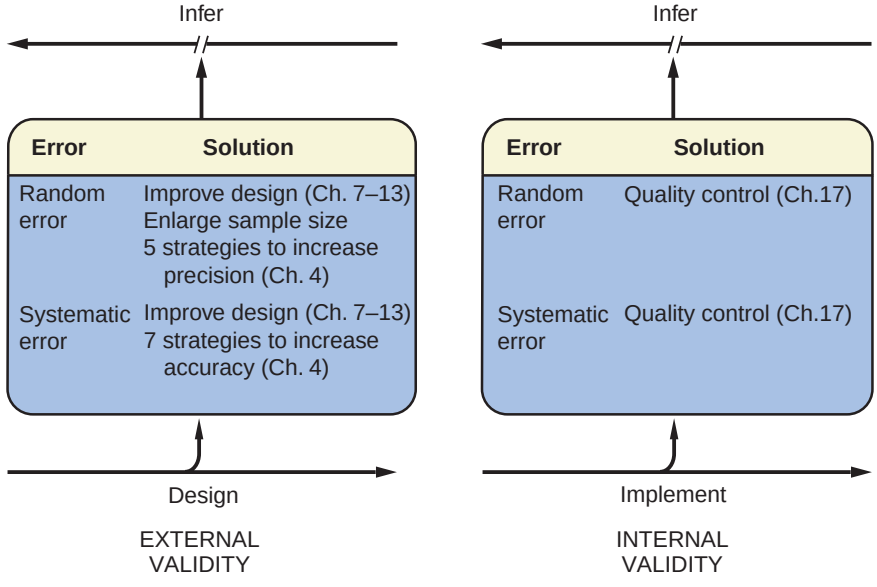
Erroneous inferences can be addressed in the analysis phase of research, but a better strategy is to focus on design and implementation (Figure 1.4), preventing errors from occurring in the first place to the extent that this is practical.

The two main kinds of errors that interfere with research inferences are random error and systematic error. The distinction is important because the strategies for minimizing them are quite different.

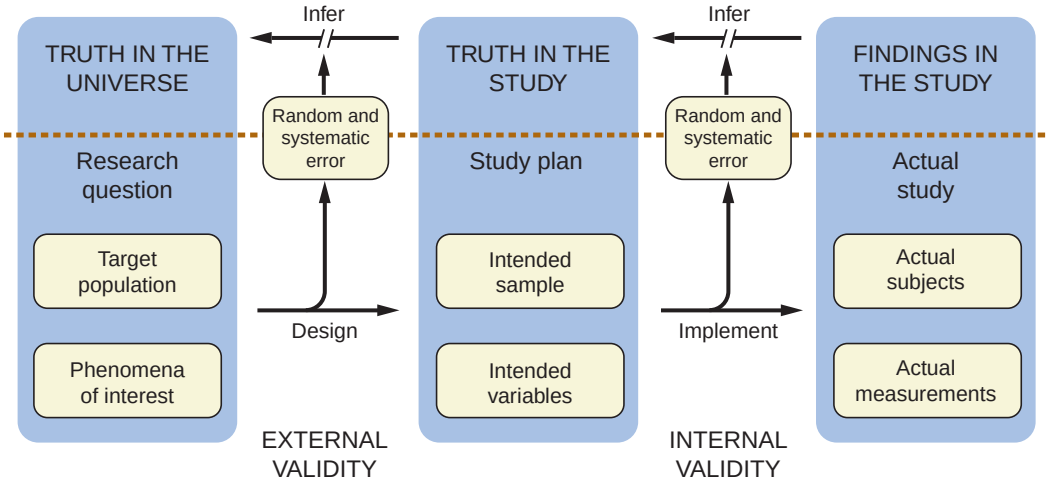
Random error is a wrong result due to **chance**—sources of variation that are equally likely to distort measurements from the study in either direction. If the true prevalence of daily fish oil supplement use in the several hundred 50- to 69-year-old patients with CHD in the investigator's clinic is 20%, a well-designed sample of 100 patients from that population might contain exactly 20 patients who use these supplements. More likely, however, the sample would contain a nearby number such as 18, 19, 21, or 22. Occasionally, chance would produce a substantially different number, such as 12 or 28. Among several techniques for reducing the influence of random error (Chapter 4), the simplest is to increase the sample size. The use of a larger sample diminishes the likelihood of a substantially wrong result by increasing the **precision** of the estimate—the degree to which the observed prevalence approximates 20% each time a sample is drawn.

Systematic error is a wrong result due to **bias**—sources of variation that distort the study findings in one direction. An illustration is the decision in Figure 1.2 to study patients in the investigator's clinic, where the local treatment patterns have responded to her interest in the topic and her fellow doctors are more likely than the average doctor to recommend fish oil. Increasing the sample size has no effect on systematic error. The best way to improve the **accuracy** of the estimate (the degree to which it approximates the true value) is to design the study in a way that reduces the size of the various biases. Alternatively, the investigator can seek additional information to assess the importance of possible biases. An example would be to compare results with those from a second sample of patients with CHD drawn from another setting, for example, examining whether the findings of such patients seen in a cardiology clinic are different from those seen in a primary care clinic.

The examples of random and systematic error in the preceding two paragraphs are components of **sampling error**, which threatens inferences from the study subjects to the population.



■ **FIGURE 1.4** Research errors. This blown-up detail of the error boxes in Figures 1.2 and 1.3 reveals strategies for controlling random and systematic error in the design and implementation phases of the study.



■ **FIGURE 1.5** Physiology of research—how it works.

Both random and systematic errors can also contribute to **measurement error**, threatening the inferences from the study measurements to the phenomena of interest. An illustration of random measurement error is the variation in the response when the diet questionnaire is administered to the patient on several occasions. An example of systematic measurement error is underestimation of the prevalence of fish oil use due to lack of clarity in how the question is phrased. Additional strategies for controlling all these sources of error are presented in Chapters 3 and 4.

The concepts presented in the last several pages are summarized in Figure 1.5 Getting the right answer to the research question is a matter of designing and implementing the study in a fashion that minimizes the magnitude of inferential errors.

■ DESIGNING THE STUDY

Study Plan

The process of developing the **study plan** begins with the one-sentence **research question** that specifies the main predictor and outcome variables and the population. Three versions of the study plan are then produced in sequence, each larger and more detailed than the preceding one.

- **Study outline** (Table 1.1 and Appendix 1). This one-page summary of the design serves as a standardized checklist to remind the investigator to address all the components. As important, the sequence has an orderly logic that helps clarify the investigator's thinking on the topic.
- **Study protocol.** This expansion on the study outline usually ranges from 5 to 15 pages, and is used to plan the study and to apply for IRB approval and grant support. The protocol parts are discussed throughout this book and summarized in Chapter 19.
- **Operations manual.** This collection of specific procedural instructions, questionnaires, and other materials is designed to ensure a uniform and standardized approach to carrying out the study with good quality control (Chapters 4 and 17).

The research question and study outline should be written out at an early stage. Putting thoughts down on paper leads the way from vague ideas to specific plans and provides a concrete basis for getting advice from colleagues and consultants. It is a challenge to do it (ideas are easier to talk about than to write down), but the rewards are a faster start and a better project.

Appendix 1 is an example of a study outline. This one-page outline deals more with the anatomy of research (Table 1.1) than with its physiology (Figure 1.5), so the investigator must remind herself to worry about the errors that may result when it is time to draw inferences

from measurements in the study sample to phenomena of interest in the population. A study's virtues and problems can be revealed by explicitly considering how the question the study is likely to answer differs from the research question, given the plans for acquiring subjects and making measurements, and given the likely problems of implementation.

With the study outline in hand and the intended inferences in mind, the investigator can proceed with the details of her protocol. This includes getting advice from colleagues, drafting specific recruitment and measurement methods, considering scientific and ethical appropriateness, modifying the study question and outline as needed, pretesting specific recruitment and measurement methods, making more changes, getting more advice, and so forth. This iterative process is the nature of research design and the topic of the rest of this book.

Trade-offs

Regretably, errors are an inherent part of all studies. The main issue is whether the errors will be large enough to change the conclusions in important ways. When designing a study, the investigator is in much the same position as a labor union official bargaining for a new contract. The union official begins with a wish list—shorter hours, more money, health care benefits, and so forth. She must then make concessions, holding on to the things that are most important and relinquishing those that are not essential or realistic. At the end of the negotiations is a vital step: She looks at the best contract she could negotiate and decides if it has become so bad that it is no longer worth having.

The same sort of concessions must be made by an investigator when she transforms the research question to the study plan and considers potential problems in implementation. On one side are the issues of internal and external validity; on the other, feasibility. The vital last step of the union negotiator is sometimes omitted. Once the study plan has been formulated, the investigator must decide whether it adequately addresses the research question and whether it can be implemented with acceptable levels of error. Often the answer is no, and there is a need to begin the process anew. But take heart! Good scientists distinguish themselves not so much by their uniformly good research ideas as by their alacrity in turning over those that won't work and moving on to better ones.

SUMMARY

1. The **anatomy** of research is the set of tangible elements that make up the study plan: the **research question** and its **significance**, and the **design**, **study subjects**, and **measurement approaches**. The challenge is to design elements that are relatively **inexpensive** and **easy** to implement.
2. The **physiology** of research is how the study works. The study findings are used to draw **inferences** about what happened in the study sample (**internal validity**), and about events in the world outside (**external validity**). The challenge here is to **design** and **implement** a study plan with adequate control over two major threats to these inferences: **random error** (chance) and **systematic error** (bias).
3. In designing a study the investigator may find it helpful to consider Figure 1.5, the relationships between the **research question** (what she wants to answer), the **study plan** (what the study is designed to answer), and the **actual study** (what the study will actually answer, given the errors of implementation that can be anticipated).
4. A good way to develop the **study plan** is to begin with a one-sentence version of the **research question** that specifies the main variables and population, and expand this into a one-page **outline** that sets out the study elements in a standardized sequence. Later on the study plan will be expanded into the **protocol** and the **operations manual**.
5. Good **judgment** by the investigator and **advice** from colleagues are needed for the many **trade-offs** involved, and for determining the overall viability of the project.

APPENDIX 1
Outline of a Study

This is the one-page study plan of a project carried out by Valerie Flaherman, MD, MPH, begun while she was a general pediatrics fellow at UCSF. Most beginning investigators find observational studies easier to pull off, but in this case a randomized clinical trial of modest size and scope was feasible, the only design that could adequately address the research question, and ultimately successful—see publication by Flaherman et al (1) for the findings, which, if confirmed, could alter policy on how best to initiate breast feeding.

TITLE: EFFECT OF EARLY LIMITED FORMULA USE ON BREASTFEEDING

Research question:

Among term newborns who have lost ≥ 5% of their birth weight before 36 hours of age, does feeding 10 cc of formula by syringe after each breastfeeding before the onset of mature milk production increase the likelihood of subsequent successful breastfeeding?

Significance:

1. Breast milk volume is low until mature milk production begins 2–5 days after birth.
2. Some mothers become worried if the onset of mature milk production is late and their baby loses a lot of weight, leading them to abandon breastfeeding within the first week. A strategy that increased the proportion of mothers who succeed in breastfeeding would have many health and psycho-social benefits to mother and child.
3. Observational studies have found that formula feeding in the first few days after birth is associated with decreased breastfeeding duration. Although this could be due to confounding by indication (see Chapter 9), the finding has led to WHO and CDC guidelines aimed at reducing the use of formula during the birth hospitalization.
4. However, a small amount of formula combined with breastfeeding and counseling might make the early breastfeeding experience more positive and increase the likelihood of success. A clinical trial is needed to assess possible benefits and harms of this strategy.

Study design:

Unblinded randomized control trial with blinded outcome ascertainment

Subjects:

- **Entry criteria:** Healthy term newborns 24–48 hours old who have lost ≥ 5% of their birth weight in the first 36 hours after birth
- **Sampling design:** Consecutive sample of consenting patients in two Northern California academic medical centers

Predictor variable, randomly assigned but not blinded:

- **Control:** Parents are taught infant soothing techniques.
- **Intervention:** Parents are taught to syringe-feed 10 cc of formula after each breastfeeding until the onset of mature milk production.

Outcome variables, blindly ascertained:

1. Any formula feeding at 1 week and 1, 2, and 3 months
2. Any breastfeeding at 1 week and 1, 2, and 3 months
3. Weight nadir

Primary null hypothesis:

Early limited formula does not affect the proportion of women who are breastfeeding their baby at 3 months.

REFERENCE

1. Flaherman VJ, Aby J, Burgos AE, et al. Effect of early limited formula on duration and exclusivity of breastfeeding in at-risk infants: an RCT. *Pediatrics*, in press.

CHAPTER 2

Conceiving the Research Question and Developing the Study Plan

Steven R. Cummings, Warren S. Browner, and Stephen B. Hulley

The **research question** is the uncertainty that the investigator wants to resolve by performing her study. There is no shortage of good research questions, and even as we succeed in answering some questions, we remain surrounded by others. Clinical trials, for example, established that treatments that block the synthesis of estradiol (aromatase inhibitors) reduce the risk of breast cancer in women who have had early stage cancer (1). But this led to new questions: How long should treatment be continued; does this treatment prevent breast cancer in patients with BRCA 1 and BRCA 2 mutations; and what is the best way to prevent the osteoporosis that is an adverse effect of these drugs? Beyond that are primary prevention questions: Are these treatments effective and safe for preventing breast cancer in healthy women?

The challenge in finding a research question is defining an important one that can be transformed into a feasible and valid **study plan**. This chapter presents strategies for accomplishing this (Figure 2.1).

ORIGINS OF A RESEARCH QUESTION

For an established investigator the best research questions usually emerge from the findings and problems she has observed in her own **prior studies** and in those of other workers in the field. A new investigator has not yet developed this base of experience. Although a fresh perspective is sometimes useful by allowing a creative person to conceive new approaches to old problems, lack of experience is largely an impediment.

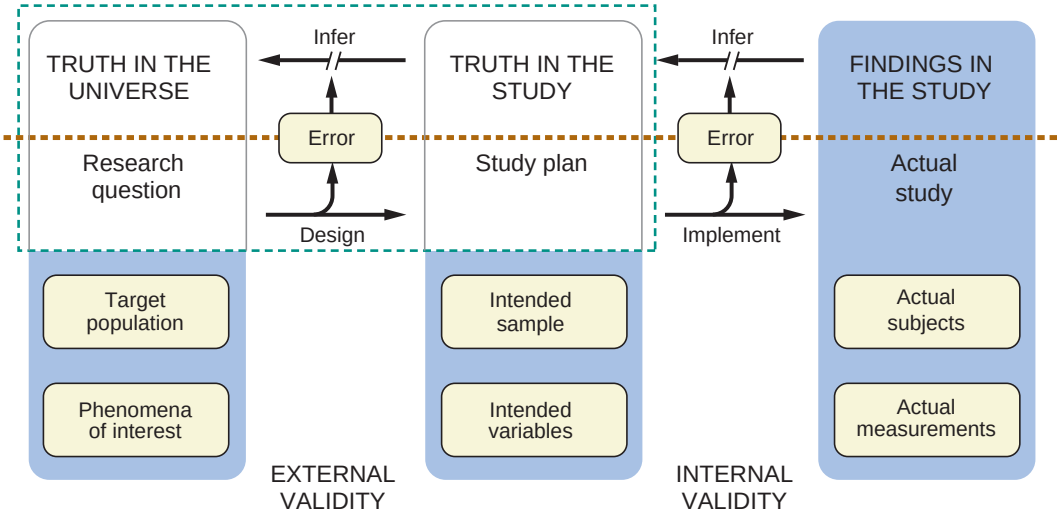


FIGURE 2.1 This chapter focuses on the area within the dashed green line, the challenge of choosing a research question that is of interest and can be tackled with a feasible study plan.

A good way to begin is to clarify the difference between a **research question** and a **research interest**. Consider this research question:

- Does participation in group counseling sessions reduce the likelihood of domestic violence among women who have recently immigrated from Central America?

This might be asked by someone whose research interest involves the efficacy of group counseling, or the prevention of domestic violence, or improving health in recent immigrants. The distinction between research questions and research interests matters because it may turn out that the specific research question cannot be transformed into a viable study plan, but the investigator can still address her research interest by asking a different question.

Of course, it's impossible to formulate a research question if you are not even sure about your research interest (beyond knowing that you're supposed to have one). If you find yourself in this boat, you're not alone: Many new investigators have not yet discovered a topic that interests them and is susceptible to a study plan they can design. You can begin by considering what sorts of research studies have piqued your interest when you've seen them in a journal. Or perhaps you were bothered by a specific patient whose treatment seemed inadequate or inappropriate: What could have been done differently that might have improved her outcome? Or one of your attending physicians told you that hypokalemia always caused profound thirst, and another said the opposite, just as dogmatically.

Mastering the Literature

It is important to master the published literature in an area of study: **Scholarship** is a necessary precursor to good research. A new investigator should conduct a thorough search of published literature in the areas pertinent to the research question and critically read important original papers. Carrying out a **systematic review** is a great next step for developing and establishing expertise in a research area, and the underlying literature review can serve as background for grant proposals and research reports. Recent advances may be known to active investigators in a particular field long before they are published. Thus, mastery of a subject entails participating in meetings and building relationships with **experts** in the field.

Being Alert to New Ideas and Techniques

In addition to the medical literature as a source of ideas for research questions, it is helpful to attend **conferences** in which new work is presented. At least as important as the formal presentations are the opportunities for informal conversations with other scientists at posters and during the breaks. A new investigator who overcomes her shyness and engages a speaker at the coffee break may find the experience richly rewarding, and occasionally she will have a new senior colleague. Even better, for a speaker known in advance to be especially relevant, it may be worthwhile to look up her recent publications and contact her in advance to arrange a meeting during the conference.

A **skeptical attitude** about prevailing beliefs can stimulate good research questions. For example, it was widely believed that lacerations which extend through the dermis required sutures to assure rapid healing and a satisfactory cosmetic outcome. However, Quinn et al. noted personal experience and case series evidence that wounds of moderate size repair themselves regardless of whether wound edges are approximated (2). They carried out a randomized trial in which all patients with hand lacerations less than 2 cm in length received tap water irrigation and a 48-hour antibiotic dressing. One group was randomly assigned to have their wounds sutured, and the other group did not receive sutures. The suture group had a more painful and time-consuming treatment in the emergency room, but blinded assessment revealed similar time to healing and similar cosmetic results. This has now become a standard approach used in clinical practice.

The application of **new technologies** often generates new insights and questions about familiar clinical problems, which in turn can generate new paradigms (3). Advances in imaging and in molecular and genetic technologies, for example, have spawned translational research studies that have led to new treatments and tests that have changed clinical medicine. Similarly, taking a new concept, technology, or finding from one field and applying it to a problem in a different field can lead to good research questions. Low bone density, for example, is a risk factor for fractures. Investigators applied this technology to other outcomes and found that women with low bone density have higher rates of cognitive decline (4), stimulating research for factors, such as low endogenous levels of estrogen, that could lead to loss of both bone and memory.

Keeping the Imagination Roaming

Careful **observation** of patients has led to many descriptive studies and is a fruitful source of research questions. **Teaching** is also an excellent source of inspiration; ideas for studies often occur while preparing presentations or during discussions with inquisitive students. Because there is usually not enough time to develop these ideas on the spot, it is useful to keep them in a **computer file** or notebook for future reference.

There is a major role for **creativity** in the process of conceiving research questions, imagining new methods to address old questions, and playing with ideas. Some creative ideas come to mind during informal conversations with colleagues over lunch; others arise from discussing recent research or your own ideas in small groups. Many inspirations are solo affairs that strike while preparing a lecture, showering, perusing the Internet, or just sitting and thinking. Fear of criticism or seeming unusual can prematurely quash new ideas. The trick is to put an unresolved problem clearly in view and allow the mind to run freely around it. There is also a need for **tenacity**, returning to a troublesome problem repeatedly until a resolution is reached.

Choosing and Working with a Mentor

Nothing substitutes for experience in guiding the many judgments involved in conceiving a research question and fleshing out a study plan. Therefore an essential strategy for a new investigator is to apprentice herself to an experienced **mentor** who has the time and interest to work with her regularly.

A good mentor will be available for regular meetings and informal discussions, encourage creative ideas, provide wisdom that comes from experience, help ensure protected time for research, open doors to networking and funding opportunities, encourage the development of independent work, and put the new investigator's name first on grants and publications whenever appropriate. Sometimes it is desirable to have more than one mentor, representing different disciplines. Good relationships of this sort can also lead to tangible resources that are needed—office space, access to clinical populations, data sets and specimen banks, specialized laboratories, financial resources, and a research team.

A bad mentor, on the other hand, can be a barrier. A mentor can harm the career of the new investigator, for example, by taking credit for findings that arise from the new investigator's work, or assuming the lead role on publishing or presenting it. More commonly, many mentors are simply too busy or distracted to pay attention to the new investigator's needs. In either case, once discussions with the mentor have proved fruitless, we recommend finding a way to move on to a more appropriate advisor, perhaps by involving a neutral senior colleague to help in the negotiations. **Changing mentors** can be hazardous, emphasizing the importance of choosing a good mentor in the first place; it is perhaps the *single most important decision* a new investigator makes.

Your mentor may give you a database and ask you to come up with a research question. In that situation, it's important to identify (1) the overlap between what's in the database and your own research interests, and (2) the quality of the database. If there isn't enough overlap or the data are irrevocably flawed, find a way to move on to another project.

CHARACTERISTICS OF A GOOD RESEARCH QUESTION

The characteristics of a research question that lead to a good study plan are that it be Feasible, Interesting, Novel, Ethical, and Relevant (which form the mnemonic **FINER**; Table 2.1).

Feasible

It is best to know the practical limits and problems of studying a question early on, before wasting much time and effort along unworkable lines.

- **Number of subjects.** Many studies do not achieve their intended purposes because they cannot enroll enough subjects. A preliminary calculation of the sample size requirements of the study early on can be quite helpful (Chapter 6), together with an estimate of the number of subjects likely to be available for the study, the number who would be excluded or refuse to participate, and the number who would be lost to follow-up. Even careful planning often produces estimates that are overly optimistic, and the investigator should assure that there are enough eligible and willing subjects. It is sometimes necessary to carry out a pilot survey or chart review to be sure. If the number of subjects appears insufficient, the investigator can consider several strategies: expanding the inclusion criteria, eliminating unnecessary exclusion criteria, lengthening the time frame for enrolling subjects, acquiring additional sources of subjects, developing more precise measurement approaches, inviting colleagues to join in a multicenter study, and using a different study design.
- **Technical expertise.** The investigators must have the skills, equipment, and experience needed for designing the study, recruiting the subjects, measuring the variables, and managing and analyzing the data. Consultants can help to shore up technical aspects that are unfamiliar to the investigators, but for major areas of the study it is better to have an experienced colleague steadily involved as a coinvestigator; for example, it is wise to include a statistician as a member of the research team from the beginning of the planning process. It is best to use familiar and established approaches, because the process of developing new

TABLE 2.1 FINER CRITERIA FOR A GOOD RESEARCH QUESTION AND STUDY PLAN

Feasible	
	Adequate number of subjects
	Adequate technical expertise
	Affordable in time and money
	Manageable in scope
	Fundable
Interesting	
	Getting the answer intrigues the investigator and her colleagues
Novel	
	Provides new findings
	Confirms, refutes, or extends previous findings
	May lead to innovations in concepts of health and disease, medical practice, or methodologies for research
Ethical	
	A study that the institutional review board will approve
Relevant	
	Likely to have significant impacts on scientific knowledge, clinical practice, or health policy
	May influence directions of future research

methods and skills is time-consuming and uncertain. When a new approach is needed, such as measurement of a new biomarker, expertise in how to accomplish the innovation should be sought.

- **Cost in time and money.** It is important to estimate the costs of each component of the project, bearing in mind that the time and money needed will generally exceed the amounts projected at the outset. If the projected costs exceed the available funds, the only options are to consider a less expensive design or to develop additional sources of funding. Early recognition of a study that is too expensive or time-consuming can lead to modification or abandonment of the plan before expending a great deal of effort.
- **Scope.** Problems often arise when an investigator attempts to accomplish too much, making many measurements at repeated contacts with a large group of subjects in an effort to answer too many research questions. The solution is to narrow the scope of the study and focus only on the most important goals. Many scientists find it difficult to give up the opportunity to answer interesting side questions, but the reward may be a better answer to the main question at hand.
- **Fundability.** Few investigators have the personal or institutional resources to fund their own research projects, particularly if subjects need to be enrolled and followed, or expensive measurements must be made. The most elegantly designed research proposal will not be feasible if no one will pay for it. Finding sources of funding is discussed in Chapter 19.

Interesting

An investigator may have many motivations for pursuing a particular research question: because it will provide financial support, because it is a logical or important next step in building a career, or because getting at the truth of the matter is interesting. We like this last reason; it is one that grows as it is exercised and that provides the intensity of effort needed for overcoming the many hurdles and frustrations of the research process. However, it is wise to confirm that you are not the only one who finds a question interesting. Speak with mentors, outside experts, and representatives of potential funders such as NIH project officers before devoting substantial energy to develop a research plan or grant proposal that peers and funding agencies may consider dull.

Novel

Good clinical research contributes new information. A study that merely reiterates what is already established is not worth the effort and cost and is unlikely to receive funding. The novelty of a proposed study can be determined by thoroughly reviewing the literature, consulting with experts who are familiar with unpublished ongoing research, and searching for abstracts of projects in your area of interest that have been funded using the NIH Research Portfolio Online Reporting Tools (**RePORT**) website (http://report.nih.gov/categorical_spending.aspx.) Reviews of studies submitted to NIH give considerable weight to whether a proposed study is **innovative** (5) such that a successful result could shift paradigms of research or clinical practice through the use of new concepts, methods, or interventions (Chapter 19). Although novelty is an important criterion, a research question need not be totally original—it can be worthwhile to ask whether a previous observation can be replicated, whether the findings in one population also apply to others, or whether a new measurement method can clarify the relationship between known risk factors and a disease. A confirmatory study is particularly useful if it avoids the weaknesses of previous studies or if the result to be confirmed was unexpected.

Ethical

A good research question must be ethical. If the study poses unacceptable physical risks or invasion of privacy (Chapter 14), the investigator must seek other ways to answer the question.

If there is uncertainty about whether the study is ethical, it is helpful to discuss it at an early stage with a representative of the institutional review board (IRB).

Relevant

A good way to decide about relevance is to imagine the various outcomes that are likely to occur and consider how each possibility might advance scientific knowledge, influence practice guidelines and health policy, or guide further research. NIH reviewers emphasize the **significance** of a proposed study: the importance of the problem, how the project will improve scientific knowledge, and how the result will change concepts, methods, or clinical services.

■ DEVELOPING THE RESEARCH QUESTION AND STUDY PLAN

It helps a great deal to write down the research question and a brief (one-page) outline of the **study plan** at an early stage (Appendix 1). This requires some self-discipline, but it forces the investigator to clarify her ideas about the plan and to discover specific problems that need attention. The outline also provides a basis for specific suggestions from colleagues.

Problems and Approaches

Two complementary approaches to the problems involved in developing a research question deserve special emphasis.

The first is the importance of getting good **advice**. We recommend a research team that includes representatives of each of the major disciplines involved in the study, and that includes at least one senior scientist. In addition, it is a good idea to consult with specialists who can guide the discovery of previous research on the topic and the choice and design of measurement techniques. Sometimes a local **expert** will do, but it is often useful to contact individuals in other institutions who have published pertinent work on the subject. A new investigator may be intimidated by the prospect of writing or calling someone she knows only as an author in the *Journal of the American Medical Association,* but most scientists respond favorably to such requests for advice.

The second approach is to allow the study plan to gradually emerge from an **iterative process** of making incremental changes in the study's design, estimating the sample size, reviewing with colleagues, pretesting key features, and revising. Once the one-page study outline is specified, formal review by colleagues will usually result in important improvements. As the protocol takes shape pilot studies of the availability and willingness of sufficient numbers of subjects may lead to changes in the recruitment plan. The preferred imaging test may turn out to be prohibitively costly and a less expensive alternative sought.

Primary and Secondary Questions

Many studies have more than one research question. Experiments often address the effect of the intervention on more than one outcome; for example, the Women's Health Initiative was designed to determine whether reducing dietary fat intake would reduce the risk of breast cancer, but an important secondary hypothesis was to examine the effect on coronary events (5). Almost all cohort and case–control studies look at several risk factors for each outcome. The advantage of designing a study with several research questions is the efficiency that can result, with several answers emerging from a single study. The disadvantages are the increased complexity of designing and implementing the study and of drawing statistical inferences when there are multiple hypotheses (Chapter 5). A sensible strategy is to establish a **single primary research question** around which to focus the study plan and sample size estimate, adding **secondary research questions** about other predictors or outcomes that may also produce valuable conclusions.

TRANSLATIONAL RESEARCH

Translational research refers to studies of how to translate findings from the ivory tower into the "real world," how to assure that scientific creativity has a favorable impact on public health. Translational research (6) comes in two main flavors (Figure 2.2):

- Applying basic science findings from laboratory research in clinical studies of patients (sometimes abbreviated as **T1** research), and
- Applying the findings of these clinical studies to alter health practices in the community (sometimes abbreviated as **T2** research).

Both forms of translational research require identifying a "translation" opportunity. Just as a literary translator first needs to find a novel or poem that merits translating, a translational research investigator must first target a scientific finding or new technology that could have an important impact on clinical research, practice, or public health. Among the strategies for making this challenging choice, it may be helpful to pay attention to colleagues when they talk about their latest findings, to presentations at national meetings about novel methods, and to speculation about mechanisms in published reports.

Translating from Laboratory to Clinical Research (T1)

A host of **tools** have become available for clinical investigations, including DNA sequencing, gene expression arrays, molecular imaging, and proteomics. From the viewpoint of a clinical investigator, there is nothing epidemiologically different about these novel measurements, technologies, or test results. The chapter on measurements will be useful in planning studies involving these types of measurements (Chapter 4), as will the advice about study design (Chapters 7–12), population samples (Chapter 3), and sample size (Chapter 6). Especially relevant to genomics and other "omics" will be the concern with multiple hypothesis testing (Chapter 5).

Compared with ordinary clinical research, being a successful T1 translational investigator often requires having an additional skill set or working with a collaborator with those skills. **Bench-to-bedside** research necessitates a thorough understanding of the underlying basic science. Although many clinical researchers believe that they can master this knowledge—just like many laboratory-based researchers believe doing clinical research requires no special training—in reality, the skills hardly overlap. For example, suppose a basic scientist has identified a gene that

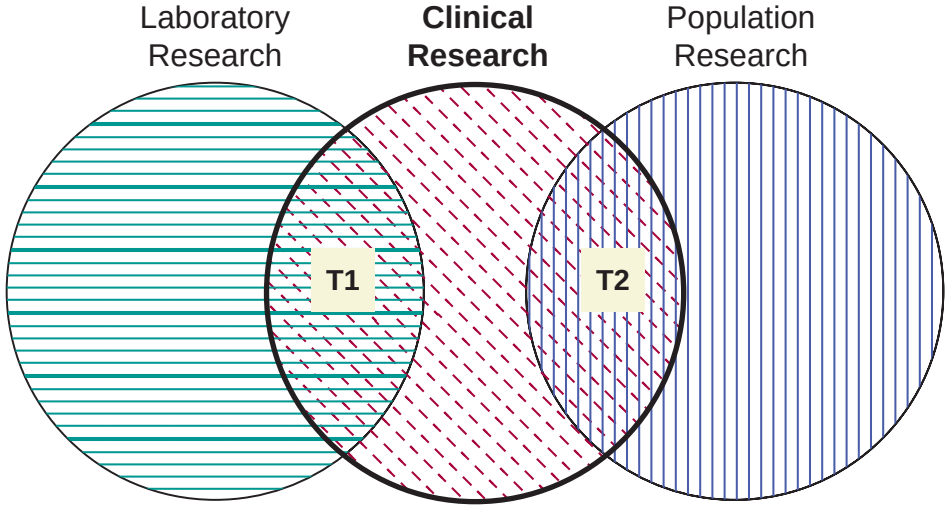


FIGURE 2.2 Translational research is the component of clinical research that interacts with basic science research (hatched area T1) or with population research (hatched area T2).

affects circadian rhythm in mice. A **clinical investigator** whose expertise is in sleep has access to a cohort study with data on sleep cycles and a bank of stored DNA, and wants to study whether there is an association between variants in the human homolog of that gene and sleep. In order to propose a T1 study of that association she needs collaborators who are familiar with that gene, as well as the advantages and limitations of the various methods of genotyping.

Similarly, imagine that a **laboratory-based investigator** has discovered a unique pattern of gene expression in tissue biopsy samples from patients with breast cancer. She should not propose a study of its use as a test for predicting the risk of recurrence of breast cancer without collaborating with someone who understands the importance of clinical research issues, such as test-retest reliability, sampling and blinding, and the effects of prior probability of disease on the applicability of her discovery. Good translational research requires expertise in more than one area. Thus a research team interested in testing a new drug may need scientists familiar with molecular biology, pharmacokinetics, pharmacodynamics, phase I and II clinical trials, and practice patterns in the relevant field of medicine.

Translating from Clinical to Population Research (T2)

Studies that attempt to apply findings from clinical trials to larger and more **diverse populations** often require expertise in identifying high-risk or underserved groups, understanding the difference between screening and diagnosis, and knowing how to implement changes in health care delivery systems. On a practical level, this kind of research usually needs access to large groups of patients (or clinicians), such as those enrolled in health plans or large clinics. Support and advice from the department chair, the chief of the medical staff at an affiliated hospital, the leader of a managed care organization, or a representative from a community organization may be helpful when planning these studies.

Some investigators take a short cut when doing this type of translational research, expanding a study in their own clinic by studying patients in their colleagues' practices (e.g., a house staff-run clinic in an academic medical center) rather than involving practitioners in the community. This is a bit like translating Aristophanes into modern Greek—it will still not be very useful for English-speaking readers. Chapter 18 emphasizes the importance of getting as far into the community as possible.

Testing research findings in larger populations often requires adapting methods to fit organizations. For example, in a study of whether a new office-based diet and exercise program will be effective in the community, it may not be possible to randomly assign individual patients. One solution would be to randomly assign physician practices instead. This may require collaborating with an expert on cluster sampling and clustered analyses. Many T2 research projects aimed to improve medical care use proxy "process" variables as their outcomes. For example, if clinical trials have established that a new treatment reduces mortality from sepsis, a translational research study comparing two programs for implementing and promoting use of the new treatment might not need to have mortality as the outcome. Rather, it might just compare the percentages of patients with sepsis who received the new treatment. Moving research from settings designed for research into organizations designed for medical care or other purposes requires flexibility and creativity in applying principles that assure as much rigor and validity of the study results as possible.

■ SUMMARY

1. All studies should start with a **research question** that addresses what the investigator would like to know. The goal is to find one that can be developed into a good **study plan**.
2. **Scholarship** is essential to developing research questions that are worth pursuing. A **systematic review** of research pertinent to an area of research interest is a good place to start. Attending **conferences** and staying alert to new results extends the investigator's expertise beyond what is already published.

3. The single most important decision a new investigator makes is her choice of one or two senior scientists to serve as her **mentor(s)**: experienced investigators who will take time to **meet**, provide **resources** and **connections**, encourage **creativity**, and promote the **independence** and visibility of their junior scientists.
4. Good research questions arise from finding new collaborators at **conferences**, from critical thinking about clinical practices and problems, from applying **new methods** to old issues, and from considering ideas that emerge from **teaching**, **daydreaming**, and **tenacious pursuit** of solutions to vexing problems.
5. Before committing much time and effort to writing a proposal or carrying out a study, the investigator should consider whether the research question and study plan are "FINER": **feasible**, **interesting**, **novel**, **ethical**, and **relevant**. Those who fund research give priority to proposals that may have innovative and **significant impacts** on science and health.
6. Early on, the research question should be developed into a one-page written **study outline** that specifically describes how many subjects will be needed, how the subjects will be selected, and what measurements will be made.
7. Developing the research question and study plan is an **iterative process** that includes consultations with advisors and friends, a growing familiarity with the literature, and **pilot studies** of the recruitment and measurement approaches.
8. Most studies have more than one question, and it is useful to focus on a **single primary question** in designing and implementing the study.
9. **Translational research** is a type of clinical research that studies the application of basic science findings in clinical studies of patients (**T1**) and how to apply these findings to improve health practices in the community (**T2**); it requires collaborations between **laboratory** and **population-based investigators**, using the **clinical research methods** presented in this book.

REFERENCES

1. The ATAC Trialists Group. Anastrazole alone or in combination with tamoxifen versus tamoxifen alone for adjuvant treatment of postmenopausal women with early breast cancer: first results of the ATAC randomized trials. *Lancet* 2002;359:2131–2139.
2. Quinn J, Cummings S, Callaham M, et al. Suturing versus conservative management of lacerations of the hand: randomized controlled trial. *BMJ* 2002;325:299–301.
3. Kuhn TS. *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press, 1962.
4. Yaffe K, Browner W, Cauley J, et al. Association between bone mineral density and cognitive decline in older women. *J Am Geriatr Soc* 1999;47:1176–1182.
5. Prentice RL, Caan B, Chlebowski RT, et al. Low-fat dietary pattern and risk of invasive breast cancer. *JAMA* 2006;295:629–642.
6. Zerhouni EA. US biomedical research: basic, translational and clinical sciences. *JAMA* 2005;294:1352–1358.

CHAPTER 3

Choosing the Study Subjects: Specification, Sampling, and Recruitment

Stephen B. Hulley, Thomas B. Newman, and Steven R. Cummings

A good choice of study subjects serves the vital purpose of ensuring that the findings in the study accurately represent what is going on in the **population** of interest. The protocol must specify a **sample** of subjects that can be studied at an acceptable cost in time and money (i.e., modest in size and convenient to access), yet large enough to control random error and representative enough to allow generalizing study findings to populations of interest. An important precept here is that **generalizability** is rarely a simple yes-or-no matter; it is a complex qualitative judgment that depends on the investigator's choice of population and of sampling design.

We will come to the issue of choosing the appropriate *number* of study subjects in Chapter 6. In this chapter we address the process of **specifying** and **sampling** the *kinds* of subjects who will be representative and feasible (Figure 3.1). We also discuss strategies for **recruiting** these people to participate in the study.

BASIC TERMS AND CONCEPTS

Populations and Samples

A **population** is a complete set of people with specified characteristics, and a **sample** is a subset of the population. In lay usage, the characteristics that define a population tend to be

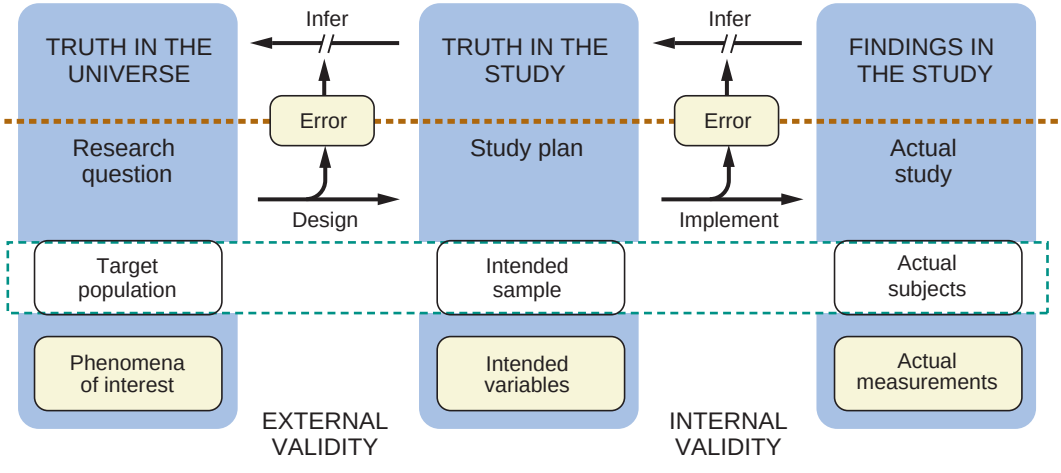


FIGURE 3.1 This chapter focuses on choosing a sample of study subjects that represent the population of interest for the research question.

geographic—for example, the population of Canada. In research, the defining characteristics are also clinical, demographic, and temporal:

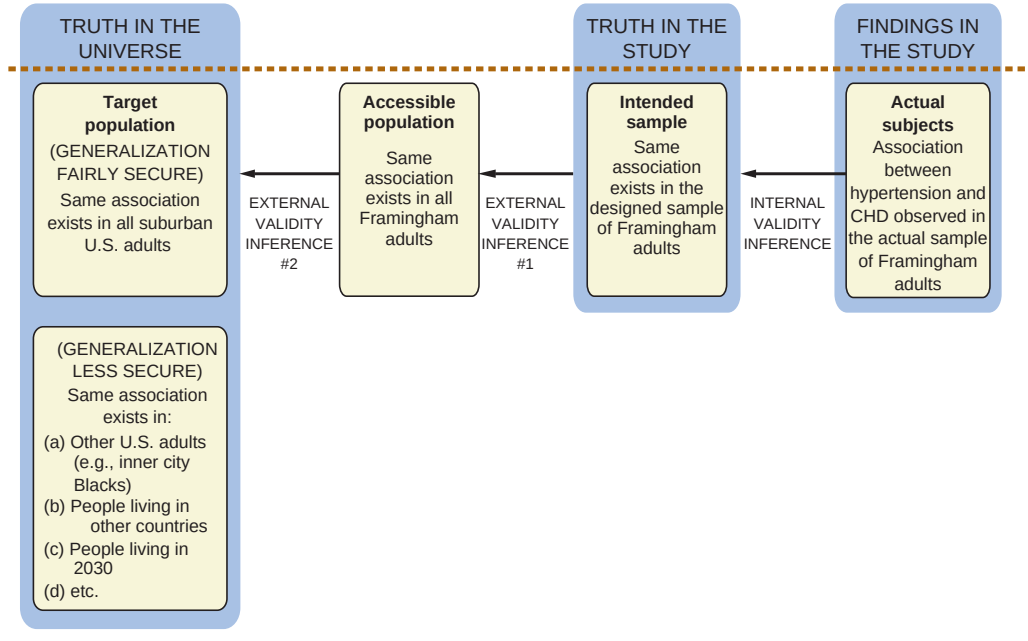
- Clinical and demographic characteristics define the **target population**, the large set of people throughout the world to which the results may be generalized—teenagers with asthma, for example.
- The **accessible population** is a geographically and temporally defined subset of the target population that is available for study—teenagers with asthma living in the investigator's town this year.
- The **intended study sample** is the subset of the accessible population that the investigator seeks to include in the study.
- The **actual study sample** is the group of subjects that does participate in the study.

Generalizing the Study Findings

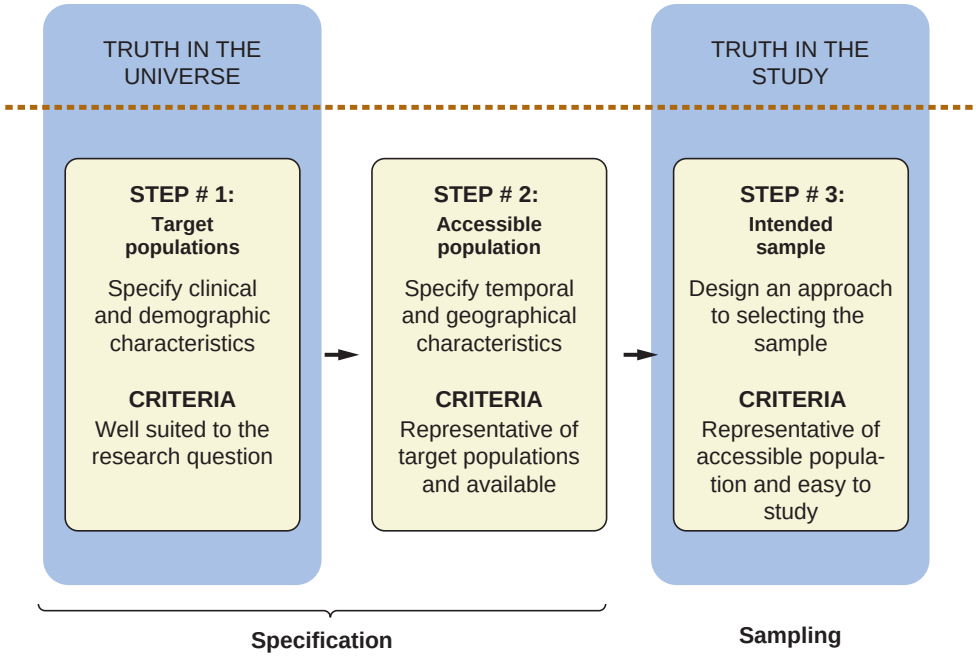
The classic Framingham Study was an early approach to scientifically designing a study to allow **inferences** from findings observed in a sample to be applied to a population (Figure 3.2).

The sampling design called for identifying all the families in Framingham with at least one person aged 30–59, listing the families in order by address, and then asking age-eligible persons in the first two of every set of three families to participate. This "systematic" sampling design is not as tamperproof as choosing each subject by a random process (as discussed later in this chapter), but two more serious concerns were the facts that one-third of the Framingham residents selected for the study refused to participate, and that in their place the investigators accepted age-eligible residents who were not in the sample and volunteered (1).

Because respondents are often healthier than nonrespondents, especially if they are volunteers, the characteristics of the actual sample undoubtedly differed from those of the intended sample. Every sample has some errors, however, and the issue is how much damage has been



■ **FIGURE 3.2** Inferences in generalizing from the study subjects to the target populations proceed from right to left.



■ **FIGURE 3.3** Steps in designing the protocol for choosing the study subjects.

done. The Framingham Study sampling errors do not seem large enough to invalidate the conclusion that risk relationships observed in the study—for example, that hypertension is a risk factor for coronary heart disease (CHD)—can be generalized to all the residents of Framingham.

The next concern is the validity of generalizing the finding that hypertension is a risk factor for CHD from the accessible population of Framingham residents to target populations elsewhere. This inference is more subjective. The town of Framingham was selected not with a scientific sampling design, but because it seemed fairly typical of middle-class white communities in the United States and was convenient to the investigators. The validity of generalizing the Framingham risk relationships to populations in other parts of the country involves the precept that, in general, analytic studies and clinical trials that address biologic relationships produce more widely generalizable results across diverse populations than descriptive studies that address distributions of characteristics. Thus, the strength of hypertension as a risk factor for CHD is similar in Caucasian Framingham residents to that observed in inner city African Americans, but the prevalence of hypertension is much higher in the latter population.

Steps in Designing the Protocol for Acquiring Study Subjects

The inferences in Figure 3.2 are presented from right to left, the sequence used for interpreting the findings of a completed study. An investigator who is planning a study reverses this sequence, beginning on the left (Figure 3.3). She begins by specifying the clinical and demographic characteristics of the target population that will serve the research question well. She then uses **geographic** and **temporal criteria** to specify a study sample that is representative and practical.

■ SELECTION CRITERIA

If an investigator wants to study the efficacy of low dose testosterone supplements versus placebo for enhancing libido in postmenopausal women, she can begin by creating selection criteria that define the population to be studied.

Establishing Selection Criteria

Inclusion criteria define the main characteristics of the target population that pertain to the research question (Table 3.1). Age is often a crucial factor, and in this study the investigator might decide to focus on women in their fifties, speculating that in this group the benefit-to-harm ratio of the drug might be optimal; another study might make a different decision and focus on older decades. The investigator also might incorporate African American, Hispanic, and Asian women in the study in an effort to expand generalizability. This is generally a good idea, but it's important to realize that the increase in generalizability is illusory if there is other evidence to suggest that the effects differ by race. In that case the investigator would need enough women of each race to statistically test for the presence of **effect modification** (an effect in one race that is different from that in other races, also known as "an interaction"; Chapter 9); the number needed is generally large, and most studies are not powered to detect effect modification.

Inclusion criteria that address the geographic and temporal characteristics of the accessible population often involve trade-offs between scientific and practical goals. The investigator may find that patients at her own hospital are an available and inexpensive source of subjects. But she must consider whether peculiarities of the local referral patterns might interfere with generalizing the results to other populations. On these and other decisions about inclusion criteria, there is no single course of action that is clearly right or wrong; the important thing is to make decisions that are sensible, that can be used consistently throughout the study, and that can be clearly described to others who will be deciding to whom the published conclusions apply.

TABLE 3.1 DESIGNING SELECTION CRITERIA FOR A CLINICAL TRIAL OF LOW DOSE TESTOSTERONE VERSUS PLACEBO TO ENHANCE LIBIDO IN MENOPAUSE

	DESIGN FEATURE	EXAMPLE
Inclusion criteria (be specific)	Specifying populations relevant to the research question and efficient for study:	
	Demographic characteristics	Women 50 to 59 years old
	Clinical characteristics	Good general health Has a sexual partner Is concerned about decreased libido
	Geographic (administrative) characteristics	Patients attending clinic at the investigator's hospital
	Temporal characteristics	Between January 1 and December 31 of specified year
Exclusion criteria (be parsimonious)	Specifying subsets of the population that will *not* be studied because of:	
	A high likelihood of being lost to follow-up	Alcoholic Plans to move out of state
	An inability to provide good data	Disoriented Has a language barrier*
	Being at high risk of possible adverse effects	History of myocardial infarction or stroke

*Alternatives to excluding those with a language barrier (when these subgroups are sizeable and important to the research question) would be collecting nonverbal data or using bilingual staff and questionnaires.

Specifying clinical characteristics for selecting subjects often involves difficult judgments, not only about which factors are important to the research question, but about how to define them. How, for example, would an investigator put into practice the criterion that the subjects be in "good health"? She might decide not to include patients with any self-reported illness, but this would likely exclude large numbers of subjects who are perfectly suitable for the research question at hand.

More reasonably, she might exclude only those with diseases that could interfere with follow-up, such as metastatic cancer. This would be an example of "**exclusion criteria**," which indicate individuals who meet the inclusion criteria and would be suitable for the study were it not for characteristics that might interfere with the success of follow-up efforts, the quality of the data, or the acceptability of randomized treatment (Table 3.1). Difficulty with the English language, psychological problems, alcoholism, and serious illness are examples of exclusion criteria. **Clinical trials** differ from observational studies in being more likely to have exclusions mandated by concern for the safety of an intervention in certain patients; for example, the use of drugs in pregnant women (Chapter 10). A good general rule that keeps things simple and preserves the number of potential study subjects is to have as *few exclusion criteria* as possible.

Clinical Versus Community Populations

If the research question involves patients with a disease, hospitalized or clinic-based patients are easier to find, but selection factors that determine who comes to the hospital or clinic may have an important effect. For example, a specialty clinic at a tertiary care medical center attracts patients from afar with serious forms of the disease, giving a distorted impression of the features and prognosis that are seen in ordinary practice. Sampling from primary care practices can be a better choice.

Another common option in choosing the sample is to select subjects in the community who represent a healthy population. These samples are often recruited using mail, e-mail, or advertising via Internet, broadcast, or print media; they are not fully representative of a general population because some kinds of people are more likely than others to volunteer or be active users of Internet or e-mail. True "population-based" samples are difficult and expensive to recruit, but useful for guiding public health and clinical practice in the community. One of the largest and best examples is the National Health and Nutrition Examination Survey (NHANES), a **representative** sample of U.S. residents.

The size and diversity of a sample can be increased by collaborating with colleagues in other cities, or by using preexisting data sets such as NHANES and Medicare data. **Electronically accessible data sets** from public health agencies, healthcare providing organizations, and medical insurance companies have come into widespread use in clinical research and may be more representative of national populations and less time-consuming than other possibilities (Chapter 13).

SAMPLING

Often the number of people who meet the selection criteria is too large, and there is a need to select a **sample** (subset) of the population for study.

Nonprobability Samples

In clinical research the study sample is often made up of people who meet the entry criteria and are easily accessible to the investigator. This is termed a **convenience sample**. It has obvious advantages in cost and logistics, and is a good choice for some research questions.

A **consecutive sample** can minimize volunteerism and other selection biases by consecutively selecting subjects who meet the entry criteria. This approach is especially desirable, for example, when it amounts to taking the entire accessible population over a long enough period to include seasonal variations or other temporal changes that are important to the research question.

The validity of drawing inferences from any sample is the premise that, for the purpose of answering the research question at hand, it sufficiently represents the accessible population. With convenience samples this requires a subjective judgment.

Probability Samples

Sometimes, particularly with descriptive research questions, there is a need for a scientific basis for generalizing the findings in the study sample to the population. Probability sampling, the gold standard for ensuring generalizability, uses a random process to guarantee that each unit of the population has a specified chance of being included in the sample. It is a scientific approach that provides a rigorous basis for estimating the fidelity with which phenomena observed in the sample represent those in the population, and for computing statistical significance and confidence intervals. There are several versions of this approach.

- A **simple random sample** is drawn by enumerating (listing) all the people in the population from which the sample will be drawn, and selecting a subset at random. The most common use of this approach in clinical research is when the investigator wishes to select a representative subset from a population that is larger than she needs. To take a random sample of the cataract surgery patients at her hospital, for example, the investigator could list all such patients on the operating room schedules for the period of study, then use a table of random numbers to select individuals for study (Appendix 3).
- A **systematic sample** resembles a simple random sample in the first step, enumerating the population, but differs in that the sample is selected by a preordained periodic process (e.g., the Framingham approach of taking the first two out of every three families from a list of town families ordered by address). Systematic sampling is susceptible to errors caused by natural periodicities in the population, and it allows the investigator to predict and perhaps manipulate those who will be in the sample. It offers no logistic advantages over simple random sampling, and in clinical research it is rarely a better choice.
- A **stratified random sample** begins by dividing the population into subgroups according to characteristics such as sex or race, and taking a random sample from each of these "strata." The Stratified subsamples can be weighted to draw disproportionately from subgroups that are less common in the population but of special interest to the investigator. In studying the incidence of toxemia in pregnancy, for example, the investigator could stratify the population by race and then sample equal numbers from each stratum. Less common races would then be overrepresented, yielding incidence estimates of comparable precision from each racial group.
- A **cluster sample** is a random sample of natural groupings (clusters) of individuals in the population. Cluster sampling is useful when the population is widely dispersed and it is impractical to list and sample from all its elements. Consider, for example, the problem of interviewing patients with lung cancer selected randomly from a statewide database of discharge diagnoses; patients could be studied at lower cost by choosing a random sample of the hospitals and taking the cases from these. Community surveys often use a two-stage cluster sample: A random sample of city blocks is drawn from city blocks enumerated on a map and a field team visits the blocks in the sample, lists all the addresses in each, and selects a subsample of addresses for study by a second random process. A disadvantage of cluster sampling is the fact that naturally occurring groups are often more homogeneous for the variables of interest than the population; each city block, for example, tends to have people of similar socioeconomic status. This means that the effective sample size (after adjusting for within-cluster uniformity) will be somewhat smaller than the number of subjects, and that statistical analysis must take the clustering into account.

Summarizing the Sampling Design Options

The use of descriptive statistics and tests of statistical significance to draw inferences about the population from observations in the study sample is based on the assumption that a probability

sample has been used. But in clinical research a random sample of the whole target population is almost never possible. Convenience sampling, preferably with a consecutive design, is a practical approach that is often suitable. The decision about whether the proposed sampling design is satisfactory requires that the investigator make a **judgment**: *for the research question at hand,* will the conclusions drawn from observations in the study sample be similar to the conclusions that would result from studying a true probability sample of the accessible population? And beyond that, will the conclusions be appropriate for the target population?

■ RECRUITMENT

The Goals of Recruitment

An important factor to consider in choosing the accessible population and sampling approach is the feasibility of recruiting study participants. There are two main goals: (1) to recruit a sample that adequately **represents** the target population, minimizing the prospect of getting the wrong answer to the research question due to systematic error (bias); and (2) to recruit a sufficient **sample size** to minimize the prospect of getting the wrong answer due to random error (chance).

Achieving a Representative Sample

The approach to recruiting a representative sample begins in the design phase with wise decisions about choosing target and accessible populations, and approaches to sampling. It ends with implementation, guarding against errors in applying the entry criteria to prospective study participants, and enhancing successful strategies as the study progresses.

A particular concern, especially for descriptive studies, is the problem of **nonresponse.**[1] The proportion of subjects selected for the study who consent to be enrolled (the response rate) influences the validity of inferring that the enrolled sample represents the population. People who are difficult to reach and those who refuse to participate once they are contacted tend to be different from people who do enroll. The level of nonresponse that will compromise the generalizability of the study depends on the nature of the research question and on the reasons for not responding. A nonresponse rate of 25%, a good achievement in many settings, can seriously distort the estimate of the prevalence of a disease when the disease itself is a cause of nonresponse.

The degree to which nonresponse bias may influence the conclusions of a descriptive study can sometimes be estimated during the study by acquiring additional information on a sample of nonrespondents. The best way to deal with nonresponse bias, however, is to minimize the number of nonrespondents. The problem of failure to make contact with individuals who have been chosen for the sample can be reduced by designing a series of repeated contact attempts using various methods (mail, e-mail, telephone, home visit). Among those contacted, refusal to participate can be minimized by improving the efficiency and attractiveness of the study, by choosing a design that avoids invasive and uncomfortable tests, by using brochures and individual discussion to allay anxiety and discomfort, by providing incentives such as reimbursing the costs of transportation and providing the results of tests, and by circumventing language barriers with bilingual staff and translated questionnaires.

Recruiting Sufficient Numbers of Subjects

Falling short in the rate of recruitment is one of the commonest problems in clinical research. In planning a study it is best to assume that the number of subjects who meet the entry criteria and agree to enter the study will be fewer, sometimes by severalfold, than the number projected

[1]Concern with nonresponse in the process of *recruiting* subjects for a study (the topic of this chapter) is chiefly a concern in descriptive studies that have a primary goal of estimating distributions of variables in particular populations. Nonresponse in the *follow-up* process is often a major issue in any study that follows a cohort over time, and particularly in a clinical trial of an intervention that may alter the response rate (Chapter 10).

at the outset. The approaches to this problem are to estimate the magnitude of the recruitment problem empirically with a pretest, to plan the study with an accessible population that is larger than believed necessary, and to make contingency plans should the need arise for additional subjects. While recruitment is ongoing it is important to closely monitor progress in meeting the recruitment goals and tabulate reasons for falling short of the goals. Understanding why potential subjects are lost to the study at various stages can lead to strategies for reducing these losses.

Sometimes recruitment involves selecting subjects who are already known to the members of the research team (e.g., in a study of a new treatment in patients attending the investigator's clinic). Here the chief concern is to present the opportunity for participation in the study fairly, making clear the advantages and disadvantages. In discussing participation, the investigator must recognize the ethical dilemmas that arise when her advice as the patient's physician might conflict with her interests as an investigator (Chapter 14).

Often recruitment involves contacting populations that are not known to the members of the research team. It is helpful if at least one member of the research team has previous experience with the approaches for contacting the prospective subjects. These include screening in work settings or public places such as shopping malls; sending out large numbers of mailings to listings such as driver's license holders; advertising on the Internet; inviting referrals from clinicians; carrying out retrospective record reviews; and examining lists of patients seen in clinic and hospital settings. Some of these approaches, particularly the latter two, involve concerns with privacy invasion that must be considered by the institutional review board.

It may be helpful to prepare for recruitment by getting the support of important organizations. For example, the investigator can meet with hospital administrators to discuss a clinic-based sample, and with community leaders, the medical society and county health department to plan a community screening operation or mailing to physicians. Written endorsements can be included as an appendix in applications for funding. For large studies it may be useful to create a favorable climate in the community by giving public lectures or by advertising through radio, TV, newspapers, fliers, websites, and mass mailings.

SUMMARY

1. Most clinical research is based, philosophically and practically, on the use of a **sample** to represent a **population**.
2. The advantage of sampling is **efficiency**: It allows the investigator to draw inferences about a large population by examining a subset at relatively small cost in time and effort. The disadvantage is the sources of **error** it introduces: If the sample is not sufficiently representative for the research question at hand the findings may not **generalize** well to the target population, and if it is not large enough the findings may not sufficiently minimize the role of **chance**.
3. In designing a sample, the investigator begins by conceptualizing the **target population** with a specific set of **inclusion criteria** that establish demographic and clinical characteristics of subjects well suited to the research question.
4. She then selects an appropriate **accessible population** that is geographically and temporally convenient, and defines a parsimonious set of **exclusion criteria** that eliminate subjects who are unethical or inappropriate to study.
5. The next step is to design an approach to **sampling** the population. A **convenience sample** may be adequate, especially for initial studies of some questions, and a **consecutive sample** is often a good choice. **Simple random sampling** can be used to reduce the size of the sample if necessary, and **other probability sampling** strategies (**stratified** and **cluster**) are useful in certain situations.
6. Finally, the investigator must design and implement strategies for **recruiting** a sample of subjects that is sufficiently **representative** of the target population to control systematic sources of error, and **large enough** to control random sources of error.

APPENDIX 3

This table provides a simple paper-based way to select a 10% random sample from a table of random numbers. Begin by enumerating (listing and numbering) every person in the population to be sampled. Then decide on a rule for obtaining an appropriate series of numbers; for example, if your list has 741 elements (which you have numbered 1 to 741), your rule might be to go vertically down each column in this table using the first three digits of each number (beginning at the upper left, the numbers are 104, 223, etc.) and to select the first 74 different numbers that fall in the range of 1 to 741. Finally, pick a starting point by an arbitrary process (closing your eyes and putting your pencil on some number in the table is one way to do it) and begin applying the rule. The modern approach, with a computerized series of random numbers, basically works the same way.

TABLE 3.2 SELECTING A RANDOM SAMPLE FROM A TABLE OF RANDOM NUMBERS

10480	15011	01536	81647	91646	02011
22368	46573	25595	85393	30995	89198
24130	48390	22527	97265	78393	64809
42167	93093	06243	61680	07856	16376
37570	33997	81837	16656	06121	91782
77921	06907	11008	42751	27756	53498
99562	72905	56420	69994	98872	31016
96301	91977	05463	07972	18876	20922
89572	14342	63661	10281	17453	18103
85475	36857	53342	53998	53060	59533
28918	79578	88231	33276	70997	79936
63553	40961	48235	03427	49626	69445
09429	93969	52636	92737	88974	33488
10365	61129	87529	85689	48237	52267
07119	97336	71048	08178	77233	13916
51085	12765	51821	51259	77452	16308
02368	21382	52404	60268	89368	19885
01011	54092	33362	94904	31273	04146
52162	53916	46369	58569	23216	14513
07056	97628	33787	09998	42698	06691
48663	91245	85828	14346	09172	30163
54164	58492	22421	74103	47070	25306
32639	32363	05597	24200	38005	13363
29334	27001	87637	87308	58731	00256
02488	33062	28834	07351	19731	92420
81525	72295	04839	96423	24878	82651
29676	20591	68086	26432	46901	20949
00742	57392	39064	66432	84673	40027
05366	04213	25669	26422	44407	44048
91921	26418	64117	94305	26766	25940

REFERENCE

1. www.framinghamheartstudy.org/about/background.html, accessed 7/23/12.

CHAPTER 4

Planning the Measurements: Precision, Accuracy, and Validity

Stephen B. Hulley, Thomas B. Newman, and Steven R. Cummings

Measurements describe phenomena in terms that can be analyzed statistically, and the validity of a study depends on how well the variables designed for the study represent the phenomena of interest (Figure 4.1). How well does a handheld glucometer measure blood glucose, for example, or an insomnia questionnaire detect amount and quality of sleep?

This chapter begins by considering how the choice of **measurement scale** influences the information content of the measurement. We then turn to the central goal of minimizing measurement error: how to design measurements that are relatively **precise** (free of random error) and **accurate** (free of systematic error), thereby enhancing the appropriateness of drawing inferences from these measurements to the phenomena of interest. We address the concept of **validity**, a qualitative relative of accuracy, before concluding with some considerations for clinical and translational research, noting especially the advantages of storing specimens for later measurements.

MEASUREMENT SCALES

Table 4.1 presents a simplified classification of measurement scales and the information that results. The classification is important because some types of variables are **more informative** than others, adding power or reducing sample size requirements, and revealing more detailed distribution patterns.

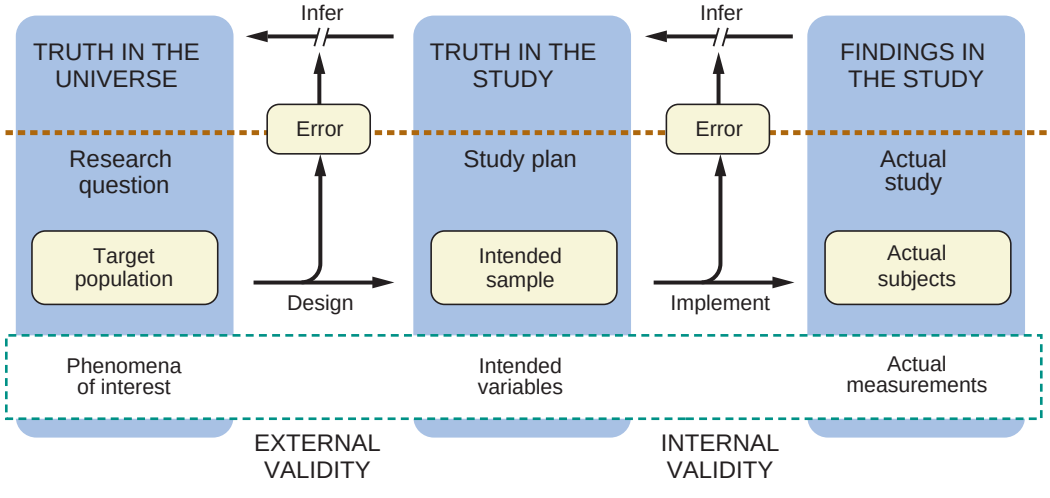


FIGURE 4.1 Designing measurements that represent the phenomena of interest.

TABLE 4.1 MEASUREMENT SCALES

TYPE OF MEASUREMENT	CHARACTERISTICS OF VARIABLE	EXAMPLE	DESCRIPTIVE STATISTICS	STATISTICAL POWER
		Categorical		
Dichotomous	Two categories	Vital status (alive or dead)	Counts, proportions	Low
Nominal	Unordered categories	Race; blood type	Same as above	Low
Ordinal	Ordered categories with intervals that are not quantifiable	Degree of pain; social class	In addition to the above: medians	Intermediate
		Numeric		
Continuous or discrete†	Ranked spectrum with quantifiable intervals	Weight; number of cigarettes/day	In addition to the above: means, standard deviations	High

†Continuous variables have an infinite number of values (e.g., weight), whereas discrete numeric variables are more limited (e.g., number of cigarettes/day). Discrete variables that have a large number of possible values resemble continuous variables for practical purposes of power and analysis.

Numeric Variables: Continuous and Discrete

Numeric variables can be quantified with a number that expresses how much or how many. **Continuous variables** quantify how much on an infinite scale; the number of possible values of body weight, for example, is limited only by the sensitivity of the machine that is used to measure it. Continuous variables are rich in information. **Discrete numeric variables** quantify how many on a scale with fixed units, usually integers, such as the number of times a woman has been pregnant. Discrete variables that have a considerable number of possible values can resemble continuous variables in statistical analyses and be equivalent for the purpose of designing measurements.

Categorical Variables: Dichotomous, Nominal, and Ordinal

Phenomena that are not suitable for quantification are measured by classifying them in categories. **Categorical variables** with two possible values (e.g., dead or alive) are termed **dichotomous**. Categorical variables with more than two categories (polychotomous) can be further characterized according to the type of information they contain. Among these, **nominal variables** have categories that are not ordered; type O blood, for example, is neither more nor less than type B blood; nominal variables tend to have an absolute qualitative character that makes them straightforward to measure. The categories of **ordinal variables** do have an order, such as severe, moderate, and mild pain. The additional information is an advantage over nominal variables, but because ordinal variables do not specify a numerical or uniform difference between one category and the next, the information content is less than that of discrete or continuous numeric variables.

Choosing a Measurement Scale

A good general rule is to **prefer continuous over categorical** variables when there is a choice, because the additional information they contain improves statistical efficiency. In a study comparing the antihypertensive effects of several treatments, for example, measuring blood pressure in millimeters of mercury allows the investigator to observe the magnitude of the change in every subject, whereas measuring it as hypertensive versus normotensive limits the

assessment. The continuous variable contains more information, and the result is a study with more power and/or a smaller sample size (Chapter 6).

Continuous variables also allow for more flexibility than categorical variables in fitting the data to the nature of the variable or the shape of the association, especially when the relationship might have a complex pattern. For example, a study of the relationship of vitamin D to various cancers would need to measure vitamin D as a continuous variable to be able to detect a possible **U-shaped pattern**, the higher mortality that has been observed in subjects with low or high levels of vitamin D than in those with intermediate levels (1). And a study of predictors of low birth weight babies should record actual birth weight rather than above or below the conventional 2,500 g **threshold**; this leaves the analytic options open, to change the cutoff that defines low birth weight, or to develop an ordinal scale with several categories of birth weight (e.g., >2,500 g, 2,000–2,499 g, 1,500–1,999 g, and <1,500 g).

Similarly, when there is the option of designing the number of response categories in an ordinal scale, as in a question about food preferences, it is often useful to provide a half-dozen categories that range from "strongly dislike" to "extremely fond of." The results can later be collapsed into a dichotomy (dislike and like), but not vice versa.

Many characteristics, particularly symptoms like pain or aspects of lifestyle, are difficult to describe with categories or numbers. But these phenomena often have important roles in diagnostic and treatment decisions, and the attempt to measure them is an essential part of the scientific approach to description and analysis. This is illustrated by the Short Form (SF)-36, a standardized questionnaire for assessing **quality of life** that produces discrete numerical ratings (2). The process of classification and measurement, if done well, can increase the objectivity of our knowledge, reduce bias, and provide a means of communication.

PRECISION

The **precision** of a variable is the degree to which it is reproducible, with nearly the same value each time it is measured. A beam scale can measure body weight with great precision, whereas an interview to measure quality of life is more likely to produce values that vary from one observer or occasion to another. Precision has a very important influence on the power of a study. The more precise a measurement, the greater the statistical power at a given sample size to estimate mean values and to test hypotheses (Chapter 6).

Precision (also called **reproducibility**, **reliability**, and **consistency**) is a function of **random error** (chance variability); the greater the error, the less precise the measurement. There are three main sources of random error in making measurements.

- **Observer variability** is due to the observer, and includes such things as choice of words in an interview and skill in using a mechanical instrument.
- **Instrument variability** is due to the instrument, and includes changing environmental factors (e.g., temperature), aging mechanical components, different reagent lots, and so on.
- **Subject variability** is due to intrinsic biologic variability in the study subjects unrelated to variables under study, such as variablity due to time of day of measurements or time since last food or medication.

Assessing Precision

Precision is assessed as the **reproducibility** of repeated measurements, either comparing measurements made by the same person (within-observer reproducibility) or different people (between-observer reproducibility). Similarly, it can be assessed within or between instruments. The reproducibility of continuous variables is often expressed as either the within-subject standard deviation or the **coefficient of variation** (within-subject standard deviation divided by

the mean).[1] For categorical variables, percent agreement, the interclass correlation coefficient, and the **kappa** statistic are often used (3–5).

Strategies for Enhancing Precision

There are five approaches to minimizing random error and increasing the precision of measurements (Table 4.2):

1. **Standardizing the measurement methods.** All study protocols should include specific instructions for making the measurements (**operational definitions**). This may include written directions on how to prepare the environment and the subject, how to carry out and record the interview, how to calibrate the instrument, and so forth (Appendix 4). This set of materials, part of the **operations manual**, is essential for large and complex studies and recommended for smaller ones. Even when there is only a single observer, specific written guidelines for making each measurement will help her performance to be uniform over the duration of the study and serve as the basis for describing the methods when the results are published.

TABLE 4.2 STRATEGIES FOR REDUCING RANDOM ERROR IN ORDER TO INCREASE PRECISION, WITH ILLUSTRATIONS FROM A STUDY OF ANTIHYPERTENSIVE TREATMENT

STRATEGY TO REDUCE RANDOM ERROR	SOURCE OF RANDOM ERROR	EXAMPLE OF RANDOM ERROR	EXAMPLE OF STRATEGY TO PREVENT THE ERROR
1. Standardizing the measurement methods in an operations manual	Observer	Variation in blood pressure (BP) measurement due to variable rate of cuff deflation (often too fast)	Specify that the cuff be deflated at 2 mm Hg/second
	Subject	Variation in BP due to variable length of quiet sitting before measurement	Specify that subject sit in a quiet room for 5 minutes before BP measurement
2. Training and certifying the observer	Observer	Variation in BP due to variable observer technique	Train observer in standard techniques
3. Refining the instrument	Instrument and observer	Variation in BP due to malfunctioning manometer	Purchase new high quality manometer
4. Automating the instrument	Observer	Variation in BP due to variable observer technique	Use automatic BP measuring device
	Subject	Variation in BP due to subject's emotional reaction to observer	Use automatic BP measuring device
5. Repeating the measurement	Observer, subject, and instrument	All measurements and all sources of variation	Use mean of two or more BP measurements

[1] When there are two measurements of a continuous variable per subject, it may be tempting to express their agreement using a **correlation coefficient**. However, because the correlation coefficient is extremely sensitive to outliers (3,4), a better approach is a "Bland-Altman" plot in which the difference between the two measurements is plotted as a function of their mean. If the absolute value of the difference between the measurements tends to increase linearly with the mean, the coefficient of variation is a better way to summarize variability than the within-subject standard deviation.

2. **Training and certifying the observers.** Training will improve the consistency of measurement techniques, especially when several observers are involved. It is often desirable to design a formal test of the mastery of the techniques specified in the operations manual and to certify that observers have achieved the prescribed level of performance (Chapter 17).
3. **Refining the instruments.** Mechanical and electronic instruments can be engineered to reduce variability. Similarly, questionnaires and interviews can be written to increase clarity and avoid potential ambiguities (Chapter 15).
4. **Automating the instruments.** Variations in the way human observers make measurements can be eliminated with automatic mechanical devices and self-response questionnaires.
5. **Repetition.** The influence of random error from any source is reduced by repeating the measurement, and using the mean of the two or more readings. Precision will be substantially increased by this strategy, the primary limitations being the added cost and practical difficulties of repeating the measurements.

For each measurement in the study, the investigator must decide how vigorously to pursue each of these strategies. This decision can be based on the importance of the variable, the magnitude of the potential problem with precision, and the feasibility and cost of the strategy. In general, the first two strategies (standardizing and training) should always be used, and the fifth (repetition) is an option that is guaranteed to improve precision when it is feasible and affordable.

TABLE 4.3 THE PRECISION AND ACCURACY OF MEASUREMENTS

	PRECISION	ACCURACY
Definition	The degree to which a variable has nearly the same value when measured several times	The degree to which a variable approximates the true value
Best way to assess	Comparison among repeated measures	Comparison with a "gold standard"
Value to study	Increase power to detect effects	Increase validity of conclusions
Threatened by	Random error (chance) contributed by	Systematic error (bias) contributed by
	The observer	The observer
	The subject	The subject
	The instrument	The instrument

ACCURACY

The **accuracy** of a variable is the degree to which it represents the true value.

Accuracy is different from precision in the ways shown in Table 4.3, and the two are not necessarily linked. If serum cholesterol were measured repeatedly using standards that had inadvertently been diluted twofold, for example, the result would be inaccurate but could still

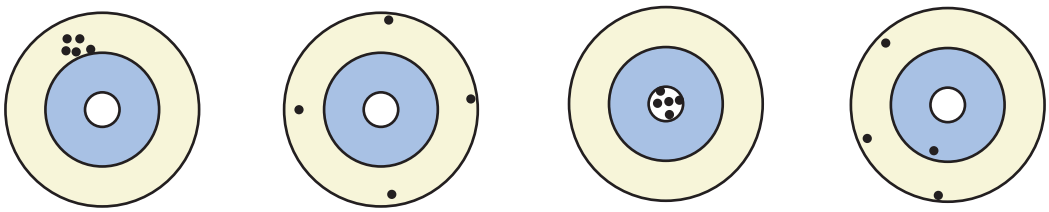


FIGURE 4.2 The difference between precision and accuracy.

be precise (consistently off by a factor of two). This concept is further illustrated in Figure 4.2. Accuracy and precision do often go hand in hand, however, in the sense that many of the strategies for increasing precision will also improve accuracy.

Accuracy is a function of **systematic error** (bias); the greater the error, the less accurate the variable. The three main classes of measurement error noted in the earlier section on precision each have counterparts here.

- **Observer bias** is a distortion, conscious or unconscious, in the perception or reporting of the measurement by the observer. It may represent systematic errors in the way an instrument is operated, such as a tendency to round down blood pressure measurements or to use leading questions in interviewing a subject.
- **Instrument bias** can result from faulty function of a mechanical instrument. A scale that has not been calibrated recently may have drifted downward, producing consistently low body weight readings.
- **Subject bias** is a distortion of the measurement by the study subject, for example, in reporting an event (respondent or recall bias). Patients with breast cancer who believe that alcohol is a cause of their cancer, for example, may exaggerate the alcohol intake they report.

The accuracy of a measurement is best assessed by comparing it, when possible, to a "**gold standard**"—a reference measurement carried out by a technique that is believed to best represent the true value of the characteristic. The decision as to what measurement approach to designate as the gold standard can be a difficult judgment that the investigator needs to make, drawing on previous work in the field.

The degree of accuracy can be expressed, for measurements on a continuous scale, as the mean difference between the measurement under investigation and the gold standard across study subjects. For measurements on a dichotomous scale, accuracy in comparison to a gold standard can be described in terms of sensitivity and specificity (Chapter 12). For measurements on categorical scales with more than two response options, the percent correct on each can be calculated.

Strategies for Enhancing Accuracy

The major approaches to increasing accuracy include the first four strategies listed earlier for precision, and three additional ones (Table 4.4):

1. **Standardizing the measurement methods.**
2. **Training and certifying the observers.**
3. **Refining the instruments.**
4. **Automating the instruments.**
5. **Making unobtrusive measurements.** It is sometimes possible to design measurements that the subjects are not aware of, thereby eliminating the possibility that they will consciously bias the variable. For example, an evaluation of the effect of placing a hand sanitizer and a hand hygiene poster in a hospital cafeteria utilized observers who blended in with cafeteria customers (6).
6. **Calibrating the instrument.** The accuracy of many instruments, especially those that are mechanical or electrical, can be increased by periodic calibration with a gold standard.
7. **Blinding.** This classic strategy does not ensure the overall accuracy of the measurements, but it can eliminate **differential bias** that affects one study group more than another. In a double-blind clinical trial the subjects and observers do not know whether active medicine or placebo has been assigned, and any inaccuracy in measuring the outcome will be the same in the two groups.

The decision on how vigorously to pursue each of these seven strategies for each measurement rests, as noted earlier for precision, on the judgment of the investigator. The considerations are the potential impact that the anticipated degree of inaccuracy will have on the conclusions of the study, and the feasibility and cost of the strategy. The first two strategies

TABLE 4.4 STRATEGIES FOR REDUCING SYSTEMATIC ERROR IN ORDER TO INCREASE ACCURACY, WITH ILLUSTRATIONS FROM A STUDY OF ANTIHYPERTENSIVE TREATMENT

STRATEGY TO REDUCE SYSTEMATIC ERROR	SOURCE OF SYSTEMATIC ERROR	EXAMPLE OF SYSTEMATIC ERROR	EXAMPLE OF STRATEGY TO PREVENT THE ERROR
1. Standardizing the measurement methods in an operations manual	Observer	Consistently high diastolic blood pressure (BP) readings due to using the point at which sounds become muffled	Specify the operational definition of diastolic BP as the point at which sounds cease to be heard
	Subject	Consistently high readings due to measuring BP right after walking upstairs to clinic	Specify that subject sit in quiet room for 5 minutes before measurement
2. Training and certifying the observer	Observer	Consistently high BP readings due to failure to follow procedures specified in operations manual	Trainer checks accuracy of observer's reading with a double-headed stethoscope
3. Refining the instrument	Instrument	Consistently high BP readings with standard cuff in subjects with very large arms	Use extra-wide BP cuff in obese patients
4. Automating the instrument	Observer	Conscious or unconscious tendency for observer to read BP lower in group randomized to active drug	Use automatic BP measuring device
	Subject	BP increase due to proximity of attractive technician	Use automatic BP measuring device
5. Making unobtrusive measurements	Subject	Tendency of subject to overestimate compliance with study drug	Measure study drug level in urine
6. Calibrating the instrument	Instrument	Consistently high BP readings due to manometer being out of adjustment	Calibrate each month
7. Blinding	Observer	Conscious or unconscious tendency for observer to read BP lower in active treatment group	Use double-blind placebo to conceal study group assignment
	Subject	Tendency of subject who knew she was on active drug to overreport side effects	Use double-blind placebo to conceal study group assignment

(standardizing and training) should always be used, calibration is needed for any instrument that has the potential to change over time, and blinding is essential whenever feasible.

VALIDITY

Validity resembles **accuracy**, but we like to think of it as adding a qualitative dimension to considering how well a measurement represents the phenomena of interest. For example, measurements of creatinine and cystatin C in the blood, two chemicals excreted by the kidneys,

might be equally *accurate* (e.g., within 1% of the true level), but cystatin C may be more *valid* as a measure of kidney function because creatinine levels are also influenced by the amount of muscle (7). In Figure 4.2, we can think of validity as describing whether the bull's-eye is in the right target.

Validity is often not amenable to assessment with a gold standard, particularly for measurements aimed at subjective and abstract phenomena such as pain or quality of life. Social scientists have created qualitative and quantitative constructs for addressing the validity of these measurement approaches.

- **Content validity** examines how well the measurement represents all aspects of the phenomena under study; for example, including questions on social, physical, emotional, and intellectual functioning to assess quality of life.
- **Face validity** describes whether the measurement seems inherently reasonable, such as measuring pain on a 10-point scale or social class by household income.
- **Construct validity** is the degree to which a specific measuring device agrees with a theoretical construct; for example, an IQ test should distinguish between people that theory or other measures suggest have different levels of intelligence.
- **Predictive validity** is the ability of the measurement to predict an outcome; for example, how well a questionnaire designed to assess depression predicts job loss or suicide.
- **Criterion-related validity** is the degree to which a new measurement correlates with well accepted existing measures.

The general approach to measuring subjective and abstract phenomena is to begin by searching the literature and consulting with experts in an effort to find a suitable **instrument** (typically a questionnaire) that has already been validated. Using such an instrument has the advantage of making the results of the new study comparable to earlier work in the area, and may simplify and strengthen the process of applying for grants and publishing the results. Its disadvantages, however, are that the validation process may have been suboptimal, and that an instrument taken off the shelf may be outmoded or not optimal for the research question.

If existing instruments are not suitable for the needs of the study, then the investigator may decide to develop a new measurement approach and validate it herself. This can be an interesting challenge and even lead to a worthwhile contribution to the literature, but it generally requires a lot of time and effort (Chapter 15). It is fair to say that the process is often less conclusive than the word "validation" connotes.

■ OTHER FEATURES OF MEASUREMENT APPROACHES

Measurements should be **sensitive** enough to detect differences in a characteristic that are important to the investigator. Just how much sensitivity is needed depends on the research question. For example, a study of whether a new medication helps people to quit smoking could use an outcome measure that is not very sensitive to the *number* of cigarettes smoked each day. On the other hand, if the question is the effect of reducing the nicotine content of cigarettes on the number of cigarettes smoked, the method should be sensitive to differences in daily habits of just a few cigarettes.

An ideal measurement is **specific**, representing only the characteristic of interest. The carbon monoxide level in expired air is a measure of smoking habits that is only moderately specific because it can also be affected by other exposures such as automobile exhaust. The specificity of assessing smoking habits can be increased by adding measurements (such as self-report and serum cotinine level) that are not affected by air pollution.

Measurements should be **appropriate** to the objectives of the study. A study of stress as an antecedent to myocardial infarction, for example, would need to consider which kind of stress (psychological or physical, acute or chronic) was of interest before setting out the operational definitions for measuring it.

Measurements should provide an adequate **distribution of responses** in the study sample. A measure of functional status is most useful if it produces values that range from high in some subjects to low in others. A major reason for pretesting is to ensure that the actual responses do not all cluster around one end of the possible range of responses (Chapter 17).

Whenever possible, measurements should be designed in a way that minimizes subjective judgments. **Objectivity** is achieved by reducing the involvement of the observer and by using automated instruments. One danger in these strategies, however, is the consequent tunnel vision that limits the scope of the observations and the ability to discover unanticipated phenomena. This can be addressed by including some open-ended questions, and an opportunity for acquiring subjective and qualitative data, in addition to the main objective and quantitative measurements.

In designing a study there is a tendency to keep adding items that are not central to the research question but *could* be of interest. It is true that additional measurements increase the likelihood of interesting findings, including some that were not anticipated at the outset. However, it is important to keep in mind the value of **efficiency** and **parsimony**. The full set of measurements should be designed to collect useful data at an affordable cost in time and money. Collecting too much information is a common error that can tire subjects, overwhelm the team making the measurements, and clutter data management and analysis. The result may be a more expensive study that paradoxically is less successful in answering the main research questions.

MEASUREMENTS ON STORED MATERIALS

Clinical research involves measurements on people that range across many domains. Some of these measurements can only be made during contact with the study subject, but many can be carried out later on biological **specimens** banked for chemical or genetic analysis, or on **images** from radiographic and other procedures filed electronically (Table 4.5).

One advantage of such storage is the opportunity to reduce the cost of the study by making measurements only on individuals who turn out during follow-up to have an outcome of interest. A terrific approach to doing this is the nested case–control design (Chapter 8), especially if paired blinded measurements can be made in a single analytic batch eliminating the batch-to-batch component of random error. This approach also has the advantage that scientific advances years after the study is begun may lead to new ideas and measurement techniques that can then be employed, funded by newly submitted grants.

TABLE 4.5 COMMON TYPES OF MEASUREMENTS THAT CAN BE MADE ON STORED MATERIALS

TYPE OF MEASUREMENT	EXAMPLES	BANK FOR LATER MEASUREMENT
Medical history	Diagnoses, medications, operations, symptoms, physical findings	Paper or electronic medical records
Psychosocial factors	Depression, family history	Voice recordings, videotapes
Anthropometric	Height, weight, body composition	Photographs
Biochemical measures	Serum cholesterol, plasma fibrinogen	Serum, plasma, urine, pathology specimens
Genetic/molecular tests	Single nucleotide polymorphisms	DNA
Imaging	Bone density, coronary calcium	X-rays, CT scans, MRIs
Electromechanical	Arrhythmia, congenital heart disease	Electrocardiogram, echocardiogram

The growing interest in **translational research** (Chapter 2) takes advantage of new measurements that have greatly expanded clinical research, for example, in the areas of **genetic and molecular epidemiology** (8, 9) and **imaging**. Measurements on specimens that contain DNA (e.g., saliva and blood) can provide information on genotypes that contribute to the occurrence of disease or modify a patient's response to treatment. Measurements on serum can be used to study molecular causes or consequences of disease; for example, inflammatory markers provide useful information in the pathophysiology of many diseases. It is important to consult with experts regarding the proper collection tubes and storage conditions in order to preserve the quality of the specimens and make them available for the widest spectrum of subsequent use. It is also important to obtain informed consent from participants that covers the scope of potential uses of the specimens.

SUMMARY

1. Variables are either **numerical** or **categorical**. Numerical variables are **continuous** (quantified on an infinite scale) or **discrete** (quantified on a finite scale such as integers); categorical variables are **nominal** (unordered) or **ordinal** (ordered), and those that have only two categories are termed **dichotomous**.
2. Variables that contain more **information** provide greater power and/or allow smaller sample sizes, according to the following **hierarchy**: continuous variables > discrete numeric variables > ordinal variables > nominal and dichotomous variables.
3. The **precision** of a measurement (i.e., the **reproducibility** of replicate measures) is another major determinant of power and sample size. Precision is reduced by **random error (chance)** from three **sources of variability**: the observer, the subject, and the instrument.
4. Strategies for **increasing precision** that should be part of every study are to **operationally define** and **standardize methods** in an **operations manual**. Other strategies that are often useful are **training and certifying observers**, **refining** and **automating** the **instruments**, and **repetition**—using the mean of repeated measurements.
5. The **accuracy** of a measurement is the degree to which it approximates a gold standard. Accuracy is reduced by **systematic error** (**bias**) from the same three sources: the observer, subject, and instrument.
6. The **strategies for increasing accuracy** include all those listed for precision with the exception of repetition. In addition, accuracy is enhanced by **unobtrusive measures**, by **calibration**, and (in comparisons between groups) by **blinding**.
7. **Validity** is the degree to which a measurement represents the phenomena it is intended to measure; it is commonly used for more abstract and subjective variables, and is assessed by **content validity**, **face validity**, **construct validity**, **predictive validity**, and **criterion-related validity**.
8. Individual measurements should be **sensitive**, **specific**, **appropriate**, and **objective**, and they should produce a **range of values**. In the aggregate, they should be **broad** but **parsimonious**, serving the research question at moderate cost in time and money.
9. Investigators should consider **storing images** and other **materials** for later measurements that can take advantage of **new technologies** as they are developed and the efficiency of **nested case–control** designs.

APPENDIX 4

OPERATIONAL DEFINITION OF A MEASUREMENT OF GRIP STRENGTH

The **operations manual** describes the method for conducting and recording the results of all the measurements made in the study. This example, from the operations manual of the Study of Osteoporotic Fractures, describes the use of a dynamometer to measure grip strength. To standardize instructions from examiner to examiner and from subject to subject, the protocol includes a script of instructions to be read to the participant verbatim.

PROTOCOL FOR MEASURING GRIP STRENGTH WITH THE DYNAMOMETER

Grip strength will be measured in both hands. The handle should be adjusted so that the participant holds the dynamometer comfortably. Place the dynamometer in the right hand with the dial facing the palm. The participant's arm should be flexed 90° at the elbow with the forearm parallel to the floor.

1. Demonstrate the test to the subject. While demonstrating, use the following description: "This device measures your arm and upper body strength. We will measure your grip strength in both arms. I will demonstrate how it is done. Bend your elbow at a 90° angle, with your forearm parallel to the floor. Don't let your arm touch the side of your body. Lower the device and squeeze as hard as you can while I count to three. Once your arm is fully extended, you can loosen your grip."
2. Allow one practice trial for each arm, starting with the right if she is right handed. On the second trial, record the kilograms of force from the dial to the nearest 0.5 kg.
3. Reset the dial. Repeat the procedure for the other arm.

The arm should not contact the body. The gripping action should be a slow, sustained squeeze rather than an explosive jerk.

REFERENCES

1. Michaelsson K, Baron JA, Snellman G, et al. Plasma vitamin D and mortality in older men: a community-based prospective cohort study. *Am J Clin Nutr* 2010;92:841–848.
2. Ware JE, Gandek B Jr. Overview of the SF-36 health survey and the International Quality of Life Assessment Project. *J Clin Epidemiol* 1998;51:903–912.
3. Bland JM, Altman DG. Measurement error and correlation coefficients. *BMJ* 1996;313:41–42; also, Measurement error proportional to the mean. *BMJ* 1996;313:106.
4. Newman TB, Kohn M. Evidence-based diagnosis. New York: Cambridge University Press, 2009.
5. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37– 46.
6. Filion K, Kukanich KS, Chapman B, et al. Observation-based evaluation of hand hygiene practices and the effects of an intervention at a public hospital cafeteria. *Am J Infect Control* 2011;39:464–470.
7. Peralta CA, Shlipak MG, Judd S, et al. Detection of chronic kidney disease with creatinine, cystatin C, and urine albumin-to-creatinine ratio and association with progression to end-stage renal disease and mortality. *JAMA* 2011;305:1545–1552.
8. Guttmacher AE, Collins FS. Genomic medicine: a primer. *NEJM* 2002;347:1512–1520.
9. Healy DG. Case–control studies in the genomic era: a clinician's guide. *The Lancet Neurology* 2006;5:701–707.

CHAPTER 5

Getting Ready to Estimate Sample Size: Hypotheses and Underlying Principles

Warren S. Browner, Thomas B. Newman, and Stephen B. Hulley

After an investigator has decided whom and what she is going to study and the design to be used, she must decide how many subjects to sample. Even the most rigorously executed study may fail to answer its research question if the sample size is too small. On the other hand, a study with too large a sample will be more difficult and costly than necessary. The goal of sample size planning is to estimate an **appropriate number** of subjects for a given study design.

Although a useful guide, sample size calculations give a deceptive impression of statistical objectivity. They are only as accurate as the data and estimates on which they are based, which are often just informed guesses. **Sample size planning** is best thought of as a mathematical way of making a ballpark estimate. It often reveals that the research design is not feasible or that different predictor or outcome variables are needed. Therefore, sample size should be estimated early in the design phase of a study, when major changes are still possible.

Before setting out the specific approaches to calculating sample size for several common research designs in Chapter 6, we will spend some time considering the **underlying principles**. Readers who find some of these principles confusing will enjoy discovering that sample size planning does not require their total mastery. However, just as a recipe makes more sense if the cook is somewhat familiar with the ingredients, sample size calculations are easier if the investigator is acquainted with the basic concepts. Even if you plan to ask a friendly biostatistician to calculate your study's sample size, having some understanding of how the process works will allow you to participate more actively in considering the assumptions and estimates involved in the calculation.

HYPOTHESES

The process begins by restating your research question as a **research hypothesis** that summarizes the main elements of the study—the sample, and the predictor and outcome variables. For example, suppose your research question is whether people who do crossword puzzles are less likely to develop dementia. Your research hypothesis would need to specify the sample (for example, people living in a retirement community who have normal cognitive function), the predictor variable (doing crossword puzzles at least once a week on average), and the outcome variable (an abnormal score on a standard test of cognitive function after two years of follow-up).

Hypotheses per se are not needed in descriptive studies that describe how characteristics are distributed in a population, such as the prevalence of abnormal cognitive function in the retirement community. (This does not mean, however, that you won't need to do a sample size estimate for a descriptive study, just that the methods for doing so, described in Chapter 6, are different.) Hypotheses are needed for studies that will use tests of statistical significance to compare findings among groups, such as whether elderly people who do crossword puzzles

regularly are less likely to become demented. Because most observational studies and all experiments address research questions that involve making comparisons, most studies need to specify at least one hypothesis. If any of the following terms appear in the research question, then the study is not simply descriptive and a research hypothesis should be formulated: greater than, less than, more likely than, associated with, compared with, related to, similar to, correlated with, causes, and leads to.

Characteristics of a Good Research Hypothesis

A good hypothesis must be based on a good research question. It should also be simple, specific, and stated in advance.

Simple Versus Complex

A **simple hypothesis** contains one predictor and one outcome variable:

> Among patients with Type II diabetes, a sedentary lifestyle is associated with an increased risk of developing proteinuria.

A **complex hypothesis** contains more than one predictor variable:

> Among patients with Type II diabetes, a sedentary lifestyle and alcohol consumption are associated with an increased risk of developing proteinuria.

Or more than one outcome variable:

> Among patients with Type II diabetes, alcohol consumption is associated with increased risks of developing proteinuria and neuropathy.

Complex hypotheses like these are not readily tested with a single statistical test and are more easily approached as two or more simple hypotheses. Sometimes, however, a combined predictor or outcome variable can be used:

> Among patients with Type II diabetes, alcohol consumption is associated with an increased risk of developing a microvascular complication (i.e., proteinuria, neuropathy, or retinopathy).

In this last example the investigator has decided that what matters is whether a participant has a complication, not what type of complication occurs.

Specific Versus Vague

A **specific hypothesis** leaves no ambiguity about the subjects and variables or about how the test of statistical significance will be applied. It uses concise operational definitions that summarize the nature and source of the subjects and how variables will be measured:

> Prior use of tricyclic antidepressant medications for at least 6 weeks is more common in patients hospitalized for myocardial infarction at Longview Hospital than in controls hospitalized for pneumonia.

This is a long sentence, but it communicates the nature of the study in a clear way that minimizes any opportunity for testing something a little different once the study findings have been examined. It would be incorrect to substitute, during the analysis phase of the study, a different measurement of the predictor, such as the self-reported depression, without considering the issue of multiple hypothesis testing (a topic we discuss at the end of the chapter). Usually, to keep the research hypothesis concise, some of these details are made explicit in the study plan rather than being stated in the research hypothesis. But they should always be clear in the investigator's conception of the study, and spelled out in the protocol.

It is often obvious from the research hypothesis whether the predictor variable and the outcome variable are dichotomous, continuous, or categorical. If it is not clear, then the type of variables can be specified:

> Among non-obese men 35 to 59 years of age, at least weekly participation in a bowling league is associated with a increased risk of developing obesity (body mass index > 30 kg/m^2) during 10 years of follow-up.

Again, if the research hypothesis gets too cumbersome, the definitions can be left out, as long as they are clarified elsewhere.

In-Advance Versus After-the-Fact

The hypothesis should be stated in writing at the outset of the study. This will keep the research effort focused on the primary objective. A single prestated hypothesis also creates a stronger basis for interpreting the study results than several hypotheses that emerge as a result of inspecting the data. Hypotheses that are formulated after examination of the data are a form of multiple hypothesis testing that can lead to overinterpreting the importance of the findings.

The Null and Alternative Hypotheses

Warning: If you have never had any formal training in statistics, or you have forgotten what you did learn, the next few paragraphs may not make sense the first time(s) you read them. Try to work through the terminology even if it seems cumbersome or silly.

The process begins by restating the research hypothesis to one that proposes no difference between the groups being compared. This restatement, called the **null hypothesis**, will become the formal basis for testing statistical significance when you analyze your data at the end of the study. By assuming that there really is no association in the population, statistical tests will help to estimate the probability that an association observed in a study may be due to chance.

For example, suppose your research question is whether drinking unpurified tap water is associated with an increased risk of developing peptic ulcer disease (perhaps because of a greater likelihood of *H. pylori* contamination). Your null hypothesis—that there is no association between the predictor and outcome variables in the population—would be:

> People in Phnom Penh who drink tap water have the *same risk* of developing peptic ulcer disease as those who drink bottled water.

The proposition that there is an association ("People in Phnom Penh who drink tap water have a greater risk of developing peptic ulcer disease than those who drink bottled water.") is called the **alternative hypothesis**. The alternative hypothesis cannot be tested directly; it is accepted by default if the test of statistical significance rejects the null hypothesis (see later).

Another piece of confusing terminology is needed. The alternative hypothesis can be either one-sided or two-sided. A **one-sided alternative hypothesis** specifies the direction of the association between the predictor and outcome variables. The hypothesis that drinking tap water increases the risk of peptic ulcer disease (compared with bottled water) is a one-sided hypothesis. A **two-sided alternative hypothesis** states only that there is an association; it does not specify the direction. For example, "Drinking tap water is associated with a different risk of peptic ulcer disease—either increased or decreased—than drinking bottled water."

One-sided hypotheses may be appropriate in selected circumstances, such as when only one direction for an association is clinically important or biologically meaningful. An example is the one-sided hypothesis that a new drug for hypertension is more likely to cause rashes than a placebo; the possibility that the drug causes fewer rashes than the placebo is not usually worth testing (however, it might be if the drug had anti-inflammatory properties!). A one-sided hypothesis may also be appropriate when there is very strong evidence from prior studies that an association is unlikely to occur in one of the two directions, such as a study to test whether

cigarette smoking affects the risk of brain cancer. Because smoking has been associated with an increased risk of many different types of cancers, a one-sided alternative hypothesis (e.g., that smoking increases the risk of brain cancer) might suffice. However, investigators should be aware that many well-supported hypotheses (e.g., that β-carotene therapy will reduce the risk of lung cancer, or that treatment with drugs that reduce the number of ventricular ectopic beats will reduce sudden death among patients with ventricular arrhythmias) turn out to be wrong when tested in randomized trials. Indeed, in these two examples, the results of well-done trials revealed a statistically significant effect that was opposite in direction from the one the investigators hoped to find (1–3). Overall, we believe that most alternative hypotheses should be two-sided.

It is important to keep in mind the difference between the research hypothesis, which is usually one-sided, and the alternative hypothesis that is used when planning sample size, which is almost always two-sided. For example, suppose the research hypothesis is that recurrent use of antibiotics during childhood is associated with an increased risk of inflammatory bowel disease. That hypothesis specifies the direction of the anticipated effect, so it is one-sided. Why use a two-sided alternative hypothesis when planning the sample size? The answer is that most of the time, both sides of the alternative hypothesis (i.e., greater risk or lesser risk) are interesting, and the investigators would want to publish the results no matter which direction was observed in the study. Statistical rigor requires the investigator to choose between one- and two-sided hypotheses before analyzing the data; switching from a two-sided to a one-sided alternative hypothesis to reduce the *P* value (see below) is not correct. In addition—and this is probably the real reason that two-sided alternative hypotheses are much more common—most grant and manuscript reviewers expect two-sided hypotheses and are critical of a one-sided approach.

UNDERLYING STATISTICAL PRINCIPLES

A research hypothesis, such as 15 minutes or more of exercise per day is associated with a lower mean fasting blood glucose level in middle-aged women with diabetes, is either true or false in the real world. Because an investigator cannot study all middle-aged women with diabetes, she must test the hypothesis in a sample of that target population. As noted in Figure 1.5, there will always be a need to draw inferences about phenomena in the population from events observed in the sample. Unfortunately, by chance alone, sometimes what happens in a sample does not reflect what would have happened if the entire population had been studied.

In some ways, the investigator's problem is similar to that faced by a jury judging a defendant (Table 5.1). The absolute truth about whether the defendant committed the crime cannot usually be determined. Instead, the jury begins by presuming innocence: The defendant did not commit the crime. The jury must then decide whether there is sufficient evidence to **reject the presumed innocence** of the defendant; the standard is known as **beyond a reasonable doubt**. A jury can err, however, by convicting an innocent defendant or by failing to convict a guilty one.

In similar fashion, the investigator starts by presuming the null hypothesis of no association between the predictor and outcome variables in the population. Based on the data collected in her sample, she uses statistical tests to determine whether there is sufficient evidence to **reject the null hypothesis** in favor of the alternative hypothesis that there is an association in the population. The standard for these tests is known as the **level of statistical significance**.

Type I and Type II Errors

Like a jury, an investigator may reach a wrong conclusion. Sometimes by chance alone a sample is not representative of the population and the results in the sample do not reflect reality in the population, leading to an erroneous inference. A **type I error** (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a **type II error** (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the

TABLE 5.1 THE ANALOGY BETWEEN JURY DECISIONS AND STATISTICAL TESTS

JURY DECISION	STATISTICAL TEST
Innocence: The defendant did not counterfeit money.	**Null hypothesis:** There is no association between dietary carotene and the incidence of colon cancer in the population.
Guilt: The defendant did counterfeit money.	**Alternative hypothesis:** There is an association between dietary carotene and the incidence of colon cancer.
Standard for rejecting innocence: Beyond a reasonable doubt.	**Standard for rejecting null hypothesis:** Level of statistical significance (α).
Correct judgment: Convict a counterfeiter.	**Correct inference:** Conclude that there is an association between dietary carotene and colon cancer when one does exist in the population.
Correct judgment: Acquit an innocent person.	**Correct inference:** Conclude that there is no association between carotene and colon cancer when one does not exist.
Incorrect judgment: Convict an innocent person.	**Incorrect inference (type I error):** Conclude that there is an association between dietary carotene and colon cancer when there actually is none.
Incorrect judgment: Acquit a counterfeiter.	**Incorrect inference (type II error):** Conclude that there is no association between dietary carotene and colon cancer when there actually is one.

population. Although type I and type II errors can never be avoided entirely, the investigator can reduce their likelihood by increasing the sample size (the larger the sample, the less likely that it will differ substantially from the population) or by adjusting the design or the measurements in other ways that we will discuss.

In this chapter and the next, we deal only with ways to reduce type I and type II errors due to **chance** variation, also known as random error. False-positive and false-negative results can also occur because of **bias**, but errors due to bias are not usually referred to as type I and type II errors. Such errors are troublesome, because they may be difficult to detect and cannot usually be quantified using statistical methods or avoided by increasing the sample size. (See Chapters 1, 3, 4, and 7–12 for ways to reduce errors due to bias.)

Effect Size

The likelihood that a study will be able to detect an association between a predictor and an outcome variable in a sample depends on the actual magnitude of that association in the population. If it is large (e.g., a 20 mg/dL difference in fasting glucose), it will be easy to detect in the sample. Conversely, if the size of the association is small (a difference of 2 mg/dL), it will be hard to detect in the sample.

Unfortunately, the investigator almost never knows the exact size of the association; one of the purposes of the study is to estimate it! Instead, the investigator must choose the size of the association in the population that she wishes to detect in the sample. That quantity is known as the **effect size**. Selecting an appropriate effect size is the most difficult aspect of sample size planning (4). The investigator should try to find data from prior studies in related areas to make an informed guess about a reasonable effect size. Alternatively, she can choose the smallest effect size that in her opinion would be clinically meaningful (say, a 10 mg/dL reduction in the fasting glucose level).

Of course, from the public health point of view, even a reduction of 2 or 3 mg/dL in fasting glucose levels might be important, especially if it was easy to achieve. The choice of the effect size is always arbitrary, and considerations of feasibility are often paramount. Indeed, when

the number of available or affordable subjects is limited, the investigator may have to work backward (Chapter 6) to determine the effect size she will be able to detect, given the number of subjects she is able to study.

Many studies have several effect sizes, because they measure several different predictor and outcome variables. When designing a study, the sample size should be determined using the desired effect size for the most important hypothesis; the detectable effect sizes for the other hypotheses can then be estimated. If there are several hypotheses of similar importance, then the sample size for the study should be based on whichever hypothesis needs the largest sample.

α, β, and Power

After a study is completed, the investigator uses statistical tests to try to reject the null hypothesis in favor of its alternative, in much the same way that a prosecuting attorney tries to convince a jury to reject innocence in favor of guilt. Depending on whether the null hypothesis is true or false in the target population, and assuming that the study is free of bias, four situations are possible (Table 5.2). In two of these, the findings in the sample and reality in the population are concordant, and the investigator's inference will be correct. In the other two situations, either a type I or type II error has been made, and the inference will be incorrect.

The investigator establishes the maximum chance that she will tolerate of making type I and type II errors in advance of the study. The maximum probability of committing a type I error (rejecting the null hypothesis when it is actually true) is called α (**alpha**). Another name for α is the **level of statistical significance**.

If, for example, a study of the effects of exercise on fasting blood glucose levels is designed with an α of 0.05, then the investigator has set 5% as the maximum chance of incorrectly rejecting the null hypothesis if it is true (and inferring that exercise and fasting blood glucose levels are associated in the population when, in fact, they are not). This is the level of reasonable doubt that the investigator will be willing to accept when she uses statistical tests to analyze the data after the study is completed.

The probability of making a type II error (failing to reject the null hypothesis when it is actually false) is called β (**beta**). The quantity $[1 - \beta]$ is called **power**, the probability of correctly rejecting the null hypothesis in the sample if the actual effect in the population is equal to (or greater than) the specified effect size.

If β is set at 0.10, then the investigator has decided that she is willing to accept a 10% chance of missing an association of the specified effect size if it exists. This represents a power of 0.90; that is, a 90% chance of finding an association of that size or greater. For example, suppose that exercise really does lead to an average reduction of 20 mg/dL in fasting glucose levels among diabetic women in the population. If the investigator replicated the study with the same 90% power on numerous occasions, we would expect that in 9 of 10 studies she would correctly reject the null hypothesis at the specified level of alpha (0.05) and conclude that exercise is associated with fasting glucose level. This does not mean that the investigator would be unable to detect a smaller effect in the population, say, a 15 mg/dL reduction; it means simply that she will have less than a 90% likelihood of doing so.

TABLE 5.2 TRUTH IN THE POPULATION VERSUS THE RESULTS IN THE STUDY SAMPLE: THE FOUR POSSIBILITIES

	TRUTH IN THE POPULATION	
RESULTS IN THE STUDY SAMPLE	ASSOCIATION BETWEEN PREDICTOR AND OUTCOME	NO ASSOCIATION BETWEEN PREDICTOR AND OUTCOME
Reject null hypothesis	Correct	Type I error
Fail to reject null hypothesis	Type II error	Correct

Ideally, α and β would be set close to zero, minimizing the possibility of false-positive and false-negative results. Reducing them, however, requires increasing the sample size or one of the other strategies discussed in Chapter 6. Sample size planning aims at choosing a sufficient number of subjects to keep α and β at an acceptably low level without making the study unnecessarily expensive or difficult.

Many studies set α at 0.05 and β at 0.20 (a power of 0.80). These are arbitrary values, and others are sometimes used: The conventional range for α is between 0.01 and 0.10, and that for β is between 0.05 and 0.20. In general, the investigator should use a low α when the research question makes it particularly important to avoid a type I (false-positive) error—for example, in testing the efficacy of a potentially dangerous medication. She should use a low β (and a small effect size) when it is especially important to avoid a type II (false-negative) error—for example, in reassuring the public that living near a toxic waste dump is safe.

P Value

Now it's time to return to the **null hypothesis**, whose underlying purpose will finally become clear. The null hypothesis has only one function: to act like a straw man. It is assumed to be true so that it can be rejected as false with a statistical test. When the data are analyzed, a statistical test is used to determine the ***P* value**, which is the probability of seeing—by chance alone—an effect as big as or bigger than that seen in the study if the null hypothesis actually were true. The key insight is to recognize that if the null hypothesis is true, and there really is no difference in the population, then the only way that the study could have found one in the sample is by chance.

If that chance is small, then the null hypothesis of no difference can be rejected in favor of its alternative, that there is a difference. By "small" we mean a *P* value that is less than α, the predetermined level of statistical significance.

However, a "**nonsignificant**" **result** (i.e., one with a *P* value greater than α) does not mean that there is no association in the population; it only means that the result observed in the sample is small compared with what could have occurred by chance alone. For example, an investigator might find that women who played intercollegiate sports were twice as likely to undergo total hip replacements later in life as those who did not, but because the number of hip replacements in the study was modest this apparent effect had a *P* value of only 0.08. This means that even if athletic activity and hip replacement were not associated in the population, there would be an 8% probability of finding an association at least as large as the one observed by the investigator *by chance alone*. If the investigator had set the significance level as a two-sided α of 0.05, she would have to conclude that the association in the sample was "not statistically significant."

It might be tempting for the investigator to change her mind and switch to a *one*-sided *P* value and report it as "$P = 0.04$." A better choice would be to report her results with the 95% confidence interval and comment that "These results, although suggestive of an association, did not achieve statistical significance ($P = 0.08$)." This solution preserves the integrity of the original two-sided hypothesis design, and also acknowledges that statistical significance is not an all-or-none situation.

Sides of the Alternative Hypothesis

Recall that an alternative hypothesis actually has two sides, either or both of which can be tested in the sample by using **one-** or **two-sided**[1] **statistical tests**. When a two-sided statistical test is used, the *P* value includes the probabilities of committing a type I error in each of the two directions, which is about twice as great as the probability in either direction alone. It is

[1]These are sometimes referred to as one- and two-*tailed* tests, after the tails (extreme areas) of statistical distributions.

easy to convert from a one-sided *P* value to a two-sided *P* value, and vice versa. A one-sided *P* value of 0.05, for example, is usually the same as a two-sided *P* value of 0.10. (Some statistical tests are asymmetric, which is why we said "usually.")

In those rare situations in which an investigator is only interested in one of the sides of the alternative hypothesis (e.g., a noninferiority trial designed to determine whether a new antibiotic is no less effective than one in current use; see Chapter 11), sample size can be calculated accordingly. A one-sided hypothesis, however, should never be used just to reduce the sample size.

Type of Statistical Test

The formulas used to calculate sample size are based on mathematical assumptions, which differ for each statistical test. Before the sample size can be calculated, the investigator must decide on the statistical approach to analyzing the data. That choice depends mainly on the type of predictor and outcome variables in the study. Table 6.1 lists some common statistics used in data analysis, and Chapter 6 provides simplified approaches to estimating sample size for studies that use these statistics.

■ ADDITIONAL POINTS

Variability

It is not simply the size of an effect that is important; its **variability** also matters. Statistical tests depend on being able to show a difference between the groups being compared. The greater the variability (or spread) in the outcome variable among the subjects, the more likely it is that the values in the groups will overlap, and the more difficult it will be to demonstrate an overall difference between them. Because measurement error contributes to the overall variability, less precise measurements require larger sample sizes (5).

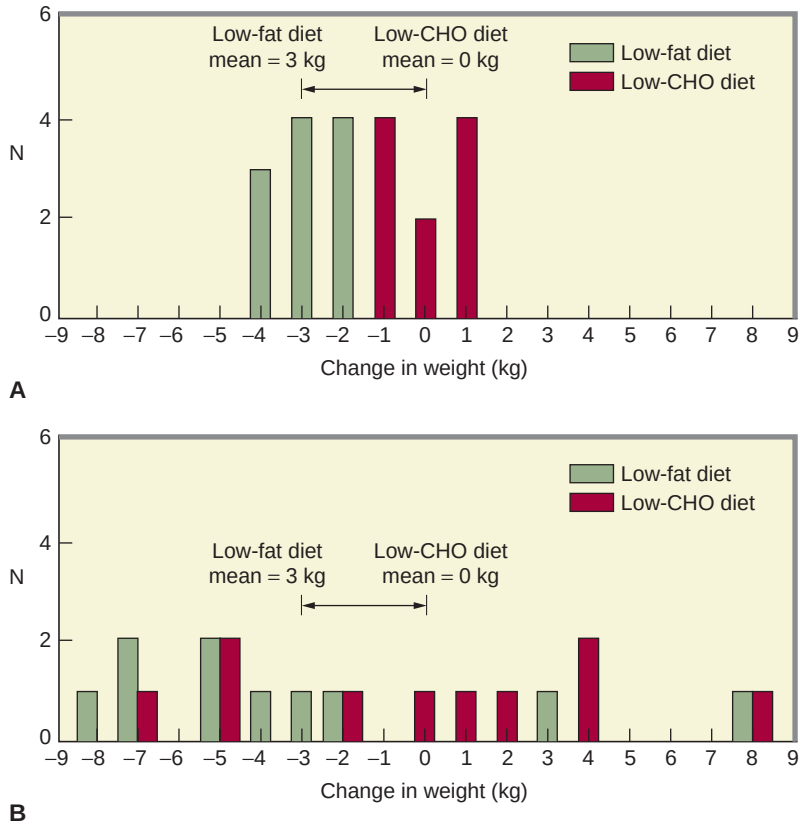
Consider a study of the effects of two diets (low fat and low carbohydrate) in achieving weight loss in 20 obese patients. If all those on the low-fat diet lost about 3 kg and all those on the low-carbohydrate diet lost little if any weight (an effect size of 3 kg), it is likely that the low-fat diet really is better (Figure 5.1A). On the other hand, if the average weight loss were 3 kg in the low-fat group and 0 kg in the low-carbohydrate group, but there was a great deal of overlap between the two groups (the situation in Figure 5.1B), the greater variability would make it more difficult to detect a difference between the diets, and a larger sample size would be needed.

When one of the variables used in the sample size estimate is continuous (e.g., body weight in Figure 5.1), the investigator will need to estimate its variability. (See the section on the *t* test in Chapter 6 for details.) In the other situations, variability is already included in the other parameters entered into the sample size formulas and tables, and need not be specified.

Multiple and *Post Hoc* Hypotheses

When more than one hypothesis is tested in a study, especially if some of those hypotheses were formulated after the data were analyzed (*post hoc* hypotheses), the likelihood that at least one will achieve statistical significance on the basis of chance alone increases. For example, if 20 independent hypotheses are tested at an α of 0.05, the likelihood is substantial (64%; $[1 - 0.95^{20}]$) that at least one hypothesis will be statistically significant by chance alone. Some statisticians advocate adjusting the level of statistical significance when more than one hypothesis is tested in a study. This keeps the overall probability of accepting any one of the alternative hypotheses, when all the findings are due to chance, at the specified level. For example, genomic studies that look for an association between thousands of genotypes and a disease need to use a much smaller α than 0.05, or they risk identifying many false-positive associations.

One approach, named after the mathematician **Bonferroni**, is to divide the significance level (say, 0.05) by the number of hypotheses tested. If there were four hypotheses, for example,



■ **FIGURE 5.1** **A:** Weight loss achieved by two diets. All subjects on the low-fat diet lost from 2 to 4 kg, whereas weight change in those on the low-carbohydrate (CHO) diet varied from –1 to +1 kg. Because there is no overlap between the two groups, it is reasonable to infer that the low-fat diet is better at achieving weight loss than the low-carbohydrate diet (as would be confirmed with a *t* test, which has a *P* value < 0.0001). **B:** Weight loss achieved by two diets. There is substantial overlap in weight change in the two groups. Although the effect size is the same (3 kg) as in **A,** there is little evidence that one diet is better than the other (as would be confirmed with a *t* test, which has a *P* value of 0.19).

each would be tested at an α of 0.0125 (i.e., $0.05 \div 4$). This requires substantially increasing the sample size over that needed for testing each hypothesis at an α of 0.05. Thus, for any particular hypothesis, the Bonferroni approach reduces the chance of a type I error at the cost of either increasing the chance of a type II error or requiring a greater sample size. If the results of a study are still statistically significant after the Bonferroni adjustment, that loss of power is not a problem. However, a result that loses statistical significance after Bonferroni adjustment, which could represent failing to support an association that was actually present in the population (a type II error), is more problematic.

Especially in these cases, the issue of what significance level to use depends more on the **prior probability** of each hypothesis than on the number of hypotheses tested, and for this reason our general view is that the mindless Bonferroni approach to multiple hypothesis testing is often too stringent. There is an analogy with the use of diagnostic tests that may be helpful (6, 7). When interpreting the results of a diagnostic test, a clinician considers the likelihood that the patient being tested has the disease in question. For example, a modestly abnormal test result in a healthy person (a serum alkaline phosphatase level that is 15% greater than the upper limit of normal) is probably a false-positive test that is unlikely to have much clinical importance. Similarly, a *P* value of 0.05 for an unlikely hypothesis is probably also a false-positive result.

However, an alkaline phosphatase level that is 10 or 20 times greater than the upper limit of normal is unlikely to have occurred by chance (although it might be a laboratory error). So too a very small *P* value (say, < 0.001) is unlikely to have occurred by chance (although it could be due to bias). It is hard to dismiss very abnormal test results as being false-positives or to dismiss very low *P* values as being due to chance, even if the prior probability of the disease or the hypothesis was low.[2]

Moreover, the number of tests that were ordered, or hypotheses that were tested, is not always relevant. The interpretation of an elevated serum uric acid level in a patient with a painful and swollen joint should not depend on whether the physician ordered just a single test (the uric acid level) or obtained the result as part of a panel of 20 tests. Similarly, when interpreting the *P* value for testing a research hypothesis that makes good sense, it should not matter that the investigator also tested several unlikely hypotheses. What matters most is the reasonableness of the research hypothesis being tested: that it has a substantial **prior probability** of being correct. (Prior probability, in this "**Bayesian**" approach, is usually a subjective judgment based on evidence from other sources.) Hypotheses that are formulated during the design of a study usually meet this requirement; after all, why else would the investigator put the time and effort into planning and doing the study?

What about unanticipated associations that appear during the collection and analysis of a study's results? This process is sometimes called **hypothesis generation** or, less favorably, "data-mining" or a "fishing expedition." The many informal comparisons that are made during data analysis are a form of multiple hypothesis testing. A similar problem arises when variables are redefined during data analysis, or when results are presented for subgroups of the sample. Significant *P* values for data-generated hypotheses that were not considered during the design of the study are all too often due to chance. They should be viewed with skepticism, and considered a source of potential research questions for future studies.

Sometimes, however, an investigator fails to specify a particular hypothesis in advance, although that hypothesis seems reasonable when it is time for the data to be analyzed. This might happen, for example, if others discover a new risk factor while the study is going on, or if the investigator just didn't happen to think of a particular hypothesis when the study was being designed. The important issue is not so much whether the hypothesis was formulated before the study began, but whether there is a reasonable prior probability based on evidence from other sources that the hypothesis is true (6, 7).

There are some definite advantages to formulating more than one hypothesis when planning a study. The use of **multiple unrelated hypotheses** increases the efficiency of the study, making it possible to answer more questions with a single research effort and to discover more of the true associations that exist in the population. It may also be a good idea to formulate several *related* hypotheses; if the findings are consistent, the study conclusions are made stronger. Studies in patients with heart failure have found that the use of angiotensin-converting enzyme inhibitors is beneficial in reducing cardiac admissions, cardiovascular mortality, and total mortality. Had only one of these hypotheses been tested, the inferences from these studies would have been less definitive. Lunch may not be free, however, when multiple hypotheses are tested. Suppose that when these related and prestated hypotheses are tested, only one turns out to be statistically significant. Then the investigator must decide (and try to convince editors and readers) whether the significant results, the nonsignificant results, or both sets of results are correct.

Primary and Secondary Hypotheses

Some studies, especially large randomized trials, specify some hypotheses as being "**secondary**." This usually happens when there is one **primary hypothesis** around which the study has been

[2]Again, the exception is some genetic studies, in which millions or even billions of associations may be examined.

designed, but the investigators are also interested in other research questions that are of lesser importance. For example, the primary outcome of a trial of zinc supplementation might be hospitalizations or emergency department visits for upper respiratory tract infections; a secondary outcome might be self-reported days missed from work or school. If the study is being done to obtain approval for a pharmaceutical agent, the primary outcome is what will matter most to the regulatory body. Stating a secondary hypothesis in advance does increase the credibility of the results when that hypothesis is tested.

A good rule, particularly for clinical trials, is to establish in advance as many hypotheses as make sense, but specify just one as the **primary hypothesis**, which can be tested statistically without argument about whether to adjust for multiple hypothesis testing. More important, having a primary hypothesis helps to focus the study on its main objective and provides a clear basis for the main sample size calculation.

Many statisticians and epidemiologists are moving away from hypothesis testing, with its emphasis on *P* values, to using confidence intervals to report the precision of the study results (8–10). Indeed, some authors believe the entire process of basing sample size planning on hypotheses is misleading, in part because it depends on quantities that are either unknown (effect size) or arbitrary (α and β) (11). However, the approach we have outlined is a practical one, and remains standard in clinical research planning.

SUMMARY

1. **Sample size planning** is an important part of the design of both analytic and descriptive studies. The sample size should be estimated early in the process of developing the research design, so that appropriate modifications can be made.
2. Analytic studies and experiments need a **hypothesis** that specifies, for the purpose of subsequent **statistical tests**, the anticipated association between the main predictor and outcome variables. Purely descriptive studies, lacking the strategy of comparison, do not require a hypothesis.
3. Good hypotheses are **specific** about how the population will be sampled and the variables measured, **simple** (there is only one predictor and one outcome variable), and **formulated in advance**.
4. The **null hypothesis**, which proposes that the predictor variable is not associated with the outcome, is the basis for tests of statistical significance. The **alternative hypothesis** proposes that they are associated. Statistical tests attempt to reject the null hypothesis of no association in favor of the alternative hypothesis that there is an association.
5. An alternative hypothesis is either **one-sided** (only one direction of association will be tested) or **two-sided** (both directions will be tested). One-sided hypotheses should only be used in unusual circumstances, when only one direction of the association is clinically or biologically meaningful.
6. For analytic studies and experiments, the sample size is an estimate of the number of subjects required to detect an association of a given **effect size** and **variability** at a specified likelihood of making **type I** (false-positive) and **type II** (false-negative) **errors**. The maximum likelihood of making a type I error is called α; that of making a type II error, β. The quantity $(1 - \beta)$ is **power**, the chance of observing an association of a given effect size or greater in a sample if one is actually present in the population.
7. It is often desirable to establish more than one hypothesis in advance, but the investigator should specify a single **primary hypothesis** as a focus and for sample size estimation. Interpretation of findings from testing **multiple hypotheses** in the sample, including unanticipated findings that emerge from the data, is based on a judgment about the **prior probability** that they represent real phenomena in the population.

REFERENCES

1. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N Engl J Med* 1994;330:1029–1035.
2. Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *N Engl J Med* 1991;324:781–788.
3. The Cardiac Arrhythmia Suppression Trial II Investigators. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med* 1992;327:227–233.
4. Van Walraven C, Mahon JL, Moher D, et al. Surveying physicians to determine the minimal important difference: implications for sample-size calculation. *J Clin Epidemiol* 1999;52:717–723.
5. McKeown-Eyssen GE, Tibshirani R. Implications of measurement error in exposure for the sample sizes of case-control studies. *Am J Epidemiol* 1994;139:415–421.
6. Browner WS, Newman TB. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. *JAMA* 1987;257:2459–2463.
7. Newman TB, Kohn, MA. *Evidence-based diagnosis*. New York: Cambridge University Press, 2009. Chapter 11.
8. Daly LE. Confidence limits made easy: interval estimation using a substitution method. *Am J Epidemiol* 1998;147:783–790.
9. Goodman SN. Toward evidence-based medical statistics. 1: The *P* value fallacy. *Ann Intern Med* 1999;130:995–1004.
10. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* 1999;130:1005–1013.
11. Bacchetti P. Current sample size conventions: flaws, harms, and alternatives. *BMC Med*. 2010;8:17.

6

Estimating Sample Size and Power: Applications and Examples

Warren S. Browner, Thomas B. Newman, and Stephen B. Hulley

Chapter 5 introduced the basic principles underlying sample size calculations. This chapter presents several cookbook techniques for using those principles to estimate the sample size needed for a research project. The first section deals with **sample size estimates for an analytic study or experiment**, including some special issues that apply to these studies such as multivariate analysis. The second section considers **studies that are primarily descriptive**. Subsequent sections deal with studies that have a **fixed sample size**, strategies for **maximizing the power** of a study, and how to estimate the sample size when there appears to be **insufficient information** from which to work. The chapter concludes with **common errors** to avoid.

At the end of the chapter, there are tables and formulas in the appendixes for several basic methods of estimating sample size. In addition, there is a calculator on our website (www.epibiostat.ucsf.edu/dcr/), and there are many sites on the Web that can provide instant interactive sample size calculations; try searching for "sample size calculator." Most statistical packages can also estimate sample size for common study designs.

SAMPLE SIZE TECHNIQUES FOR ANALYTIC STUDIES AND EXPERIMENTS

There are several variations on the recipe for estimating sample size in an analytic study or experiment, but they all have certain steps in common:

1. State the **null hypothesis** and either a **one-** or **two-sided alternative hypothesis**.
2. Select the appropriate **statistical test** from Table 6.1 based on the type of predictor variable and outcome variable in those hypotheses.
3. Choose a reasonable **effect size** (and **variability**, if necessary).
4. Set α and β. Specify a two-sided α unless the alternative hypothesis is clearly one-sided.
5. Use the appropriate table or formula in the appendix, an online calculator, or a statistics package to estimate the sample size.

Even if the exact value for one or more of the ingredients is uncertain, it is important to estimate the sample size early in the design phase. Waiting until the last minute to prepare the sample size can lead to a rude awakening: It may be necessary to start over with new ingredients, which may mean redesigning the entire study. This is why this subject is covered early in this book.

Not all analytic studies fit neatly into one of the three main categories of sample size calculation described in the following sections: use of the chi-squared test if both predictor and outcome are dichotomous, use of the *t* test if one is dichotomous and the other continuous, and use of the correlation coefficient if both are continuous. A few of the more common exceptions are discussed in the section called "Other Considerations and Special Issues" (page 60).

TABLE 6.1 SIMPLE STATISTICAL TESTS FOR USE IN ESTIMATING SAMPLE SIZE*

	OUTCOME VARIABLE	
PREDICTOR VARIABLE	DICHOTOMOUS	CONTINUOUS
Dichotomous	Chi-squared test†	*t* test
Continuous	*t* test	Correlation coefficient

*See section on "Other Considerations and Special Issues" for what to do about ordinal variables, or if planning to analyze the data with another type of statistical test.

†The chi-squared test is always two-sided; a one-sided equivalent is the *Z* statistic.

The *t* Test

The ***t* test** (sometimes called "Student's *t* test," after the pseudonym of its developer) is commonly used to determine whether the mean value of a continuous variable in one group differs significantly from that in another group. For example, the *t* test would be appropriate to use when comparing the mean depression scores in patients treated with two different antidepressants, or the mean body mass index among subjects who do and do not have diabetes. The *t* test assumes that the distribution (spread) of the variable in each of the two groups approximates a normal (bell-shaped) curve. However, the *t* test is remarkably robust, so it can be used for almost any distribution unless the number of subjects is small (fewer than 30 to 40) or there are extreme outliers.

Although the *t* test is usually used for comparing continuous outcomes, it can also be used to estimate the sample size for a dichotomous outcome (e.g., in a case–control study) if the study has a continuous predictor variable. In this situation, the *t* test compares the mean value of the predictor variable in the cases with that in the controls.

To estimate the sample size for a study in which mean values of a continuous outcome variable will be compared with a *t* test (see Example 6.1), the investigator must

1. State the null hypothesis and whether the alternative hypothesis is one- or two-sided.
2. Estimate the effect size (E) as the difference in the mean value of the continuous variable between the study groups.
3. Estimate variability as the standard deviation (S) of that variable.
4. Calculate the standardized effect size (E/S), defined as the effect size divided by the standard deviation of the outcome variable.
5. Set α and β.

The **effect size** and **variability** can often be estimated from previous studies in the literature and consultation with experts. Occasionally, a small pilot study will be necessary to estimate the standard deviation of the variable (also see the section "How to Estimate Sample Size When There Is Insufficient Information" on page 70). When an outcome variable is the *change* in a continuous measurement (e.g., change in weight during a study), the investigator should use the standard deviation of the *change* in that variable (not the standard deviation of the variable itself) in the sample size estimates. The standard deviation of the change in a variable is usually smaller than the standard deviation of the variable; therefore, the sample size will also be smaller.

Sometimes, an investigator cannot obtain any meaningful information about the standard deviation of a variable. In that situation, it's worthwhile to use a quantity called the **standardized effect size**, which is a unitless quantity that makes it possible to estimate a sample size; it also simplifies comparisons among effect sizes of different variables. The standardized effect size is simply the effect size divided by the standard deviation of the variable. For example, a 10 mg/dL difference in serum cholesterol level, which has a standard deviation in the population of about 40 mg/dL, would equal a standardized effect size of 0.25. The larger the standardized

EXAMPLE 6.1 Sample Size When Using the *t* Test

Problem: The research question is whether there is a difference in the efficacy of albuterol and ipratropium bromide for the treatment of asthma. The investigator plans a randomized trial of the effect of these drugs on FEV_1 (forced expiratory volume in 1 second) after 2 weeks of treatment. A previous study has reported that the mean FEV_1 in persons with treated asthma was 2.0 liters, with a standard deviation of 1.0 liter. The investigator would like to be able to detect a difference of 10% or more in mean FEV_1 between the two treatment groups. How many patients are required in each group (albuterol and ipratropium) at α (two-sided) = 0.05 and power = 0.80?

The ingredients for the sample size calculation are as follows:

1. **Null Hypothesis:** Mean FEV_1 after 2 weeks of treatment is the same in asthmatic patients treated with albuterol as in those treated with ipratropium.
   Alternative Hypothesis (two-sided): Mean FEV_1 after 2 weeks of treatment is different in asthmatic patients treated with albuterol from what it is in those treated with ipratropium.
2. Effect size = 0.2 liters (10% × 2.0 liters).
3. Standard deviation of FEV_1 = 1.0 liter.
4. Standardized effect size = effect size ÷ standard deviation = 0.2 liters ÷ 1.0 liter = 0.2.
5. α (two-sided) = 0.05; $\beta = 1 - 0.80 = 0.20$. (Recall that $\beta = 1 -$ power.)

Looking across from a standardized effect size of 0.20 in the leftmost column of Table 6A and down from α (two-sided) = 0.05 and $\beta = 0.20$, 394 patients are required per group. This is the number of patients in each group who need to complete the study; even more will need to be enrolled to account for dropouts. This sample size may not be feasible, and the investigator might reconsider the study design, or perhaps settle for only being able to detect a larger effect size. See the section on the *t* test for paired samples (Example 6.8) for a potential solution.

effect size, the smaller the required sample size. For most studies, the standardized effect size will be >0.1. Effect sizes smaller than that are difficult to detect (they require very large sample sizes) and usually not very important clinically.

Appendix 6A gives the sample size requirements for various combinations of α and β for several standardized effect sizes. To use Table 6A, look down its leftmost column for the standardized effect size. Next, read across the table to the chosen values for α and β for the sample size required per group. (The numbers in Table 6A assume that the two groups being compared are of the same size; use the formula below the table, a statistics package, or an interactive Web-based program if that assumption is not true.)

There is a convenient **shortcut** for approximating sample size using the *t* test, when more than about 30 subjects will be studied and the power is set at 0.80 ($\beta = 0.2$) and α (two-sided) is set at 0.05 (1). The formula is

$$\text{Sample size (per equal-sized group)} = 16 \div (\text{standardized effect size})^2.$$

For Example 6.1, the shortcut estimate of the sample size would be $16 \div 0.2^2 = 400$ per group.

The Chi-Squared Test

The **chi-squared (χ^2) test** can be used to compare the proportion of subjects in each of two groups who have a dichotomous outcome. For example, the proportion of men who develop coronary heart disease (CHD) while being treated with folate can be compared with the

proportion who develop CHD while taking a placebo. The chi-squared test is always two-sided; an equivalent test for one-sided hypotheses is the **one-sided Z test**.

In an experiment or cohort study, effect size is specified by the difference between P_1, the proportion of subjects expected to have the outcome in one group (i.e., the risk of the outcome), and P_2, the proportion expected in the other group. For example, in a cohort study comparing the risk of developing end-stage renal disease among men and women with hypertension, P_1 would be the proportion of men who develop end-stage renal disease, and P_2 would be the proportion of women who do so. Variability is a function of P_1 and P_2, so it need not be specified.

By contrast, for the purposes of calculating sample size for a case–control study, P_1 and P_2 have different definitions. They refer to the proportions of cases and controls expected to have a particular value of a dichotomous predictor (e.g., the proportion of cases of end-stage renal disease who were men). Thus, in a case–control study, P_1 represents the proportion of cases expected to have a particular predictor variable (i.e., the prevalence of that predictor), and P_2 represents the proportion of controls expected to have the predictor.

To estimate the sample size for a study that will be analyzed with the chi-squared test or Z test to compare two proportions (Example 6.2), the investigator must

1. State the null hypothesis and decide whether the alternative hypothesis should be one- or two-sided.
2. Estimate the effect size and variability in terms of P_1, the proportion with the outcome in one group, and P_2, the proportion with the outcome in the other group.
3. Set α and β.

Appendix 6B gives the sample size requirements for several combinations of α and β, and a range of values of P_1 and P_2. To estimate the sample size, look down the leftmost column of Tables 6B.1 or 6B.2 for the smaller of P_1 and P_2 (if necessary, rounded to the nearest 0.05). Next, read across for the difference between P_1 and P_2. Based on the chosen values for α and β, the table gives the sample size required per group.

EXAMPLE 6.2 Calculating Sample Size When Using the Chi-Squared Test

Problem: The research question is whether subjects who practice Tai Chi have a lower risk of developing back pain than those who jog for exercise. A review of the literature suggests that the 2-year risk of back pain is about 0.30 in joggers. The investigator hopes to be able to show that Tai Chi reduces that risk by at least 0.10. At α (two-sided) = 0.05 and power = 0.80, how many subjects will need to be studied to determine whether the 2-year incidence of developing back pain is 0.20 (or less) in those who do Tai Chi?

Solution: The ingredients for the sample size calculation are as follows:

1. **Null Hypothesis:** The incidence of back pain is the same in those who jog and those who practice Tai Chi.
   Alternative Hypothesis (two-sided): The incidence of back pain differs in those who jog and those who practice Tai Chi.
2. P_2 (incidence in those who jog) = 0.30; P_1 (incidence in those who practice Tai Chi) = 0.20. The smaller of these values is 0.20, and the difference between them ($P_1 - P_2$) is 0.10.
3. α (two-sided) = 0.05; $\beta = 1 - 0.80 = 0.20$.

Looking across from 0.20 in the leftmost column in Table 6B.1 and down from an expected difference of 0.10, the middle number for α (two-sided) = 0.05 and $\beta = 0.20$ is the required sample size of 313 joggers and 313 Tai Chi practitioners to complete the study.

Often the investigator specifies the effect size in terms of the **relative risk** (risk ratio) of the outcome in two groups of subjects. For example, an investigator might study whether women who use oral contraceptives are at least twice as likely as nonusers to have a myocardial infarction. In a cohort study (or experiment), it is straightforward to convert back and forth between relative risk and the two proportions (P_1 and P_2), since the relative risk is just P_1 divided by P_2 (or vice versa).

For a case–control study, however, the situation is a little more complex because the relative risk must be approximated by the **odds ratio (OR)**:

$$\text{OR} = \frac{[P_1 \times (1-P_2)]}{[P_2 \times (1-P_1)]}$$

The investigator must specify the odds ratio (OR) and P_2 (the proportion of controls exposed to the predictor variable). Then P_1 (the proportion of cases exposed to the predictor variable) is

$$P_1 = \frac{\text{OR} \times P_2}{(1-P_2) + (\text{OR} \times P_2)}$$

For example, if the investigator expects that 10% of controls will be exposed to the oral contraceptives ($P_2 = 0.1$) and wishes to detect an odds ratio of 3 associated with the exposure, then

$$P_1 = \frac{(3 \times 0.1)}{(1-0.1) + (3 \times 0.1)} = \frac{0.3}{1.2} = 0.25$$

The Correlation Coefficient

Although the **correlation coefficient (*r*)** is not used frequently in sample size calculations, it can be used when the predictor and outcome variables are both continuous. The correlation coefficient is a measure of the strength of the linear association between the two variables. It varies between −1 and +1. Negative values indicate that as one variable increases, the other decreases (like blood lead level and IQ in children). The closer the absolute value of r is to 1, the stronger the association; the closer to 0, the weaker the association. Height and weight in adults, for example, are highly correlated in some populations, with $r \approx 0.9$. Such high values, however, are uncommon; many biologic associations have much smaller correlation coefficients.

Correlation coefficients are common in some fields of clinical research, such as behavioral medicine, but using them to estimate the sample size has a disadvantage: Correlation coefficients have little intuitive meaning. When squared (r^2), a correlation coefficient represents the proportion of the spread (variance) in an outcome variable that results from its linear association with a predictor variable, and vice versa. That's why small values of r, such as ≤0.3, may be statistically significant if the sample is large enough without being very meaningful clinically or scientifically, since they "explain" at most 9% of the variance.

An alternative—and often preferred—way to estimate the sample size for a study in which the predictor and outcome variables are both continuous is to dichotomize one of the two variables (say, at its median) and use the t test calculations instead. This has the advantage of expressing the effect size as a difference between two groups (interpreting correlation coefficients, which do not convey effect size, is more vague). To estimate sample size for a study that will be analyzed with a correlation coefficient (Example 6.3), the investigator must

1. State the null hypothesis, and decide whether the alternative hypothesis is one or two-sided.
2. Estimate the effect size as the absolute value of the smallest correlation coefficient (r) that the investigator would like to be able to detect. (Variability is a function of r and is already included in the table and formula.)
3. Set α and β.

EXAMPLE 6.3 Calculating Sample Size When Using the Correlation Coefficient in a Cross-Sectional Study

Problem: The research question is whether urinary cotinine levels (a measure of the intensity of current cigarette smoking) are correlated with bone density in smokers. A previous study found a modest correlation ($r = -0.3$) between reported smoking (in cigarettes per day) and bone density (in g/cm^3); the investigator anticipates that urinary cotinine levels will be at least as well correlated. How many smokers will need to be enrolled, at α (two-sided) = 0.05 and $\beta = 0.10$?

Solution: The ingredients for the sample size calculation are as follows:

1. **Null Hypothesis:** There is no correlation between urinary cotinine level and bone density in smokers.
   Alternative Hypothesis: There is a correlation between urinary cotinine level and bone density in smokers.
2. Effect size (r) = $|-0.3| = 0.3$.
3. α (two-sided) = 0.05; $\beta = 0.10$.

Using Table 6C, reading across from $r = 0.30$ in the leftmost column and down from α (two-sided) = 0.05 and $\beta = 0.10$, 113 smokers will be required.

In Appendix 6C, look down the leftmost column of Table 6C for the effect size (r). Next, read across the table to the chosen values for α and β, yielding the total sample size required. Table 6C yields the appropriate sample size when the investigator wishes to reject the null hypothesis that there is no association between the predictor and outcome variables (e.g., $r = 0$). If the investigator wishes to determine whether the correlation coefficient in the study differs from a value other than zero (e.g., $r = 0.4$), she should see the text below Table 6C for the appropriate methodology.

■ OTHER CONSIDERATIONS AND SPECIAL ISSUES

Dropouts

Each sampling unit must be available for analysis; subjects who are enrolled in a study but in whom outcome status cannot be ascertained (such as **dropouts**) do not count in the sample size. If the investigator anticipates that any of her subjects will not be available for follow-up (as is very often the case), she should estimate the proportion that will be lost and increase the size of the enrolled sample accordingly. If, for example, the investigator estimates that 20% of her sample will be lost to follow-up, then the sample size should be increased by a factor of $(1 \div [1 - 0.20])$, or 1.25.

Categorical Variables

While there are mathematical reasons why estimating a sample size for **ordinal variables** using a test may not be appropriate, in practice ordinal variables can often be treated as continuous variables, especially if the number of categories is relatively large (six or more) and if averaging the values of the variable makes sense.

In other situations, the best strategy is to change the research hypothesis slightly by dichotomizing the categorical variable. As an example, suppose a researcher is studying whether speaking English as a second language is associated with the number of times that diabetic patients visit a podiatrist in a year. The number of visits is unevenly distributed: Many people will have no visits, some will make one visit, and only a few will make two or more visits. In this

situation, the investigator could estimate the sample size as if the outcome were dichotomous (no visits versus one or more visits).

Survival Analysis

When an investigator wishes to compare survival or other time-to-event data, such as which of two treatments is more effective in prolonging life in women with advanced breast cancer, survival analysis will be the appropriate technique for analyzing the data (2, 3). Although the outcome variable, say months of survival, *appears* to be continuous, the *t* test is not appropriate because what is actually being assessed is not time (a continuous variable) but the proportion of subjects (a dichotomous variable) still alive at each point in time. Similarly, an investigator might be comparing the rate of developing the outcome (per 100 person-years of follow-up) in two groups. A reasonable approximation can be made by simply estimating the proportions of subjects expected to ever have the outcome in the two groups and estimating the sample size with the chi-squared test. However, if the outcome is expected to occur in most of the subjects, such as death in a study of advanced breast cancer, a better strategy (because it minimizes the total sample size) is to estimate the sample size based on the proportions of subjects in each group who are expected to have the outcome at a point during follow-up when about half of the total outcomes have occurred. For example, in a study comparing disease-free survival in breast cancer patients treated with standard versus experimental treatment, in which about 60% of the subjects in the standard treatment group are expected to have died by 2 years, compared with 40% of those who received an experimental treatment, the sample size can be estimated using "survival at 2 years" as the dichotomous outcome.

Clustered Samples

Some research designs involve the use of **clustered samples**, in which subjects are sampled by groups (Chapter 11). Consider, for example, a study of whether a continuing medical education intervention for clinicians improves the rate of smoking cessation among their patients. Suppose that 20 physician practices are randomly assigned to the group that receives the intervention and 20 practices are assigned to a control group. One year later, the investigators plan to review the charts of a random sample of 50 patients who had been smokers at baseline in each practice to determine how many have quit smoking. Does the sample size equal 40 (the number of practices) or 2,000 (the number of patients)? The answer, which lies somewhere in between those two extremes, depends upon how similar the patients within a practice are (in terms of their likelihood of smoking cessation) compared with the similarity among all the patients. Estimating this quantity often requires obtaining pilot data, unless another investigator has previously done a similar study. There are several techniques for estimating the required sample size for a study using clustered samples (4–7), but they are challenging and usually require the assistance of a statistician.

Matching

For a variety of reasons, an investigator may choose to use a matched design (Chapter 9). The techniques in this chapter, which ignore any matching, nevertheless provide reasonable estimates of the required sample size unless the exposure (in a matched case–control study) or outcome (in a matched cohort study) is strongly correlated with the matching variable. More precise estimates, which require the investigator to specify the correlation between exposures or outcomes in matched pairs, can be made using standard approaches (8), statistical software, or an interactive Web-based program.

Multivariate Adjustment and Other Special Statistical Analyses

When designing an observational study, an investigator may decide that one or more variables will confound the association between the predictor and outcome (Chapter 9), and plan to use

statistical techniques to adjust for these **confounders** when she analyzes her results. When this adjustment will be included in testing the primary hypothesis, the estimated sample size needs to take this into account.

Analytic approaches that adjust for confounding variables often increase the required sample size (9, 10). The magnitude of this increase depends on several factors, including the prevalence of the confounder, the strength of the association between the predictor and the confounder, and the strength of the association between the confounder and the outcome. These effects are complex and no general rule covers all situations.

Statisticians have developed multivariate methods such as linear regression and logistic regression that allow the investigator to adjust for confounding variables. One widely used statistical technique, **Cox proportional hazards** analysis, can adjust both for confounders and for differences in length of follow-up. If one of these techniques is going to be used to analyze the data, there are corresponding approaches for estimating the required sample size (3,11–14). Sample size techniques are also available for other designs, such as studies of potential genetic risk factors or candidate genes (15–17), economic studies (18–20), dose–response studies (21), or studies that involve more than two groups (22). Again, the Internet is a useful resource for these more sophisticated approaches (e.g., search for "sample size" and "logistic regression").

It is usually easier, at least for novice investigators, to estimate the sample size assuming a simpler method of analysis, such as the chi-squared test or the *t* test. It's also a good way to check the results obtained when using more sophisticated methods. Suppose, for example, an investigator is planning a case–control study of whether serum cholesterol level (a continuous variable) is associated with the occurrence of brain tumors (a dichotomous variable). Even if the eventual plan is to analyze the data with the logistic regression technique, a ballpark sample size can be estimated with the *t* test. It turns out that the simplified approaches usually produce sample size estimates that are similar to those generated by more sophisticated techniques. An experienced biostatistician should be consulted, however, if a grant proposal that involves substantial costs is being submitted for funding: Reviewers will expect you to use a sophisticated approach even if they realize that the sample size estimates are based on guesses about the risk of the outcome, the effect size, and so on. Having your sample size estimated by a statistician also conveys the message that you have access to the collaborators who will be needed to manage and analyze the study's data. Indeed, a biostatistician will contribute in many other ways to the design and execution of the study. But she will surely appreciate working with a clinical investigator who has thought about the issues and has made at least an initial attempt to estimate the sample size.

Equivalence and Non-Inferiority Trials

Sometimes the goal of a study is to *rule out* a substantial association between the predictor and outcome variables. An **equivalence trial** tests whether a new drug has pretty much the same efficacy as an established drug. This situation poses a challenge when planning sample size, because the desired effect size is zero or very small. A **non-inferiority trial** is a one-sided version of this design that examines whether the new drug is at least not substantially worse than the established drug (Chapter 11).

Sample size calculations for these designs are complex (23–26) and the advice of an experienced statistician will be helpful. One acceptable method is to design the study to have substantial power (say, 0.90 or 0.95) to reject the null hypothesis when the effect size is small enough that it would not be clinically important (e.g., a difference of 5 mg/dL in mean fasting glucose levels). If the results of such a well-powered study show "no effect" (i.e., the 95% confidence interval excludes the prespecified difference of 5 mg/dL), then the investigator can be reasonably sure that the two drugs have similar effects. One problem with equivalence and non-inferiority trials, however, is that the additional power and the small effect size often

require a very large sample size; of the two designs, non-inferiority trials have the advantage of being one-sided, permitting either a smaller sample size or a smaller alpha.

Another problem involves the loss of the usual safeguards that are inherent in the paradigm of the null hypothesis, which protects a conventional study that compares an active drug with a placebo, against type I errors (falsely rejecting the null hypothesis). The paradigm ensures that many problems in the design or execution of a study, such as using imprecise measurements or excessive loss to follow-up, make it harder to reject the null hypothesis. Investigators in a conventional study, who are trying to reject a null hypothesis, have a strong incentive to do the best possible study. For a non-inferiority study, however, in which the goal is to find no difference, those safeguards do not apply.

■ SAMPLE SIZE TECHNIQUES FOR DESCRIPTIVE STUDIES

Estimating the sample size for descriptive studies, including studies of diagnostic tests, is also based on somewhat different principles. Such studies do not have predictor and outcome variables, nor do they compare different groups statistically. Therefore, the concepts of power and the null and alternative hypotheses do not apply. Instead, the investigator calculates descriptive statistics, such as means and proportions. Often, however, descriptive studies (What is the prevalence of depression among elderly patients in a medical clinic?) are also used to ask analytic questions (What are the predictors of depression among these patients?). In this situation, sample size should be estimated for the analytic study as well, to avoid the common problem of having inadequate power for what turns out to be the question of greater interest.

Descriptive studies commonly report **confidence intervals**, a range of values about the sample mean or proportion. A confidence interval is a measure of the precision of a sample estimate. The investigator sets the confidence level, such as 95% or 99%. An interval with a greater confidence level (say 99%) is wider, and therefore more likely to include the true population value, than an interval with a lower confidence level (90%).

The width of a confidence interval depends on the sample size. For example, an investigator might wish to estimate the mean score on the U.S. Medical Licensing Examination in a group of medical students who were taught using an alternative Web-based curriculum. From a sample of 50 students, she might estimate that the mean score in the population of all students is 215, with a 95% confidence interval from 205 to 225. A smaller study, say with 20 students, might have about the same mean score but would almost certainly have a wider 95% confidence interval.

When estimating sample size for descriptive studies, the investigator specifies the desired level and width of the confidence interval. The sample size can then be determined from the tables or formulas in the appendix.

Continuous Variables

When the variable of interest is continuous, a confidence interval around the mean value of that variable is often reported. To estimate the sample size for that confidence interval (Example 6.4), the investigator must

1. Estimate the standard deviation of the variable of interest.
2. Specify the desired precision (total width) of the confidence interval.
3. Select the confidence level for the interval (e.g., 95%, 99%).

To use Appendix 6D, standardize the total width of the interval (divide it by the standard deviation of the variable), then look down the leftmost column of Table 6D for the expected standardized width. Next, read across the table to the chosen confidence level for the required sample size.

EXAMPLE 6.4 Calculating Sample Size for a Descriptive Study of a Continuous Variable

Problem: The investigator seeks to determine the mean hemoglobin level among third graders in an urban area with a 95% confidence interval of ± 0.3 g/dL. A previous study found that the standard deviation of hemoglobin in a similar city was 1 g/dL.

Solution: The ingredients for the sample size calculation are as follows:

1. Standard deviation of variable (SD) = 1 g/dL.
2. Total width of interval = 0.6 g/dL (0.3 g/dL above and 0.3 g/dL below). The standardized width of interval = total width ÷ SD = 0.6 ÷ 1 = 0.6.
3. Confidence level = 95%.

Reading across from a standardized width of 0.6 in the leftmost column of Table 6D and down from the 95% confidence level, the required sample size is 43 third graders.

Dichotomous Variables

In a descriptive study of a dichotomous variable, results can be expressed as a confidence interval around the estimated proportion of subjects with one of the values. This includes studies of the **sensitivity** or **specificity** of a diagnostic test, which appear at first glance to be continuous variables but are actually dichotomous—proportions expressed as percentages (Chapter 12). To estimate the sample size for that confidence interval, the investigator must

1. Estimate the expected proportion with the variable of interest in the population. (If more than half of the population is expected to have the characteristic, then plan the sample size based on the proportion expected not to have the characteristic.)
2. Specify the desired precision (total width) of the confidence interval.
3. Select the confidence level for the interval (e.g., 95%).

In Appendix 6E, look down the leftmost column of Table 6E for the expected proportion with the variable of interest. Next, read across the table to the chosen width and confidence level, yielding the required sample size.

Example 6.5 provides a sample size calculation for studying the sensitivity of a diagnostic test, which yields the required number of subjects with the disease. When studying the specificity of the test, the investigator must estimate the required number of subjects who do *not* have the disease. There are also techniques for estimating the sample size for studies of receiver operating characteristic (ROC) curves (27), likelihood ratios (28), and reliability (29) (Chapter 12).

EXAMPLE 6.5 Calculating Sample Size for a Descriptive Study of a Dichotomous Variable

Problem: The investigator wishes to determine the sensitivity of a new diagnostic test for pancreatic cancer. Based on a pilot study, she expects that 80% of patients with pancreatic cancer will have positive tests. How many such patients will be required to estimate a 95% confidence interval for the test's sensitivity of 0.80 ± 0.05?

Solution: The ingredients for the sample size calculation are as follows:

1. Expected proportion = 0.20. (Because 0.80 is more than half, sample size is estimated from the proportion expected to have a falsely negative result, that is, 0.20.)

2. Total width = 0.10 (0.05 below and 0.05 above).
3. Confidence level = 95%.

Reading across from 0.20 in the leftmost column of Table 6E and down from a total width of 0.10, the middle number (representing a 95% confidence level) yields the required sample size of 246 patients with pancreatic cancer.

WHAT TO DO WHEN SAMPLE SIZE IS FIXED

Especially when doing secondary data analysis, the sample size may have been determined before you design your study. Even when you are designing a study from scratch, it's common to find that the number of participants who are available or affordable for study is limited. Indeed, most investigators, if they are being honest, will acknowledge that they often "work backwards" from a fixed or realistic sample size to determine the effect size they'll have a reasonable power to detect. That's part of the reason why it's silly to treat a sample size estimate as if it was carved into stone.

When an investigator must work backward from the fixed sample size (Example 6.6), she estimates the effect size that can be detected at a given power (usually 80%). Less commonly, she estimates the power to detect a given effect. The investigator can use the sample size tables in the chapter appendixes, interpolating when necessary, or use the sample size formulas in the appendixes for estimating the effect size.

A general rule is that a study should have a power of 80% or greater to detect a reasonable effect size. There is nothing magical about 80%: Sometimes an investigator gets lucky and finds a statistically significant result even when she had limited power to do so (even a power as low as 50% provides a 50-50 chance of observing a statistically significant effect in the sample that is actually present in the population). Thus it may be worthwhile to pursue studies that have less than 80% power if the cost of doing so is small, such as when doing an analysis of data that have already been collected. And there are some studies—for example, one showing that a novel treatment reduces pulmonary arterial pressures by more than 50% in patients with longstanding refractory pulmonary hypertension—in which a sample size of two or three subjects would suffice to indicate that further study (on safety and effects on clinical outcomes) is warranted.

The investigator should keep in mind, however, that she might face the difficulty of interpreting (and publishing) a study that failed to find an association because of insufficient power; the broad confidence intervals will reveal the possibility of a substantial effect in the population from which the small study sample was drawn. It's also important to understand that an "under-powered" study that got "lucky" and had a statistically significant result may be criticized because reviewers are skeptical as to whether the investigator really intended to look for that particular association, or whether she tested scores of hypotheses and cherry-picked the one result that had a significant *P* value.

EXAMPLE 6.6 Calculating the Detectable Effect Size When Sample Size is Fixed

Problem: An investigator estimates that she will have access to 200 new mothers of twins during her fellowship. Based on a small pilot study, she estimates that about half of those women (i.e., 100) might be willing to participate in a study of whether a 6-week meditation program reduces stress, as compared with a control group that receives a pamphlet describing relaxation. If the standard deviation of the stress score is expected to be

(continued)

EXAMPLE 6.6 Calculating the Detectable Effect Size When Sample Size is Fixed *(continued)*

5 points in both the control and the treatment groups, what size difference will the investigator be able to detect between the two groups, at α (two-sided) = 0.05 and $\beta = 0.20$?

Solution: In Table 6A, reading down from α (two-sided) = 0.05 and $\beta = 0.20$ (the rightmost column in the middle triad of numbers), 45 patients per group are required to detect a standardized effect size of 0.6, which is equal to 3 points (0.6 × 5 points). The investigator (who will have about 50 patients per group) will be able to detect a difference of a little less than 3 points between the two groups.

STRATEGIES FOR MINIMIZING SAMPLE SIZE AND MAXIMIZING POWER

When the estimated sample size is greater than the number of subjects that can be studied realistically, the investigator should proceed through several steps. First, the calculations should be checked, as it is easy to make mistakes. Next, the "ingredients" should be reviewed. Is the effect size unreasonably small or the variability unreasonably large? Is α or β unreasonably small or is the confidence level too high or the interval unreasonably narrow?

These technical adjustments can be useful, but it is important to realize that statistical tests ultimately depend on the information contained in the data. Many changes in the ingredients, such as reducing power from 90% to 80%, do not improve the quantity or quality of the data that will be collected. There are, however, several strategies for reducing the required sample size or for increasing power for a given sample size that actually increase the information content of the collected data. Many of these strategies involve modifications of the research hypothesis; the investigator should carefully consider whether the new hypothesis still answers the research question that she wishes to study.

Use Continuous Variables

When continuous variables are an option, they usually permit smaller sample sizes than dichotomous variables. Blood pressure, for example, can be expressed either as millimeters of mercury (continuous) or as the presence or absence of hypertension (dichotomous). The former permits a smaller sample size for a given power or a greater power for a given sample size.

In Example 6.7, the continuous outcome addresses the effect of nutrition supplements on muscle strength among the elderly. The dichotomous outcome is concerned with its effects on the proportion of subjects who have at least a minimal amount of strength, which may be a more valid surrogate for potential fall-related morbidity.

EXAMPLE 6.7 Use of Continuous Versus Dichotomous Variables

Problem: Consider a placebo-controlled trial to determine the effect of nutrition supplements on strength in elderly nursing home residents. Previous studies have established that quadriceps strength (as peak torque in newton-meters) is approximately normally distributed, with a mean of 33 N·m and a standard deviation of 10 N·m, and that about 10% of the elderly have very weak muscles (strength <20 N·m). Nutrition supplements for 6 months are thought to be worthwhile if they can increase strength by 5 N·m as compared with the usual diet. This change in mean strength can be estimated, based on the distribution of quadriceps strength in the elderly, to correspond to a reduction in the proportion of the elderly who are very weak to about 5%.

One design might treat strength as a dichotomous variable: very weak versus not very weak. Another might use all the information contained in the measurement and treat

strength as a continuous variable. How many subjects would each design require at α (two-sided) = 0.05 and β = 0.20?

Solution: The ingredients for the sample size calculation using a **dichotomous outcome variable** (very weak versus not very weak) are as follows:

1. **Null Hypothesis:** The proportion of elderly nursing home residents who are very weak (peak quadriceps torque <20 N·m) after receiving 6 months of nutrition supplements is the same as the proportion who are very weak among those on a usual diet.
   Alternative Hypothesis: The proportion of elderly nursing home residents who are very weak (peak quadriceps torque <20 N·m) after receiving 6 months of nutrition supplements differs from the proportion among those on a usual diet.
2. P_1 (proportion very weak on usual diet) = 0.10; P_2 (in supplement group) = 0.05. The smaller of these values is 0.05, and the difference between them ($P_1 - P_2$) is 0.05.
3. α (two-sided) = 0.05; β = 0.20.

Using Table 6B.1, reading across from 0.05 in the leftmost column and down from an expected difference of 0.05, the middle number (for α [two-sided] = 0.05 and β = 0.20), this design would require 473 subjects per group.

The ingredients for the sample size calculation using a **continuous outcome variable** (quadriceps strength as peak torque) are as follows:

1. **Null Hypothesis:** Mean quadriceps strength (as peak torque in N·m) in elderly nursing home residents after receiving 6 months of nutrition supplements is the same as mean quadriceps strength in those on a usual diet.
   Alternative Hypothesis: Mean quadriceps strength (as peak torque in N·m) in elderly nursing home residents after receiving 6 months of nutrition supplements differs from mean quadriceps strength in those on a usual diet.
2. Effect size = 5 N·m.
3. Standard deviation of quadriceps strength = 10 N·m.
4. Standardized effect size = effect size ÷ standard deviation = 5 N·m ÷ 10 N·m = 0.5.
5. α (two-sided) 0.05; β = 0.20.

Using Table 6A, reading across from a standardized effect size of 0.50, with α (two-sided) = 0.05 and β = 0.20, this design would require about 64 subjects in each group. (In this example, the shortcut sample size estimate from page 57 of 16 ÷ (standardized effect size)^2, or 16 ÷ $(0.5)^2$, gives the same estimate of 64 subjects per group.) The bottom line is that the use of a continuous outcome variable resulted in a substantially smaller sample size.

Use Paired Measurements

In some experiments or cohort studies with continuous outcome variables, paired measurements—one at baseline, another at the conclusion of the study—can be made in each subject. The outcome variable is the change between these two measurements. In this situation, a *t* test on the paired measurements can be used to compare the mean value of this change in the two groups. This technique often permits a smaller sample size because, by comparing each subject with herself, it removes the baseline between-subject part of the variability of the outcome variable. For example, the change in weight on a diet has less variability than the final weight, because final weight is highly correlated with initial weight. Sample size for this type of *t* test is estimated in the usual way (Example 6.8), except that the standardized effect size (*E*/*S* in Table 6A) is the anticipated difference in the *change* in the variable divided by the standard deviation *of that change*.

EXAMPLE 6.8 Use of the *t* Test with Paired Measurements

Problem: Recall Example 6.1, in which the investigator studying the treatment of asthma is interested in determining whether albuterol can improve FEV_1 by 200 mL compared with ipratropium bromide. Sample size calculations indicated that 394 subjects per group are needed, more than are likely to be available. Fortunately, a colleague points out that asthmatic patients have great differences in their FEV_1 values before treatment. These between-subject differences account for much of the variability in FEV_1 after treatment, therefore obscuring the effect of treatment. She suggests using a (two-sample) paired *t* test to compare the *changes* in FEV_1 in the two groups. A pilot study finds that the standard deviation of the change in FEV_1 is only 250 mL. How many subjects would be required per group, at α (two-sided) = 0.05 and $\beta = 0.20$?

Solution: The ingredients for the sample size calculation are as follows:

1. **Null Hypothesis:** The change in mean FEV_1 after 2 weeks of treatment is the same in asthmatic patients treated with albuterol as in those treated with ipratropium bromide. **Alternative Hypothesis:** The change in mean FEV_1 after 2 weeks of treatment is different in asthmatic patients treated with albuterol from what it is in those treated with ipratropium bromide.
2. Effect size = 200 mL.
3. Standard deviation of the outcome variable = 250 mL.
4. Standardized effect size = effect size ÷ standard deviation = 200 mL ÷ 250 mL = 0.8.
5. α (two-sided) = 0.05; $\beta = 1 - 0.80 = 0.20$.

Using Table 6A, this design would require about 26 participants per group, a much more reasonable sample size than the 394 per group in Example 6.1. In this example, the shortcut sample size estimate of 16 ÷ (standardized effect size)^2, or 16 ÷ $(0.8)^2$, gives a similar estimate of 25 subjects per group.

A Brief Technical Note

This chapter always refers to **two-sample *t* tests**, which are used when comparing the mean values of continuous variables in two groups of subjects. A two-sample *t* test can be **unpaired**, if the variable itself is being compared between two groups (Example 6.1), or **paired** if the variable is the change in a pair of measurements, say before and after an intervention (e.g., Example 6.8).

A third type of *t* test, the ***one-sample* paired *t* test**, compares the mean change in a pair of measurements within a single group to zero change. This type of analysis is reasonably common in time series designs (Chapter 11), a before–after approach to examining treatments that are difficult to randomize (for example, the effect of elective hysterectomy, a decision few women are willing to leave to a coin toss, on quality of life). However, it is a weaker design because the absence of a comparison group makes it difficult to know what would have happened had the subjects been left untreated. When planning a study that will be analyzed with a one-sample paired *t* test, the total sample size is just half of the sample size per group listed in Appendix 6A. For example, for $\alpha = 0.05$ (two-sided) and $\beta = 0.2$, to detect a 0.5 standard deviation difference ($E/S = 0.5$) would require 64/2 = 32 subjects. Appendix 6F presents additional information on the use and misuse of one- and two-sample *t* tests.

Use More Precise Variables

Because they reduce variability, more precise variables permit a smaller sample size in both analytic and descriptive studies. Even a modest change in precision can have a substantial effect on sample size. For example, when using the *t* test to estimate sample size, a 20% decrease

in the standard deviation of the outcome variable results in a 36% decrease in the sample size. Techniques for increasing the precision of a variable, such as making measurements in duplicate, are presented in Chapter 4.

Use Unequal Group Sizes

Because an equal number of subjects in each of two groups usually gives the greatest power for a given total number of subjects, Tables 6A, 6B.1, and 6B.2 in the appendixes assume equal sample sizes in the two groups. Sometimes, however, the distribution of subjects is not equal in the two groups, or it is easier or less expensive to recruit study subjects for one group than the other. It may turn out, for example, that an investigator wants to estimate sample size for a study comparing the 30% of the subjects in a cohort who smoke cigarettes with the 70% who do not. Or, in a case–control study, the number of persons with the disease may be small, but it may be possible to sample a much larger number of controls. In general, the gain in power when the size of one group is increased to twice the size of the other is considerable; tripling and quadrupling one of the groups provide progressively smaller gains. Sample sizes for unequal groups can be computed from the formulas found in the text to Appendixes 6A and 6B or from the sample size calculators in statistical software or on the Web.

Here is a useful approximation (30) for estimating sample size for case–control studies of dichotomous risk factors and outcomes using c controls per case (Example 6.9). If n represents the number of cases that would have been required for one control per case (at a given α, β, and effect size), then the approximate number of cases (n') with cn' controls that will be required is

$$n' = [(c + 1) \div 2c] \times n$$

For example, with $c = 2$ controls per case, then $[(2 + 1) \div (2 \times 2)] \times n = 3/4 \times n$, and only 75% as many cases are needed. As c gets larger, n' approaches 50% of n (when $c = 10$, for example, $n' = 11/20 \times n$).

Use a More Common Outcome

When planning a study of a dichotomous outcome, the more frequently that outcome occurs, up to a frequency of about 0.5, the greater the power. So changing the definition of an outcome is one of the best ways to increase power: If an outcome occurs more often, there is more of a chance to detect its predictors. Indeed, power depends more on the number of subjects with a specified outcome than it does on the total number of subjects in the study. Studies with rare outcomes, like the occurrence of breast cancer in healthy women, require very large sample sizes to have adequate power.

One of the best ways to have an outcome occur more frequently is to enroll subjects at greater risk of developing that outcome (such as women with a family history of breast cancer).

EXAMPLE 6.9 Use of Multiple Controls per Case in a Case–Control Study

Problem: An investigator is studying whether exposure to household insecticide is a risk factor for aplastic anemia. The original sample size calculation indicated that 25 cases would be required, using one control per case. Suppose that the investigator has access to only 18 cases. How should the investigator proceed?

Solution: The investigator should consider using multiple controls per case (after all, she can find many patients who do not have aplastic anemia). By using three controls per case, for example, the approximate number of cases that will be required is $[(3 + 1) \div (2 \times 3)] \times 25 = 17$.

EXAMPLE 6.10 Use of a More Common Outcome

Problem: Suppose an investigator is comparing the efficacy of an antiseptic gargle versus a placebo gargle in preventing upper respiratory infections. Her initial calculations indicated that her anticipated sample of 200 volunteer college students was inadequate, in part because she expected that only about 20% of her subjects would have an upper respiratory infection during the 3-month follow-up period. Suggest a few changes in the study plan.

Solution: Here are three possible solutions: (1) study a sample of pediatric interns and residents, who are likely to experience a much greater incidence of upper respiratory infections than college students; or (2) do the study in the winter, when these infections are more common; or (3) follow the sample for a longer period of time, say 6 or 12 months. All of these solutions involve modification of the research hypothesis, but none of them seem sufficiently large to affect the overall research question about the efficacy of antiseptic gargle.

Others are to extend the follow-up period, so that there is more time to accumulate outcomes, or to loosen the definition of what constitutes an outcome (e.g., by including ductal carcinoma in situ). All these techniques (Example 6.10), however, may change the research question, so they should be used with caution.

■ HOW TO ESTIMATE SAMPLE SIZE WHEN THERE IS INSUFFICIENT INFORMATION

Often the investigator finds that she is missing one or more of the ingredients for the sample size calculation and becomes frustrated in her attempts to plan the study. This is an especially frequent problem when the investigator is using an instrument of her design (such as a new questionnaire comparing quality of life in women with stress versus urge incontinence). How should she go about deciding what fraction of a standard deviation of the scores on her instrument would be clinically significant?

The first strategy is an **extensive search** for previous and related findings on the topic and on similar research questions. Roughly comparable situations and mediocre or dated findings may be good enough. For example, are there data on quality of life among patients with other urologic problems, or with related conditions like having a colostomy? If the literature review is unproductive, she should contact other investigators about their judgment on what to expect, and whether they are aware of any unpublished results that may be relevant.

If there is still no information available, she may consider doing a small **pilot study** or obtaining a data set for a secondary analysis to obtain the missing ingredients before embarking on the main study. Indeed, a pilot study is highly recommended for almost all studies that involve new instruments, measurement methods, or recruitment strategies. It saves time in the end by enabling investigators to do a much better job planning the main study. Pilot studies are useful for estimating the standard deviation of a measurement, or the proportion of subjects with a particular characteristic. However, an alternative is to recognize that for continuous variables that have a roughly bell-shaped distribution, the **standard deviation** can be estimated as one-quarter of the difference between the high and low ends of the range of values that occur commonly, ignoring extreme values. For example, if most subjects are likely to have a serum sodium level between 135 and 143 mEq/L, the standard deviation of serum sodium is about 2 mEq/L (1/4 × 8 mEq/L).

Another strategy, when the mean and standard deviation of a continuous or categorical variable are in doubt, is to **dichotomize** that variable. Categories can be lumped into two groups, and continuous variables can be split at their mean or median. For example, dividing quality of

life into "better than the median" or "the median or less" avoids having to estimate its standard deviation in the sample, although one still has to estimate what proportions of subjects would be above the overall median in each of the two groups being studied. The chi-squared test can then be used to make a reasonable, albeit somewhat high, estimate of the sample size.

Often, however, the investigator must choose the detectable effect size based on a value that she considers to be **clinically meaningful**. In that situation, the investigator should vet her choice with colleagues in the field. For example, suppose that an investigator is studying a new invasive treatment for severe refractory gastroparesis, a condition in which at most 5% of patients improve spontaneously. If the treatment is shown to be effective, her gastroenterologist colleagues indicate that they would be willing to treat up to five patients to produce a sustained benefit in just one of them (because the treatment has substantial side effects and is expensive, they don't think the number would be more than five). A number needed to treat (NNT) of 5 corresponds to a risk difference of 20% (NNT = 1/risk difference), so the investigator should estimate the sample size based on a comparison of $P_1 = 5\%$ versus $P_2 = 25\%$ (i.e., 59 subjects per group at a power of 0.80 and a two-sided α of 0.05).

If all this fails, the investigator should just make an **educated guess** about the likely values of the missing ingredients. The process of thinking through the problem and imagining the findings will often result in a reasonable estimate, and that is what sample size planning is about. This is usually a better option than just deciding, in the absence of any rationale, to design the study to have 80% power at a two-sided α of 0.05 to detect a standardized effect size of, say, 0.5 between the two groups ($n = 64$, per group, by the way). Very few grant reviewers will accept that sort of entirely arbitrary decision.

COMMON ERRORS TO AVOID

Many inexperienced investigators (and some experienced ones!) make mistakes when planning sample size. A few of the more common ones follow:

1. A common error is estimating the sample size late during the design of the study. Do it early in the process, when fundamental changes can still be made.
2. Dichotomous variables can appear to be continuous when they are expressed as a percentage or rate. For example, vital status (alive or dead) might be misinterpreted as continuous when expressed as percent alive. Similarly, in a survival analysis in which not all subjects die, a dichotomous outcome can appear to be continuous (e.g., median survival in months). For all of these, the outcome itself is actually dichotomous (a proportion) and the appropriate simple approach in planning sample size would be the chi-squared test.
3. The sample size estimates the number of subjects with outcome data, not the number who need to be enrolled. The investigator should always plan for dropouts and subjects with missing data.
4. The tables at the end of the chapter assume that the two groups being studied have equal sample sizes. Often that is not the case; for example, a cohort study of whether use of vitamin supplements reduces the risk of sunburn would probably not enroll equal numbers of subjects who used, or did not use, vitamins. If the sample sizes are not equal, then the formulas that follow the tables or calculators on the Web or in statistical software should be used.
5. When using the *t* test to estimate the sample size, the standard deviation of the outcome variable is a key factor. Therefore, if the outcome is change in a continuous variable, the investigator should use the standard deviation of that change rather than the standard deviation of the variable itself.
6. Be aware of clustered data. If there appear to be two "levels" of sample size (e.g., one for physicians and another for patients), clustering is a likely problem and the tables in the appendices do not apply.

7. If you find yourself having difficulty estimating a sample size for your study, be sure that your research hypothesis meets the criteria discussed earlier in this chapter (simple, specific, and stated in advance).

SUMMARY

1. When estimating sample size for an **analytic study**, the following steps need to be taken:
   (a) state the **null and alternative hypotheses**, specifying the **number of sides**;
   (b) select a **statistical test** that could be used to analyze the data, based on the types of predictor and outcome variables (**chi-squared test** if both are dichotomous, ***t* test** if one is dichotomous and one continuous, and **correlation coefficient** if both are continuous);
   (c) estimate the **effect size** (and its **variability**, if necessary); and
   (d) specify appropriate values for **α and β**, based on the importance of avoiding **type I** and **type II errors**.
2. Other considerations in calculating sample size for analytic studies include adjusting for potential **dropouts**; strategies for dealing with **categorical variables**, **survival analysis**, **clustered samples**, **multivariate adjustment**; and special statistical approaches to equivalence and **non-inferiority trials**.
3. The steps for estimating sample size for **descriptive studies**, which do not have hypotheses, are to (a) estimate the **proportion** of subjects with a dichotomous outcome or the **standard deviation** of a continuous outcome; (b) specify the desired precision (width of the **confidence interval**); and (c) specify the **confidence level** (e.g., 95%).
4. When sample size is **predetermined**, the investigator can work backward to estimate the detectable **effect size** or, less commonly, the study's **power**.
5. Strategies to **minimize sample size** include using **continuous variables**, more **precise measurements**, **paired measurements**, and more **common outcomes**, as well as increasing the number of controls per case in case–control studies.
6. When there seems to be not enough information to estimate the sample size, the investigator should review the **literature** in related areas and consult with **colleagues** to help choose an effect size that is clinically meaningful.
7. Errors to avoid include estimating sample size **too late**, misinterpreting **proportions expressed as percentages**, not taking **missing subjects and data** into account, and not addressing **clustered and paired data** appropriately.

APPENDIX 6A
Sample Size Required per Group When Using the *t* Test to Compare Means of Continuous Variables

TABLE 6A SAMPLE SIZE *PER GROUP* FOR COMPARING TWO MEANS

ONE-SIDED α =	0.005			0.025			0.05		
TWO-SIDED α =	0.01			0.05			0.10		
*E/S** β =	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20
0.10	3,565	2,978	2,338	2,600	2,103	1,571	2,166	1,714	1,238
0.15	1,586	1,325	1,040	1,157	935	699	963	762	551
0.20	893	746	586	651	527	394	542	429	310
0.25	572	478	376	417	338	253	347	275	199
0.30	398	333	262	290	235	176	242	191	139
0.40	225	188	148	164	133	100	136	108	78
0.50	145	121	96	105	86	64	88	70	51
0.60	101	85	67	74	60	45	61	49	36
0.70	75	63	50	55	44	34	45	36	26
0.80	58	49	39	42	34	26	35	28	21
0.90	46	39	32	34	27	21	28	22	16
1.00	38	32	26	27	23	17	23	18	14

**E/S* is the standardized effect size, computed as *E* (expected effect size) divided by *S* (SD of the outcome variable). To estimate the sample size, read across from the *standardized effect size*, and down from the specified values of α and β for the required sample size in each group. For a one-sample *t* test, the total sample size is one-half of the number listed.

CALCULATING VARIABILITY

Variability is usually reported as either the standard deviation or the standard error of the mean (SEM). For the purposes of sample size calculation, the standard deviation of the variable is most useful. Fortunately, it is easy to convert from one measure to another: The standard deviation is simply the standard error times the square root of N, where N is the number of subjects that makes up the mean. Suppose a study reported that the weight loss in 25 persons on a low-fiber diet was 10 ± 2 kg (mean ± SEM). The standard deviation would be $2 \times \sqrt{25} = 10$ kg.

GENERAL FORMULA FOR OTHER VALUES

The general formula for other values of E, S, α, and β, or for unequal group sizes, is as follows. Let:

Z_α = the standard normal deviate for α (If the alternative hypothesis is two-sided, $Z_\alpha = 2.58$ when $\alpha = 0.01$, $Z_\alpha = 1.96$ when $\alpha = 0.05$, and $Z_\alpha = 1.645$ when $\alpha = 0.10$. If the alternative hypothesis is one-sided, $Z_\alpha = 1.645$ when $\alpha = 0.05$.)

Z_β = the standard normal deviate for β ($Z_\beta = 0.84$ when $\beta = 0.20$, and $Z_\beta = 1.282$ when $\beta = 0.10$)
q_1 = proportion of subjects in group 1
q_2 = proportion of subjects in group 2
N = **total** number of subjects required

Then:

$$N = [(1/q_1 + 1/q_2)\ S^2\ (Z_\alpha + Z_\beta)^2] \div E^2.$$

Readers who would like to skip the work involved in hand calculations with this formula can get an instant answer from a calculator on our website (www.epibiostat.ucsf.edu/dcr/). (Because this formula is based on approximating the t statistic with a Z statistic, it will slightly underestimate the sample size when N is less than about 30. Table 6A uses the t statistic to estimate sample size.)

APPENDIX 6B
Sample Size Required per Group When Using the Chi-Squared Statistic or *Z* Test to Compare Proportions of Dichotomous Variables

TABLE 6B.1 SAMPLE SIZE *PER GROUP* FOR COMPARING TWO PROPORTIONS

UPPER NUMBER: $\alpha = 0.05$ (ONE-SIDED) OR $\alpha = 0.10$ (TWO-SIDED); $\beta = 0.20$
MIDDLE NUMBER: $\alpha = 0.025$ (ONE-SIDED) OR $\alpha = 0.05$ (TWO-SIDED); $\beta = 0.20$
LOWER NUMBER: $\alpha = 0.025$ (ONE-SIDED) OR $\alpha = 0.05$ (TWO-SIDED); $\beta = 0.10$

SMALLER OF P_1 AND P_2*	DIFFERENCE BETWEEN P_1 AND P_2									
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.05	381	129	72	47	35	27	22	18	15	13
	473	159	88	59	43	33	26	22	18	16
	620	207	113	75	54	41	33	27	23	19
0.10	578	175	91	58	41	31	24	20	16	14
	724	219	112	72	51	37	29	24	20	17
	958	286	146	92	65	48	37	30	25	21
0.15	751	217	108	67	46	34	26	21	17	15
	944	270	133	82	57	41	32	26	21	18
	1,252	354	174	106	73	53	42	33	26	22
0.20	900	251	121	74	50	36	28	22	18	15
	1,133	313	151	91	62	44	34	27	22	18
	1,504	412	197	118	80	57	44	34	27	23
0.25	1,024	278	132	79	53	38	29	23	18	15
	1,289	348	165	98	66	47	35	28	22	18
	1,714	459	216	127	85	60	46	35	28	23
0.30	1,123	300	141	83	55	39	29	23	18	15
	1,415	376	175	103	68	48	36	28	22	18
	1,883	496	230	134	88	62	47	36	28	23
0.35	1,197	315	146	85	56	39	29	23	18	15
	1,509	395	182	106	69	48	36	28	22	18
	2,009	522	239	138	90	62	47	35	27	22
0.40	1,246	325	149	86	56	39	29	22	17	14
	1,572	407	186	107	69	48	35	27	21	17
	2,093	538	244	139	90	62	46	34	26	21

(continued)

TABLE 6B.1 SAMPLE SIZE *PER GROUP* FOR COMPARING TWO PROPORTIONS *(CONTINUED)*

UPPER NUMBER: $\alpha = 0.05$ (ONE-SIDED) OR $\alpha = 0.10$ (TWO-SIDED); $\beta = 0.20$
MIDDLE NUMBER: $\alpha = 0.025$ (ONE-SIDED) OR $\alpha = 0.05$ (TWO-SIDED); $\beta = 0.20$
LOWER NUMBER: $\alpha = 0.025$ (ONE-SIDED) OR $\alpha = 0.05$ (TWO-SIDED); $\beta = 0.10$

SMALLER OF P_1 AND P_2*	DIFFERENCE BETWEEN P_1 AND P_2									
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.45	1,271	328	149	85	55	38	28	21	16	13
	1,603	411	186	106	68	47	34	26	20	16
	2,135	543	244	138	88	60	44	33	25	19
0.50	1,271	325	146	83	53	36	26	20	15	—
	1,603	407	182	103	66	44	32	24	18	—
	2,135	538	239	134	85	57	42	30	23	—
0.55	1,246	315	141	79	50	34	24	18	—	—
	1,572	395	175	98	62	41	29	22	—	—
	2,093	522	230	127	80	53	37	27	—	—
0.60	1,197	300	132	74	46	31	22	—	—	—
	1,509	376	165	91	57	37	26	—	—	—
	2,009	496	216	118	73	48	33	—	—	—
0.65	1,123	278	121	67	41	27	—	—	—	—
	1,415	348	151	82	51	33	—	—	—	—
	1,883	459	197	106	65	41	—	—	—	—
0.70	1,024	251	108	58	35	—	—	—	—	—
	1,289	313	133	72	43	—	—	—	—	—
	1,714	412	174	92	54	—	—	—	—	—
0.75	900	217	91	47	—	—	—	—	—	—
	1,133	270	112	59	—	—	—	—	—	—
	1,504	354	146	75	—	—	—	—	—	—
0.80	751	175	72	—	—	—	—	—	—	—
	944	219	88	—	—	—	—	—	—	—
	1,252	286	113	—	—	—	—	—	—	—
0.85	578	129	—	—	—	—	—	—	—	—
	724	159	—	—	—	—	—	—	—	—
	958	207	—	—	—	—	—	—	—	—
0.90	381	—	—	—	—	—	—	—	—	—
	473	—	—	—	—	—	—	—	—	—
	620	—	—	—	—	—	—	—	—	—

The one-sided estimates use the *Z* statistic.

*P_1 represents the proportion of subjects expected to have the outcome in one group; P_2 in the other group. (In a case–control study, P_1 represents the proportion of cases with the predictor variable; P_2 the proportion of controls with the predictor variable.) To estimate the sample size, read across from the smaller of P_1 and P_2, and down the expected *difference* between P_1 and P_2. The three numbers represent the sample size required in each group for the specified values of α and β.

Additional detail for P_1 and P_2 between 0.01 and 0.10 is given in Table 6B.2.

TABLE 6B.2 SAMPLE SIZE *PER GROUP* FOR COMPARING TWO PROPORTIONS, THE SMALLER OF WHICH IS BETWEEN 0.01 AND 0.10

UPPER NUMBER: $\alpha = 0.05$ (ONE-SIDED) OR $\alpha = 0.10$ (TWO-SIDED); $\beta = 0.20$
MIDDLE NUMBER: $\alpha = 0.025$ (ONE-SIDED) OR $\alpha = 0.05$ (TWO-SIDED); $\beta = 0.20$
LOWER NUMBER: $\alpha = 0.025$ (ONE-SIDED) OR $\alpha = 0.05$ (TWO-SIDED); $\beta = 0.10$

SMALLER OF P_1 AND P_2	EXPECTED DIFFERENCE BETWEEN P_1 AND P_2									
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
0.01	2,019	700	396	271	204	162	134	114	98	87
	2,512	864	487	332	249	197	163	138	120	106
	3,300	1,125	631	428	320	254	209	178	154	135
0.02	3,205	994	526	343	249	193	157	131	113	97
	4,018	1,237	651	423	306	238	192	161	137	120
	5,320	1,625	852	550	397	307	248	207	177	154
0.03	4,367	1,283	653	414	294	224	179	148	126	109
	5,493	1,602	813	512	363	276	220	182	154	133
	7,296	2,114	1,067	671	474	359	286	236	199	172
0.04	5,505	1,564	777	482	337	254	201	165	139	119
	6,935	1,959	969	600	419	314	248	203	170	146
	9,230	2,593	1,277	788	548	410	323	264	221	189
0.05	6,616	1,838	898	549	380	283	222	181	151	129
	8,347	2,308	1,123	686	473	351	275	223	186	159
	11,123	3,061	1,482	902	620	460	360	291	242	206
0.06	7,703	2,107	1,016	615	422	312	243	197	163	139
	9,726	2,650	1,272	769	526	388	301	243	202	171
	12,973	3,518	1,684	1,014	691	508	395	318	263	223
0.07	8,765	2,369	1,131	680	463	340	263	212	175	148
	11,076	2,983	1,419	850	577	423	327	263	217	183
	14,780	3,965	1,880	1,123	760	555	429	343	283	239
0.08	9,803	2,627	1,244	743	502	367	282	227	187	158
	12,393	3,308	1,562	930	627	457	352	282	232	195
	16,546	4,401	2,072	1,229	827	602	463	369	303	255
0.09	10,816	2,877	1,354	804	541	393	302	241	198	167
	13,679	3,626	1,702	1,007	676	491	377	300	246	207
	18,270	4,827	2,259	1,333	893	647	495	393	322	270
0.10	11,804	3,121	1,461	863	578	419	320	255	209	175
	14,933	3,936	1,838	1,083	724	523	401	318	260	218
	19,952	5,242	2,441	1,434	957	690	527	417	341	285

The one-sided estimates use the *Z* statistic.

GENERAL FORMULA FOR OTHER VALUES

The general formula for calculating the ***total*** sample size (N) required for a study using the Z statistic, where P_1 and P_2 are defined above, is as follows (see Appendix 6A for definitions of Z_α and Z_β). Let

q_1 = proportion of subjects in group 1
q_2 = proportion of subjects in group 2
N = total number of subjects
$P = q_1 P_1 + q_2 P_2$

Then

$$N = \frac{[Z_\alpha \sqrt{P(1-P)(1/q_1 + 1/q_2)} + Z_\beta \sqrt{P_1(1-P_1)(1/q_1) + P_2(1-P_2)(1/q_2)}]^2}{(P_1 - P_2)^2}$$

Readers who would like to skip the work involved in hand calculations with this formula can get an instant answer from a calculator on our website (www.epibiostat.ucsf.edu/dcr/). (This formula does not include the Fleiss-Tytun-Ury continuity correction and therefore underestimates the required sample size by up to about 10%. Tables 6B.1 and 6B.2 do include this continuity correction.)

APPENDIX 6C
Total Sample Size Required When Using the Correlation Coefficient (*r*)

TABLE 6C SAMPLE SIZE FOR DETERMINING WHETHER A CORRELATION COEFFICIENT DIFFERS FROM ZERO

ONE-SIDED α =	0.005			0.025			0.05		
TWO-SIDED α =	0.01			0.05			0.1		
β =	0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20
r*									
0.05	7,118	5,947	4,663	5,193	4,200	3,134	4,325	3,424	2,469
0.10	1,773	1,481	1,162	1,294	1,047	782	1,078	854	616
0.15	783	655	514	572	463	346	477	378	273
0.20	436	365	287	319	259	194	266	211	153
0.25	276	231	182	202	164	123	169	134	98
0.30	189	158	125	139	113	85	116	92	67
0.35	136	114	90	100	82	62	84	67	49
0.40	102	86	68	75	62	47	63	51	37
0.45	79	66	53	58	48	36	49	39	29
0.50	62	52	42	46	38	29	39	31	23
0.60	40	34	27	30	25	19	26	21	16
0.70	27	23	19	20	17	13	17	14	11
0.80	18	15	13	14	12	9	12	10	8

*To estimate the total sample size, read across from *r* (the expected correlation coefficient) and down from the specified values of α and β.

GENERAL FORMULA FOR OTHER VALUES

The general formula for other values of r, α, and β is as follows (see Appendix 6A for definitions of Z_α and Z_β). Let

r = expected correlation coefficient
$C = 0.5 \times \ln [(1 + r)/(1 - r)]$
N = Total number of subjects required

Then

$$N = [(Z_\alpha + Z_\beta) \div C]^2 + 3.$$

ESTIMATING SAMPLE SIZE FOR DIFFERENCE BETWEEN TWO CORRELATIONS

If testing whether a correlation, r_1, is different from r_2 (i.e., the null hypothesis is that $r_1 = r_2$; the alternative hypothesis is that $r_1 \neq r_2$), let

$C_1 = 0.5 \times \ln [(1 + r_1)/(1 - r_1)]$
$C_2 = 0.5 \times \ln [(1 + r_2)/(1 - r_2)]$

Then

$$N = [(Z_\alpha + Z_\beta) \div (C_1 - C_2)]^2 + 3.$$

APPENDIX 6D
Sample Size for a Descriptive Study of a Continuous Variable

TABLE 6D SAMPLE SIZE FOR COMMON VALUES OF *W/S

	CONFIDENCE LEVEL		
W/S	90%	95%	99%
0.10	1,083	1,537	2,665
0.15	482	683	1,180
0.20	271	385	664
0.25	174	246	425
0.30	121	171	295
0.35	89	126	217
0.40	68	97	166
0.50	44	62	107
0.60	31	43	74
0.70	23	32	55
0.80	17	25	42
0.90	14	19	33
1.00	11	16	27

**W/S* is the standardized width of the confidence interval, computed as *W* (desired total width) divided by *S* (standard deviation of the variable). To estimate the total sample size, read across from the *standardized width* and down from the specified confidence level.

GENERAL FORMULA FOR OTHER VALUES

For other values of W, S, and a confidence level of $(1 - \alpha)$, the total number of subjects required (N) is

$$N = 4Z_{\alpha} S^2 \div W^2$$

(see Appendix 6A for the definition of Z_{α}).

APPENDIX 6E
Sample Size for a Descriptive Study of a Dichotomous Variable

TABLE 6E SAMPLE SIZE FOR PROPORTIONS

UPPER NUMBER: 90% CONFIDENCE LEVEL
MIDDLE NUMBER: 95% CONFIDENCE LEVEL
LOWER NUMBER: 99% CONFIDENCE LEVEL

EXPECTED PROPORTION (*P*)*	TOTAL WIDTH OF CONFIDENCE INTERVAL (*W*)						
	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0.10	98	44	—	—	—	—	—
	138	61	—	—	—	—	—
	239	106	—	—	—	—	—
0.15	139	62	35	22	—	—	—
	196	87	49	31	—	—	—
	339	151	85	54	—	—	—
0.20	174	77	44	28	19	14	—
	246	109	61	39	27	20	—
	426	189	107	68	47	35	—
0.25	204	91	51	33	23	17	13
	288	128	72	46	32	24	18
	499	222	125	80	55	41	31
0.30	229	102	57	37	25	19	14
	323	143	81	52	36	26	20
	559	249	140	89	62	46	35
0.40	261	116	65	42	29	21	16
	369	164	92	59	41	30	23
	639	284	160	102	71	52	40
0.50	272	121	68	44	30	22	17
	384	171	96	61	43	31	24
	666	296	166	107	74	54	42

*To estimate the sample size, read across the *expected proportion* (*P*) who have the variable of interest and down from the desired *total width* (*W*) of the confidence interval. The three numbers represent the sample size required for 90%, 95%, and 99% confidence levels.

GENERAL FORMULA FOR OTHER VALUES

The general formula for other values of P, W, and a confidence level of $(1 - \alpha)$, where P and W are defined above, is as follows. Let

Z_α = the standard normal deviate for a two-sided α, where $(1 - \alpha)$ is the confidence level (e.g., since $\alpha = 0.05$ for a 95% confidence level, $Z_\alpha = 1.96$; for a 90% confidence level $Z_\alpha = 1.65$, and for a 99% confidence level $Z_\alpha = 2.58$).

Then the total number of subjects required is:

$$N = 4Z_\alpha^2 \, P \, (1 - P) \div W^2$$

APPENDIX 6F
Use and Misuse of *t* Tests

Two-sample *t* tests, the primary focus of this chapter, are used when comparing the mean values of a variable in two groups of subjects. The two groups can be defined by a predictor variable—active drug versus placebo in a randomized trial, or presence versus absence of a risk factor in a cohort study—or they can be defined by an outcome variable, as in a case–control study. A two-sample *t* test can be **unpaired**, if measurements obtained on a single occasion are being compared between two groups, or **paired** if the change in measurements made at two points in time, say before and after an intervention, are being compared between the groups. A third type of *t* test, the **one-sample paired *t* test**, compares the mean change in measurements at two points in time within a single group to zero or some other specified change.

Table 6F illustrates the misuse of one-sample paired *t* tests in a study designed for between-group comparisons—a randomized blinded trial of the effect of a new sleeping pill on quality of life. In situations like this, some investigators have performed (and published!) findings with two separate one-sample *t* tests—one each in the treatment and placebo groups.

In the table, the *P* values designated with a dagger (†) are from one-sample paired *t* tests. The first *P* (0.05) shows a significant change in quality of life in the treatment group during the study; the second *P* value (0.16) shows no significant change in the control group. However, this analysis does not permit inferences about differences between the groups, and it would be wrong to conclude that there was a significant effect of the treatment.

The *P* values designated with a (*) represent the appropriate two-sample *t* test results. The first two *P* values (0.87 and 0.64) are two-sample unpaired *t* tests that show no statistically significant between-group differences in the initial or final measurements for quality of life. The last *P* value (0.17) is a two-sample paired *t* test; it is closer to 0.05 than the *P* value for the end of study values (0.64) because the paired mean differences have smaller standard deviations. However, the improved quality of life in the treatment group (1.3) was not significantly different from that in the placebo group (0.9), and the correct conclusion is that the study did not find the treatment to be effective.

TABLE 6F CORRECT (AND INCORRECT) WAYS TO ANALYZE PAIRED DATA

	QUALITY OF LIFE, AS MEAN ± SD		
TIME OF MEASUREMENT	TREATMENT (*N* = 100)	CONTROL (*N* = 100)	*P* VALUE
Baseline	7.0 ± 4.5	7.1 ± 4.4	0.87*
End of study	8.3 ± 4.7	8.0 ± 4.6	0.64*
P value	0.05†	0.16†	
Difference	1.3 ± 2.1	0.9 ± 2.0	0.17*

*Comparing treatment with control.

†Comparing baseline with end of study.

REFERENCES

1. Lehr R. Sixteen S-squared over D-squared: a relation for crude sample size estimates. *Stat Med* 1992;11:1099–1102.
2. Barthel FM, Babiker A, Royston P, Parmar MK. Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. *Stat Med* 2006;25(15):2521–2542.
3. Ahnn S, Anderson SJ. Sample size determination in complex clinical trials comparing more than two groups for survival endpoints. *Stat Med* 1998;17(21):2525–2534.
4. Donner A. Sample size requirements for stratified cluster randomization designs [published erratum appears in *Stat Med* 1997;30(16):2927]. *Stat Med* 1992;11:743–750.
5. Kerry SM, Bland JM. Trials which randomize practices II: sample size. *Fam Pract* 1998;15:84–87.
6. Hemming K, Girling AJ, Sitch AJ, et al. Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Med Res Methodol* 2011;11:102.
7. Jahn-Eimermacher A, Ingel K, Schneider A. Sample size in cluster-randomized trials with time to event as the primary endpoint. *Stat Med* 2013;32(5):739–751.
8. Edwardes MD. Sample size requirements for case–control study designs. *BMC Med Res Methodol* 2001;1:11.
9. Drescher K, Timm J, Jöckel KH. The design of case–control studies: the effect of confounding on sample size requirements. *Stat Med* 1990;9:765–776.
10. Lui KJ. Sample size determination for case–control studies: the influence of the joint distribution of exposure and confounder. *Stat Med* 1990;9:1485–1493.
11. Latouche A, Porcher R, Chevret S. Sample size formula for proportional hazards modelling of competing risks. *Stat Med* 2004;23(21):3263–3274.
12. Novikov I, Fund N, Freedman LS. A modified approach to estimating sample size for simple logistic regression with one continuous covariate. *Stat Med* 2010;29(1):97–107.
13. Vaeth M, Skovlund E. A simple approach to power and sample size calculations in logistic regression and Cox regression models. *Stat Med* 2004;23(11):1781–1792.
14. Dupont WD, Plummer WD Jr. Power and sample size calculations for studies involving linear regression. *Control Clin Trials* 1998;19:589–601.
15. Murcray CE, Lewinger JP, Conti DV, et al. Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genet Epidemiol* 2011;35(3):201–210.
16. Wang S, Zhao H. Sample size needed to detect gene-gene interactions using linkage analysis. *Ann Hum Genet* 2007;71(Pt 6):828–842.
17. Witte JS. Rare genetic variants and treatment response: sample size and analysis issues. *Stat Med* 2012; 31(25):3041–3050.
18. Willan AR. Sample size determination for cost-effectiveness trials. *Pharmacoeconomics* 2011;29(11):933–949.
19. Glick HA. Sample size and power for cost-effectiveness analysis (Part 2): the effect of maximum willingness to pay. *Pharmacoeconomics* 2011;29(4):287–296.
20. Glick HA. Sample size and power for cost-effectiveness analysis (Part 1). *Pharmacoeconomics* 2011;29(3):189–198.
21. Patel HI. Sample size for a dose-response study [published erratum appears in *J Biopharm Stat* 1994;4:127]. *J Biopharm Stat* 1992;2:l–8.
22. Day SJ, Graham DF. Sample size estimation for comparing two or more treatment groups in clinical trials. *Stat Med* 1991;10:33–43.
23. Guo JH, Chen HJ, Luh WM. Sample size planning with the cost constraint for testing superiority and equivalence of two independent groups. *Br J Math Stat Psychol* 2011;64(3):439–461.
24. Zhang P. A simple formula for sample size calculation in equivalence studies. *J Biopharm Stat* 2003;13(3):529–538.
25. Stucke K, Kieser M. A general approach for sample size calculation for the three-arm 'gold standard' non-inferiority design. *Stat Med* 2012;31(28):3579–3596.
26. Julious SA, Owen RJ. A comparison of methods for sample size estimation for non-inferiority studies with binary outcomes. *Stat Methods Med Res* 2011;20(6):595–612.
27. Obuchowski NA. Sample size tables for receiver operating characteristic studies. *AJR Am J Roentgenol* 2000;175(3):603–608.
28. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol* 1991;44:763–770.
29. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;85(3):257–268.
30. Jewell NP. *Statistics for epidemiology*. Boca Raton: Chapman and Hall, 2004, p. 68.

SECTION II

Study Designs

CHAPTER 7

Designing Cross-Sectional and Cohort Studies

Stephen B. Hulley, Steven R. Cummings, and Thomas B. Newman

Observational studies have two primary purposes: **descriptive**, examining the distributions of predictors and outcomes in a population, and **analytic**, characterizing associations between these predictor and outcome variables. In this chapter we present two basic observational designs, which are categorized according to the **time frame** for making the measurements.

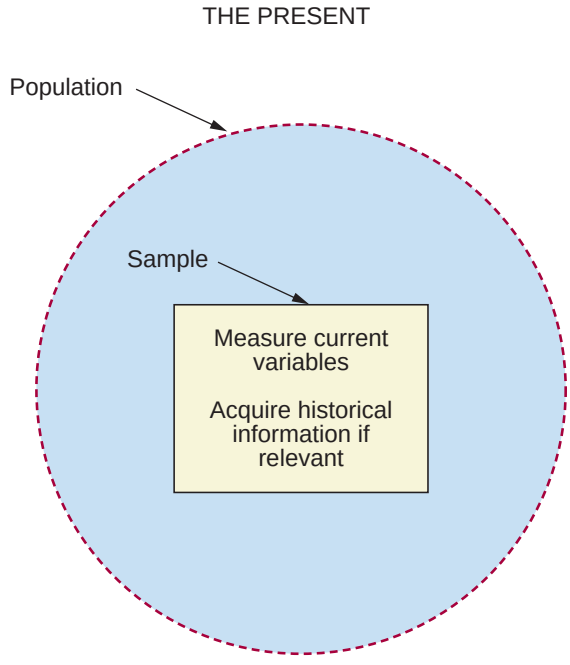
In a **cross-sectional** study, the investigator makes all of her measurements on a single occasion or within a short period of time. She draws a sample from the population and looks at distributions of variables within that sample, sometimes designating them as predictors and outcomes based on biologic plausibility and historical information. For example, if she is interested in studying the relationship between body weight and blood pressure she could measure these variables at a single clinic visit for each study subject and examine whether subjects with higher body weights were more likely to have hypertension.

In a **cohort study** measurements take place over a period of time in a group of participants who have been identified at the beginning of the study ("the cohort"). Thus, the defining characteristic of cohort studies is that a group **assembled at the outset** is followed **longitudinally**. For example the investigator could measure body weight and blood pressure on a cohort of study subjects at an initial clinic visit and then follow them for 5 years to determine the relationship between baseline weight and the incidence of hypertension. In this chapter we discuss **prospective** and **retrospective cohort** designs and **multiple-cohort** designs. We also address **statistical analysis** approaches, and the importance of optimizing **cohort retention** during follow-up.

CROSS-SECTIONAL STUDIES

In a cross-sectional study all the measurements are made at about the same time, with no follow-up period (Figure 7.1). Cross-sectional designs are well suited to the goal of describing variables and their distribution patterns. In the National Health and Nutrition Examination Survey (NHANES), for example, a sample designed to represent the entire U.S. population aged 1–74 was interviewed and examined in the early 1970s. This cross-sectional study was a major source of information about the health and habits of the U.S. population in the year it was carried out, providing estimates of such things as the prevalence of smoking in various demographic groups. Subsequent cross-sectional NHANES surveys have been carried out periodically, and all NHANES data sets are available for public use (www.cdc.gov/nchs/nhanes.htm).

Cross-sectional studies can be used for examining associations, although the choice of which variables to label as predictors and which as outcomes depends on the cause-and-effect hypotheses of the investigator rather than on the study design. This choice is easy for constitutional factors such as age, race, and sex; these cannot be altered by other variables and therefore are always predictors. For other variables, however, the choice can go either way. For example, in NHANES III there was a cross-sectional association between childhood obesity and hours spent



■ **FIGURE 7.1** In a cross-sectional study, the steps are to:

- Define selection criteria and recruit a sample from the population.
- Measure current values of predictor and outcome variables, often supplemented by historical information.

watching television (1). Whether to label obesity or television-watching as the predictor and the other as the outcome depends on the causal hypothesis of the investigator.

Unlike cohort studies, which have a longitudinal time dimension and can be used to estimate **incidence** (the proportion who *develop* a disease or condition over time), cross-sectional studies provide information about **prevalence**, the proportion who *have* a disease or condition at one point in time. Prevalence matters to a clinician, who must estimate the likelihood that the patient sitting in her office has a particular disease; the greater the prevalence, the greater the "prior probability" of the disease (the probability before the results of various diagnostic tests are available; see Chapter 12). That's why more patients with knee pain have osteoarthritis than palindromic rheumatism. Prevalence is also useful to health planners who want to know how many people have certain diseases so that they can allocate enough resources to care for them. When analyzing cross-sectional studies, the prevalence of the outcome can be compared in those with and without an exposure, yielding the **relative prevalence** of the outcome, the cross-sectional equivalent of relative risk (see Appendix 8A for examples).

Sometimes cross-sectional studies describe the prevalence of ever having done something or ever having had a disease or condition. In that case, it is important to make sure that follow-up time is the same in those exposed and unexposed.This is illustrated in Example 7.1, in which the prevalence of ever having tried smoking was studied in a cross-sectional study of children with differing levels of exposure to movies in which the actors smoke. Of course, children who had seen more movies were also older, and therefore had longer to try smoking, so it was important to adjust for age in multivariate analyses (see Chapter 9).

Strengths and Weaknesses of Cross-Sectional Studies

A major advantage of cross-sectional studies is that there is no waiting around for the outcome to occur. This makes them fast and inexpensive, and avoids the problem of loss to follow-up. Another advantage is that a cross-sectional study can be included as the first step in a cohort

EXAMPLE 7.1 Cross-Sectional Study

Sargent et al. (2) sought to determine whether exposure to movies in which the actors smoke is associated with smoking initiation. The steps in performing the study were to:

1. **Define selection criteria and recruit the population sample.** The investigators did a random-digit-dial survey of 6,522 U.S. children aged 10 to 14 years.
2. **Measure the predictor and outcome variables.** They quantified smoking in 532 popular movies and for each subject asked which of a randomly selected subset of 50 movies they had seen. Subjects were also asked about a variety of covariates such as age, race, gender, parent smoking and education, sensation-seeking (e.g., "I like to do dangerous things"), and self-esteem (e.g., "I wish I were someone else"). The outcome variable was whether the child had ever tried smoking a cigarette.

The prevalence of ever having tried smoking varied from 2% in the lowest quartile of movie smoking exposure to 22% in the highest quartile. After adjusting for age and other confounders, these differences were statistically significant; the authors estimated that 38% of smoking initiation was attributable to exposure to movies in which the actors smoke.

study or clinical trial at little or no added cost. The results define the demographic and clinical characteristics of the study group at baseline and can sometimes reveal cross-sectional associations of interest.

However, as previously noted, it's often difficult to establish causal relationships from cross-sectional data. Cross-sectional studies are also impractical for the study of rare diseases, unless the sample is drawn from a population of diseased patients rather than the general population. A **case series** of this sort is better suited to describing the characteristics of the disease than to analyzing differences between these patients and healthy people, although informal comparisons with prior experience can sometimes identify very strong risk factors. In a case series of the first 1,000 patients with AIDS, for example, 727 were homosexual or bisexual males and 236 were injection drug users (3). It did not require a formal control group to conclude that these groups were at increased risk. Furthermore, within a sample of persons with a disease there may be associations of interest, e.g., the higher risk of Kaposi's sarcoma among patients with AIDS who were homosexual than among those who were injection drug users.

Because cross-sectional studies measure only prevalence, rather than incidence, it is important to be cautious when drawing inferences about the causes, prognosis, or natural history of a disease. A factor that is associated with prevalence of disease may be a cause of the disease but could also just be associated with duration of the disease. For example, the prevalence of chronic renal failure is affected not only by its incidence, but also by survival once it has occurred. Given the observation that obesity is associated with greater survival among dialysis patients (4), a cross-sectional study of the predictors of chronic renal failure might overestimate the association between obesity and renal failure.

Serial Surveys

Occasionally, investigators perform a series of cross-sectional studies in the same population, say every 5 years. This design can be used to draw inferences about changing patterns over time. For example, Zito et al. (5), using annual cross-sectional surveys, reported that the prevalence of prescription psychotropic drug use among youth (<20 years old) increased more than

threefold between 1987 and 1996 in a mid-Atlantic Medicaid population. **Serial cross-sectional surveys** have a longitudinal time frame, but they are not the same as a cohort study, because a new sample is drawn each time. As a result, changes within individuals cannot be assessed, and findings may be influenced by people entering or leaving the population (and, thus, the samples) due to births, deaths, and migration.

COHORT STUDIES

Prospective Cohort Studies

Cohort was the Roman term for a group of soldiers that marched together, and in clinical research a cohort is a group of subjects, specified at the outset and followed over time. In a **prospective cohort study**, the investigator begins by assembling a sample of subjects (Figure 7.2). She measures characteristics in each subject that might predict the subsequent outcomes, and follows these subjects with periodic measurements of the outcomes of interest (Example 7.2).

Strengths and Weaknesses of Prospective Cohort Studies

A major advantage of the **cohort design** is that, unlike cross-sectional designs, it allows the calculation of **incidence**—the number of new cases of a condition occuring over time (Table 7.1). Measuring levels of the predictor before the outcome occurs establishes the time sequence of the variables, which strengthens the process of inferring the causal basis of an association. The prospective approach also prevents the predictor measurements from being influenced by the

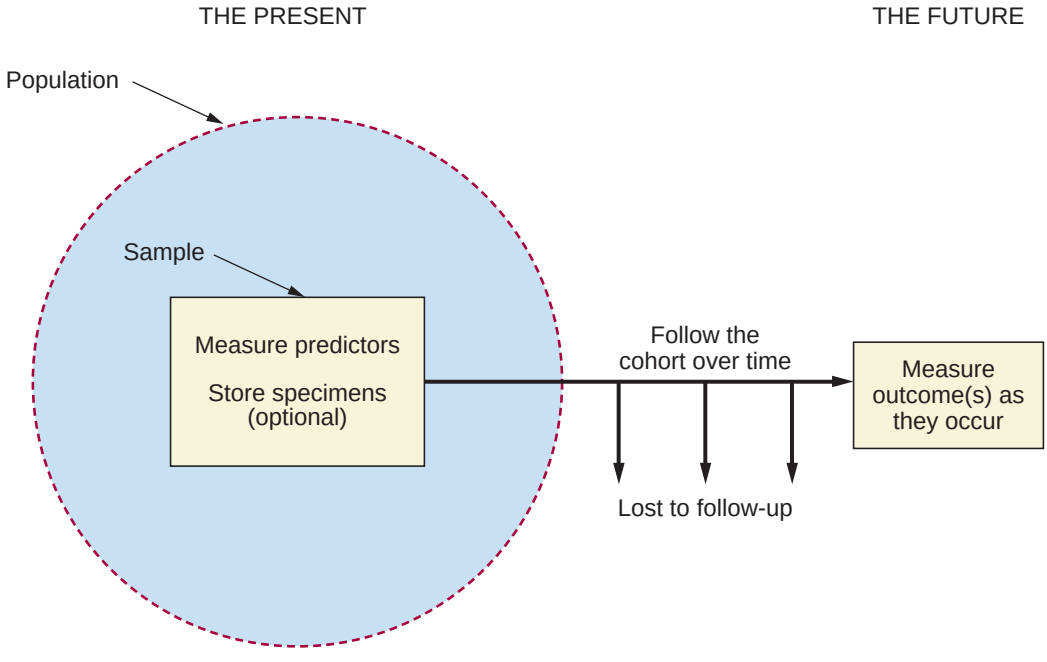


FIGURE 7.2 In a prospective cohort study, the steps are to:

- Define selection criteria and recruit a sample from the population ("the cohort").
- Measure the predictor variables and, if appropriate, the baseline level of the outcome variable.
- Consider the option to store specimens, images, etc. for later analysis of predictors.
- Follow the cohort over time, minimizing loss to follow-up.
- Measure the outcome variable(s) during follow-up.

EXAMPLE 7.2 Prospective Cohort Study

The classic Nurses' Health Study examines incidence and risk factors for common diseases in women. The steps in performing the study were to:

1. **Define selection criteria and assemble the cohort.** In 1976, the investigators obtained lists of registered nurses aged 25 to 42 in the 11 most populous states and mailed them an invitation to participate in the study; those who agreed became the cohort.
2. **Measure predictor variables, including potential confounders.** They mailed a questionnaire about weight, exercise, and other potential risk factors and obtained completed questionnaires from 121,700 nurses. They sent questionnaires periodically to ask about additional risk factors and update the status of some risk factors that had been measured previously.
3. **Follow-up the cohort and measure outcomes.** The periodic questionnaires also included questions about the occurrence of a variety of disease outcomes, which were validated by the investigators.

The prospective approach allowed investigators to make measurements at baseline and collect data on subsequent outcomes. The large size of the cohort and long period of follow-up provided substantial statistical power to study risk factors for cancers and other diseases.

For example, the investigators examined the hypothesis that gaining weight increases a woman's risk of breast cancer after menopause (6). The women reported their weight at age 18 in an early questionnaire, and follow-up weights in later questionnaires. The investigators succeeded in following 95% of the women, and 1,517 breast cancers were confirmed during the next 12 years. Heavier women had a higher risk of breast cancer after menopause, and those who gained more than 20 kg since age 18 had a twofold increased risk of developing breast cancer (relative risk = 2.0; 95% confidence interval, 1.4 to 2.8). Adjusting for potential confounding factors did not change the result.

outcome or knowledge of its occurrence and it allows the investigator to measure variables more completely and accurately than is usually possible retrospectively. This is important for predictors such as dietary habits that are difficult for a subject to remember accurately. When fatal diseases are studied retrospectively, predictor variable measurements about the decedent can only be reconstructed from indirect sources such as medical records or friends and relatives.

All cohort studies share the general disadvantage of observational studies (relative to clinical trials) that causal inference is challenging and interpretation often muddied by the influences of confounding variables (Chapter 9). A particular weakness of the prospective design is its

TABLE 7.1 STATISTICS FOR EXPRESSING DISEASE FREQUENCY IN OBSERVATIONAL STUDIES

TYPE OF STUDY	STATISTIC	DEFINITION
Cross-sectional	Prevalence	$\frac{\text{Number of people who } \textit{have} \text{ a disease or condition at a given point in time}}{\text{Number of people at risk}}$
Cohort	Incidence rate	$\frac{\text{Number of people who } \textit{get} \text{ a disease or condition}}{\text{Number of people at risk} \times \text{time period at risk}}$

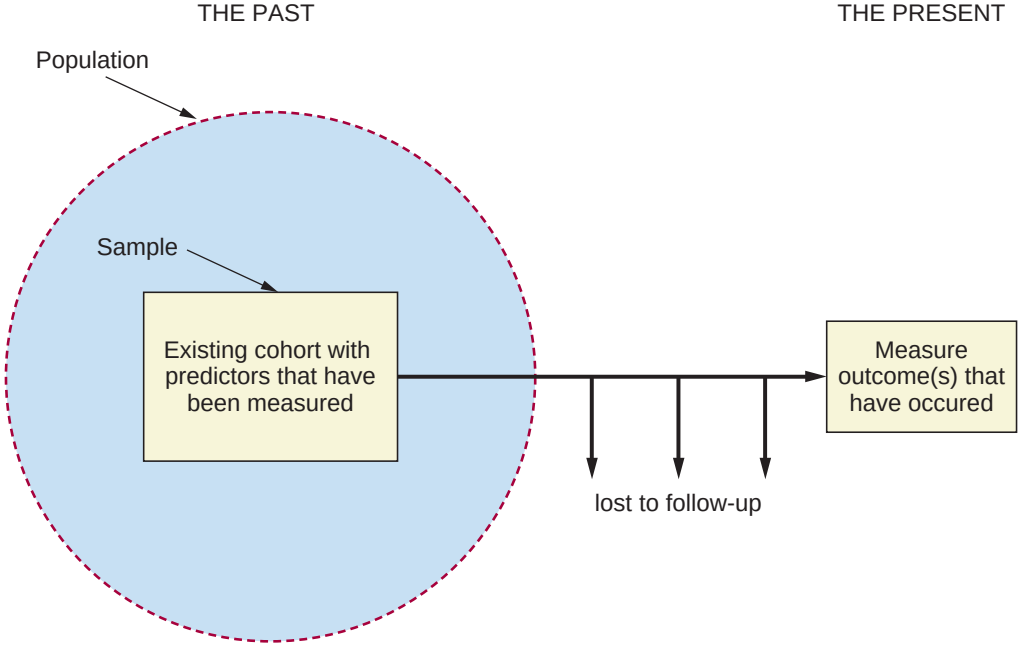
expense and inefficiency for studying rare outcomes. Even diseases we think of as relatively common, such as breast cancer, happen at such a low rate in any given year that large numbers of people must be followed for long periods of time to observe enough outcomes to produce meaningful results. Cohort designs are more efficient for dichotomous outcomes that are more common and immediate, and for continuous outcomes.

Retrospective Cohort Studies

The design of a **retrospective cohort** study (Figure 7.3) differs from that of a prospective one in that the assembly of the cohort, baseline measurements, and follow-up have all happened in the past. This type of study is only possible if adequate data about the predictors are available on a cohort of subjects that has been assembled for other purposes, such as an electronic clinical or administrative database (Example 7.3).

Strengths and Weaknesses of Retrospective Cohort Studies

Retrospective cohort studies have many of the strengths of prospective cohort studies, and they have the advantage of being much less costly and time-consuming. The subjects are already assembled, baseline measurements have already been made, and the follow-up period has already taken place. The main disadvantages are the limited control the investigator has over the approach to sampling and follow-up of the population, and over the nature and the quality of the baseline measurements. The existing data may be incomplete, inaccurate, or measured in ways that are not ideal for answering the research question.



■ **FIGURE 7.3** In a retrospective cohort study, the cohort selection and follow-up have occured in the past, so the steps are to:

- Identify an existing cohort that has some predictor information already recorded.
- Assess loss to follow-up that has occurred.
- Measure the outcome variable(s) that have already occurred.

EXAMPLE 7.3 Retrospective Cohort Study

Pearce et al. used UK National Health Service Central Registry data to describe the risk of leukemia and brain tumors associated with head CT scans in childhood (7). The steps in performing the study were to:

1. **Identify a suitable existing cohort.** The cohort consisted of 178,604 children and young adults aged <22 who received head CT scans between 1985 and 2002.
2. **Collect predictor variable data.** The investigators reviewed the records to collect gender, age, numbers, and types of radiology procedures and estimated radiation dose.
3. **Collect outcome data.** To avoid inclusion of CT scans related to cancer diagnosis, the investigators recorded leukemia occurring at least 2 years after the first CT, and brain tumors at least 5 years after the first CT, through 2008.

Childhood CT scans significantly increased the risk of leukemia and brain cancer, and the increase was dose-related; cumulative doses of 50–60 mGy tripled the risk of both leukemia and brain cancer. However, the absolute increase in risk was low, one excess case of each outcome per 10,000 head scans. The investigators, while noting that the benefits of the CT scans likely outweighed these risks, urged that radiation doses from CT scans be kept as low as possible in children, and that alternative procedures that avoid ionizing radiation be considered whenever appropriate.

Multiple-Cohort Studies and External Controls

Multiple-cohort studies begin with two or more separate samples of subjects: typically, one group with **exposure** to a potential risk factor and one or more other groups with no exposure or a lower level of exposure (Figure 7.4). After defining suitable cohorts with different levels of exposure to the predictor of interest, the investigator measures other predictor variables, follows up the cohorts, and assesses outcomes as in any other type of cohort study (Example 7.4).

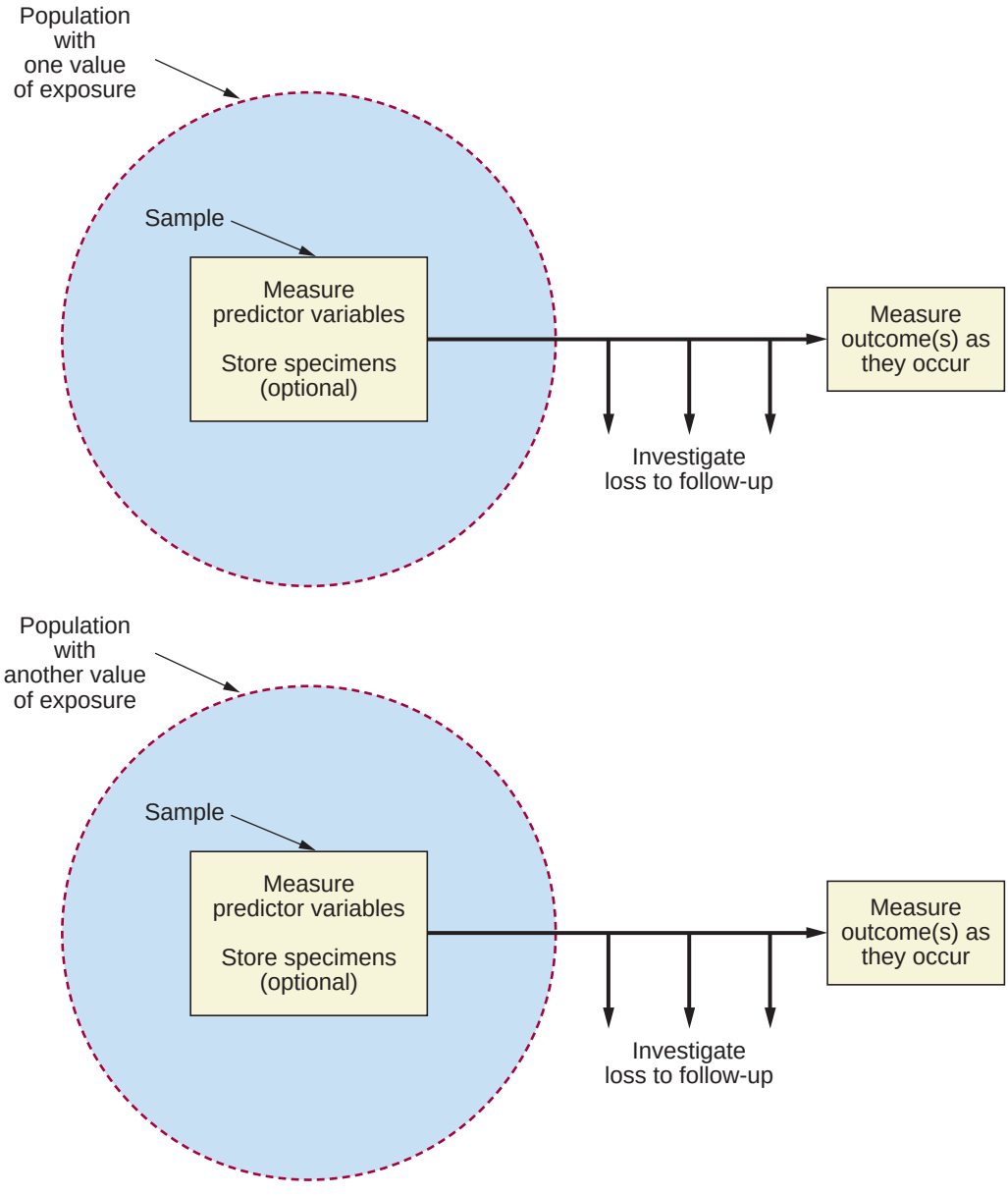
The use of two different samples of subjects in a **double-cohort design** should not be confused with the use of two samples in the case–control design (Chapter 8). In a double-cohort

EXAMPLE 7.4 Multiple-Cohort Design

To determine whether substantial neonatal jaundice or dehydration has adverse effects on neurodevelopment, investigators from UCSF and Kaiser Permanente of Northern California (8, 9) undertook a triple-cohort study. The steps in performing the study were to:

1. **Identify cohorts with different exposures.** The investigators used electronic databases to identify term and near-term newborns who
   1. had a maximum total serum bilirubin level of ≥ 25 mg/dL, or
   2. were readmitted for dehydration with a serum sodium of ≥ 150 mEq/L or weight loss of ≥ 12% from birth, or
   3. were randomly selected from the birth cohort
2. **Collect outcome data.** The investigators used electronic databases to search for diagnoses of neurological disorders and did full neurodevelopmental examinations at the age of 5 for consenting participants (blinded to which of the three cohorts the participant belonged to).

Neither hyperbilirubinemia nor dehydration was associated with adverse outcomes.



■ **FIGURE 7.4** In a double-cohort study (which can be conducted either prospectively or retrospectively) the steps are to:

- Select two or more cohorts from populations with different levels of the exposure (main predictor).
- Measure other predictors.
- Measure outcome variables during follow-up.

study the two groups of subjects are chosen based on the level of a predictor, whereas in a case–control study the two groups are chosen based on the presence or absence of an outcome.

In a variation on the multiple-cohort design, the outcome rate in a cohort can be compared with outcome rates in **census or registry** data from different populations. For example, in a classic study of whether uranium miners had an increased incidence of lung cancer, Wagoner et al. (10) compared the incidence of respiratory cancer in 3,415 uranium miners with that of

white men who lived in the same states. The increased incidence of lung cancer observed in the miners helped establish occupational exposure to ionizing radiation as an important cause of lung cancer.

Strengths and Weaknesses of Multiple-Cohort Designs

The multiple-cohort design may be the only feasible approach for studying rare exposures to potential occupational and environmental hazards. Using data from a census or registry as the external control group has the additional advantage of being population-based and economical. Otherwise, the strengths of this design are similar to those of other cohort studies.

The problem of **confounding** is accentuated in a multiple-cohort study because the cohorts are assembled from separate populations that can differ in important ways (besides exposure to the predictor variable) that influence the outcomes. Although some of these differences, such as age and race, can be matched or used to adjust the findings statistically, other characteristics may not be measurable and create problems in the interpretation of observed associations.

STATISTICAL APPROACH TO COHORT STUDIES

Risks, odds, and **rates** are estimates of the frequency of a dichotomous outcome in subjects who have been followed for a period of time. These three measures are closely related, sharing the same numerator—the number of subjects who develop the dichotomous outcome. Implicit in these three measures is the concept of being *at risk,* which means that the subject did not already have the outcome of interest at the beginning of the study. In a prospective study of the predictors of diabetes, a woman who had diabetes at baseline would not be at risk, since she already had the outcome of interest. On the other hand, there are episodic diseases, like heart failure requiring admission to a hospital, in which the outcome of interest may be the "incident" occurrence of a new episode, even if it occurs in someone who already has the disease.

Consider a study of 1,000 people who were followed for 2 years to see who developed lung cancer, and among whom eight new cases occurred each year. Risk, odds, and rate are shown in Table 7.2.

Of the three measures, risk is the easiest to understand because of its everyday familiarity—the risk of getting lung cancer in two years was 16 out of a thousand. Odds are harder to grasp intuitively—the odds of getting lung cancer were 16 to 984; fortunately, for rare outcomes (as in this case) the odds are quantitatively similar to risk and have no particular advantage. In studies comparing two groups the **odds ratio** is also similar to the **risk ratio** *when the outcome is*

TABLE 7.2 CALCULATION OF RISK, ODDS, AND RATE FOR A STUDY OF 1,000 PEOPLE FOLLOWED FOR TWO YEARS, WITH EIGHT NEW CASES OF LUNG CANCER EACH YEAR

STATISTIC	FORMULA	EXAMPLE
Risk	$\frac{N \text{ who develop the outcome}}{N \text{ at risk}}$	$\frac{16}{1{,}000} = 0.016$
Odds	$\frac{N \text{ who develop the outcome}}{N \text{ who do not develop the outcome}}$	$\frac{16}{984} = 0.0163$
Rate*	$\frac{N \text{ who develop the outcome}}{\text{Person-time at risk}}$	$\frac{16 \text{ cases}}{1{,}992 \text{ person-years}} = 0.008 \text{ cases / Person-year}$

*The denominator for the rate is the number at risk in the first year (1,000), plus the number at risk in the second (992).

rare, and this fact has unique importance in two situations: It is the basis for logistic regression calculations, and it is used to approximate relative risk in case–control studies (Appendix 8B). **Rates**, which take into account the accumulation of events over the course of time, are expressed as numbers of events divided by **person-time** at risk—the total amount of follow-up for each of the study subjects so long as that individual is alive, remains in the study, and has not yet had the outcome.

In some cohort studies significant **loss to follow-up**, unequal follow-up, or deaths or other events that preclude ascertainment of the outcome may occur. In these cases it is helpful to compare **incidence rates** between the groups—the number of outcomes divided by the person-time at risk. Each subject in the study contributes months or years of person-time from entry into the cohort until she either develops the outcome of interest or is "**censored**" due to loss to follow-up or death. The incidence rate in any group in the study is the number of outcomes in that group divided by the sum of that group's person-time at risk. As is true for the **risk ratio** (also known as relative risk), the **rate ratio** can be estimated as the quotient of rates in people who do and do not have a particular risk factor. The Cox proportional hazard model provides a method for multivariate analysis of data of this form (sometimes called "time to event" data); it allows estimation of **hazard ratios**, which are similar to rate ratios and have come into widespread use as the measure of association in **Cox regression analyses**.

Other Cohort Study Issues

The hallmark of a cohort study is the need to define the cohort of subjects at the *beginning* of a period of follow-up. The subjects should be appropriate to the research question and available for follow-up. They should sufficiently resemble the population to which the results will be generalized. The number of subjects should provide adequate power.

The quality of the study will depend on the precision and accuracy of the measurements of predictor and outcome variables (Chapter 4). The ability to draw inferences about cause and effect will depend on the degree to which the investigator has measured all **potential confounders** (Chapter 9), and the ability to generalize to subgroups of the population will depend on the degree to which the investigator has measured all **sources of effect modification**. Predictor variables may change during the study; whether and how frequently measurements should be repeated depends on cost, how much the variable is likely to change, and the importance to the research question of observing these changes. Outcomes should be assessed using standardized criteria, and when their assessment could be influenced by awareness of key risk factors, it is helpful if those making the assessments can be blinded to that predictor.

Follow-up of the entire cohort is important, and prospective studies should take a number of steps to achieve this goal (Table 7.3). Subjects who plan to move out of reach during the study or who will be difficult to follow for other reasons should be excluded at the outset. The investigator should collect information early on that she can use to find subjects who move or die, including the address, telephone number, and e-mail address of the subject, her personal physician, and at least two close friends or relatives who do not live in the same house. Mobile telephone numbers and personal e-mail addresses are particularly helpful, as they often remain unchanged when subjects, friends, or family move or change jobs. If feasible, obtaining the social security number will help in determining the vital status of those lost to follow-up, and obtaining hospital discharge information from the Social Security Administration for subjects who receive Medicare. Periodic contact with the subjects once or twice a year helps in keeping track of them, and may improve the timeliness and accuracy of recording the outcomes of interest. Finding subjects for follow-up assessments sometimes requires persistent and repeated efforts by mail, e-mail, telephone, or even house calls.

TABLE 7.3 STRATEGIES FOR MINIMIZING LOSSES DURING FOLLOW-UP

During enrollment

1. Exclude those likely to be lost:
   a. Planning to move
   b. Uncertainty about willingness to return
   c. Ill health or fatal disease unrelated to research question
2. Obtain information to allow future tracking:
   a. Address, telephone number (mobile phone numbers are particularly useful), and e-mail address of subject
   b. Social Security/Medicare number
   c. Name, address, telephone number, and e-mail address of close friends or relatives who do not live with the subject
   d. Name, address, telephone number, and email address of physician(s)

During follow-up*

1. Periodic contact with subjects to collect information, provide results, and be supportive:
   a. By telephone: may require calls during weekends and evenings
   b. By mail: repeated mailings by e-mail or with stamped, self-addressed return cards
   c. Other: newsletters, token gifts
2. For those who are not reached by phone or mail:
   a. Contact friends, relatives, or physicians
   b. Request forwarding addresses from postal service
   c. Seek address through other public sources, such as telephone directories and the Internet, and ultimately a credit bureau search
   d. For subjects receiving Medicare, collect data about hospital discharges from the Social Security Administration
   e. Determine vital status from state health department or National Death Index

At all times

1. Treat study subjects with appreciation, kindness, and respect, helping them to understand the research question so they will want to join as partners in making the study successful.

*This assumes that participants in the study have given informed consent to collect the tracking information and for follow-up contact.

SUMMARY

1. In a **cross-sectional** study, the variables are all measured at a single point in time, with no structural distinction between predictors and outcomes. Cross-sectional studies yield **weaker evidence for causality** than cohort studies because the predictor variable is not shown to precede the outcome.
2. Cross-sectional studies are valuable for providing descriptive information about **prevalence**, and have the advantage of **avoiding the time**, **expense**, and **dropout problems** of a follow-up design; they are often useful as the first step of a cohort study or experiment, and can be linked in independently sampled **serial surveys** to reveal population changes over time.
3. Cross-sectional studies require a large sample size when studying uncommon diseases and variables in the general population, but can be useful in a **case series** of an uncommon disease.
4. In **cohort studies**, a group of subjects identified at the outset is followed over time to describe the **incidence** or natural history of a condition and to discover **predictors** (risk factors) for various outcomes. The ability to measure the predictor before the outcome occurs establishes the sequence of events and controls bias in that measurement.
5. **Prospective cohort** studies begin at the outset of follow-up and may require large numbers of subjects followed for long periods of time. The latter disadvantage can sometimes be

overcome by identifying a **retrospective cohort** in which measurements of predictor variables have already occurred.

6. The **multiple-cohort** design, which compares the incidence of outcomes in cohorts that differ in the level of a predictor variable ("the **exposure**"), is useful for studying the effects of rare and occupational exposures.
7. **Risks**, **odds**, and **rates** are three ways to estimate the frequency of a dichotomous outcome during follow-up; among these, **incidence rates**, which take into account person-time of participants who remain alive and event-free in the study, are the basis for modern approaches to calculating **multivariate hazard ratios** using Cox proportional hazard models.
8. Inferences about **cause and effect** are strengthened by measuring and adjusting for all conceivable potential confounding variables. Bias in the assessment of outcomes is prevented by **standardizing** the measurements and **blinding** those assessing the outcome to the predictor variable values.
9. The strengths of a cohort design can be undermined by incomplete **follow-up** of subjects. Losses can be minimized by **excluding subjects** at the outset who may not be available for follow-up, by collecting **baseline information** that facilitates tracking, and by **staying in touch** with all subjects regularly.

REFERENCES

1. Andersen RE, Crespo CJ, Bartlett SJ, et al. Relationship of physical activity and television watching with body weight and level of fatness among children: results from the Third National Health and Nutrition Examination Survey. *JAMA* 1998;279(12):938–942.
2. Sargent JD, Beach ML, Adachi-Mejia AM, et al. Exposure to movie smoking: its relation to smoking initiation among US adolescents. *Pediatrics* 2005;116(5):1183–1191.
3. Jaffe HW, Bregman DJ, Selik RM. Acquired immune deficiency syndrome in the United States: the first 1,000 cases. *J Infect Dis* 1983;148(2):339–345.
4. Kalantar-Zadeh K, Abbott KC, Salahudeen AK, et al. Survival advantages of obesity in dialysis patients. *Am J Clin Nutr* 2005; 81: 543–554.
5. Zito JM, Safer DJ, DosReis S, et al. Psychotropic practice patterns for youth: a 10-year perspective. *Arch Pediatr Adolesc Med* 2003;157(1):17–25.
6. Huang Z, Hankinson SE, Colditz GA, et al. Dual effect of weight and weight gain on breast cancer risk. *JAMA* 1997;278:1407–1411.
7. Pearce MS, Salotti JA, Little MP, et al. Radiation exposure from CT scans in childhood and subsequent risk of leukemia and brain tumors: a retrospective cohort study. *Lancet* 2012;380:499–505.
8. Newman TB, Liljestrand P, Jeremy RJ, et al. Outcomes of newborns with total serum bilirubin levels of 25 mg/dL or more. *N Engl J Med* 2006;354:1889–1900.
9. Escobar GJ, Liljestrand P, Hudes ES, et al. Five-year neurodevelopmental outcome of neonatal dehydration. *J Pediatr* 2007;151(2):127–133, 133 e1.
10. Wagoner JK, Archer VE, Lundin FE, et al. Radiation as the cause of lung cancer among uranium miners. *N Engl J Med* 1965;273:181–187.

CHAPTER 8

Designing Case–Control Studies

Thomas B. Newman, Warren S. Browner, Steven R. Cummings, and Stephen B. Hulley

In Chapter 7 we introduced cohort studies, in which the sequence of the measurements is the same as the chronology of cause and effect: predictor variables are measured first, then outcomes are observed during follow-up. In contrast, in a **case–control** study the investigator works backward. She begins by choosing one sample of people with the outcome (the cases) and another sample of people without that outcome (the controls); she then compares the levels of predictor variables in the two samples to see which predictors are associated with the outcome. For example, a case–control study might involve assembling a group of cases of ocular melanoma and a sample of healthy controls, followed by gathering data from each group about previous exposure to arc welding to estimate how that exposure affects the risk of ocular melanoma. The case–control design is relatively **inexpensive** and uniquely **efficient** for studying **rare diseases**.

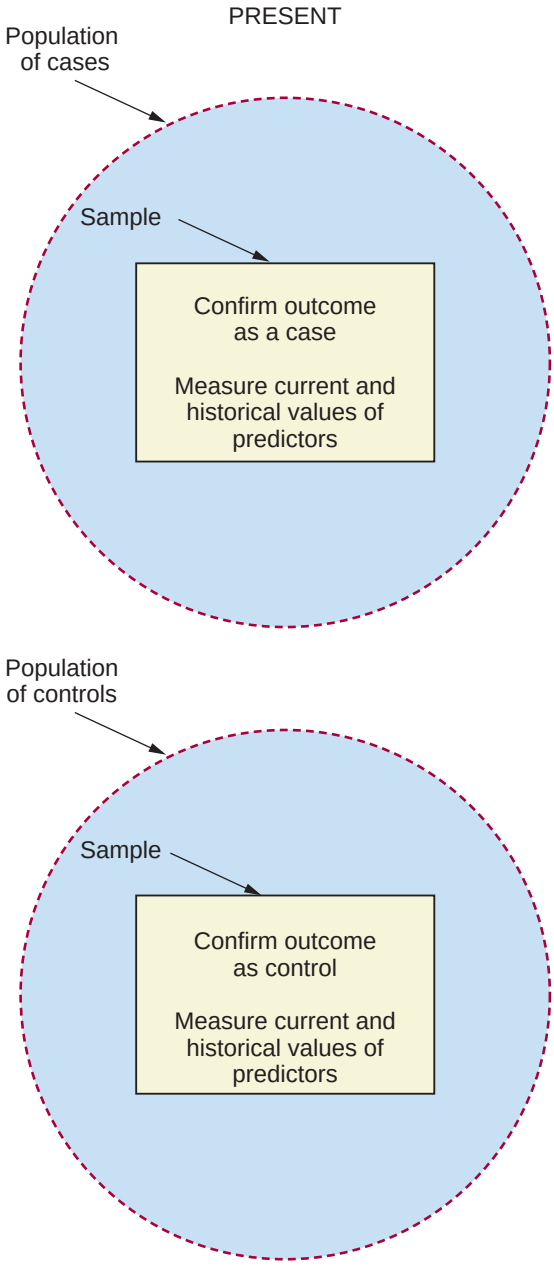
This chapter also presents several variations on the simple case–control design noted above. A ***nested* case–control** design compares the incident cases nested in a cohort study with controls drawn at random from the rest of the cohort; this design controls sampling and measurement bias and saves money if the predictors are expensive measurements that can be made on stored specimens or images collected at the outset of the cohort study. An ***incidence-density* case–control** design allows investigators to analyze risk relationships, taking into account changes over time in risk factor levels and loss to follow-up. And a **nested case–cohort** design allows a random sample of the entire cohort to serve as the control for several different sets of cases. The chapter ends with advice on choosing among the observational study designs discussed in Chapters 7 and 8.

CASE–CONTROL STUDIES

Because most diseases are relatively uncommon, both cohort and cross-sectional studies of general population samples are expensive designs, requiring thousands of subjects to identify risk factors for a rare disease like stomach cancer. As noted in Chapter 7, a **case series** of patients with the disease can identify an obvious risk factor (such as injection drug use for AIDS), using prior knowledge of the prevalence of the risk factor in the general population. For most risk factors, however, it is necessary to assemble a reference group, so that exposure to the risk factor in subjects with the disease (cases) can be compared with exposure to the risk factor among subjects without the disease (controls).

Case–control studies are **retrospective** (Figure 8.1). The study identifies one group of subjects with the disease and another without it, then looks backward to find differences in predictor variables that may explain why the cases got the disease and the controls did not (Example 8.1).

Case–control studies began as epidemiologic studies to identify risk factors for diseases. For this reason, and because it makes the discussion easier to follow, we generally refer to "cases" as those with the disease. However, the case–control design can also be used to look at other uncommon outcomes, such as disability among those who already have a disease. In addition, when



■ **FIGURE 8.1** In a case–control study, the steps are to:

- Define selection criteria and recruit one sample from a population of cases and a second sample from a population of controls.
- Measure current values of relevant variables, often supplemented by historical information.

undesired outcomes are the rule rather than the exception, the cases in a case–control study may be the rare patients who have had a good outcome, such as recovery from a usually fatal disease.

Case–control studies are the "house red" on the research design wine list: more modest and a little riskier than the other selections, but much less expensive and sometimes surprisingly good. The design of a case–control study is challenging because of the increased opportunities for bias, but there are many examples of well-designed case–control studies that have yielded important results. These include the links between maternal diethylstilbestrol use and vaginal cancer in daughters (a classic study that provided a definitive conclusion based on just seven cases!) (1), and between prone sleeping position and sudden infant death syndrome (2), a simple result that has saved thousands of lives (3).

EXAMPLE 8.1 Case–Control Study

Because intramuscular vitamin K is given routinely to newborns in the United States, a pair of studies reporting a doubling in the risk of childhood cancer among those who had received intramuscular vitamin K caused quite a stir (4, 5). To investigate this association further, German investigators (6)

1. **Selected the sample of cases.** 107 children with leukemia from the German Childhood Cancer Registry.
2. **Selected the sample of controls.** 107 children matched by sex and date of birth and randomly selected from children living in the same town as the case at the time of diagnosis (from local government residential registration records).
3. **Measured the predictor variable.** Reviewed medical records to determine which cases and controls had received intramuscular vitamin K in the newborn period.

The authors found 69 of 107 cases (64%) and 63 of 107 controls (59%) had been treated with vitamin K, for an odds ratio of 1.3 (95% confidence interval [CI], 0.7 to 2.3). (See Appendix 8A for the calculation.) Therefore, this study did not confirm the existence of an association between the receipt of vitamin K as a newborn and subsequent childhood leukemia. The point estimate and upper limit of the 95% CI leave open the possibility of a clinically important increase in leukemia in the population from which the samples were drawn, but several other studies, and an analysis using an additional control group in the cited study, also failed to confirm the association (7, 8).

Case–control studies cannot yield estimates of the incidence or prevalence of a disease because the proportion of study subjects who have the disease is determined by how many cases and how many controls the investigator chooses to sample, rather than by their proportions in the population. Case–control studies do provide descriptive information on the characteristics of the cases and, more important, an estimate of the strength of the association between each predictor variable and the outcome. These estimates are in the form of odds ratios, which approximate the relative risk if the risk of the disease in both exposed and unexposed subjects is relatively low (about 10% or less; see Appendix 8B).

Strengths of Case–Control Studies

Efficiency for Rare Outcomes

One of the major strengths of case–control studies is their rapid, high yield of information from relatively few subjects. Consider a study of the effect of circumcision on subsequent carcinoma of the penis. This cancer is very rare in circumcised men but is also rare in uncircumcised men, whose lifetime cumulative incidence is about 0.16% (9). To do a cohort study with a reasonable chance (80%) of detecting even a very strong risk factor (say a relative risk of 50) would require following more than 6,000 men for many years, assuming that roughly equal proportions were circumcised and uncircumcised. A randomized clinical trial of circumcision at birth would require the same sample size, but the cases would occur at a median of 67 years after entry into the study—it would take three generations of investigators to follow the subjects!

Now consider a case–control study of the same question. For the same chance of detecting the same relative risk, only 16 cases and 16 controls (and not much time or effort from the investigators) would be required. For diseases that are either rare or have long latent periods between exposure and disease, case–control studies are not only far more efficient than other designs, they are often the only feasible option.

Usefulness for Generating Hypotheses

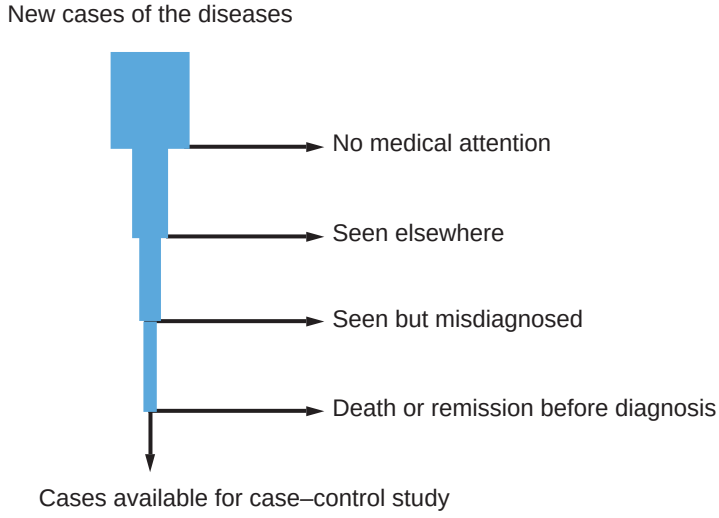
The retrospective approach of case–control studies, and their ability to examine a large number of predictor variables, makes them useful for generating hypotheses about the causes of a new outbreak of disease. For example, a case–control study of an epidemic of deaths from acute renal failure in Haitian children found an odds ratio of 53 for ingestion of locally manufactured acetaminophen syrup. Further investigation revealed that the renal failure was due to poisoning by diethylene glycol, which was found to contaminate the acetaminophen syrup (10), a problem that unfortunately has recurred (11).

Weaknesses of Case–Control Studies

Case–control studies have great strengths, but they also have major disadvantages. First, only one outcome can be studied (the presence or absence of the disease that was the criterion for drawing the two samples), whereas cohort and cross-sectional studies (and clinical trials) can study several outcome variables. Second, as mentioned, the information available in case–control studies is limited: There is no direct way to estimate the incidence or prevalence of the disease, nor the attributable or excess risk, unless the investigator also knows the exact population and time period from which the cases arose. But the biggest weakness of case–control studies is their **susceptibility to bias**. This bias comes chiefly from two sources: the **separate sampling** of the cases and controls, and the **retrospective measurement** of the predictor variables. These two problems and the strategies for dealing with them are the topic of the next two sections.

Sampling Bias and How to Control It

The sampling in a case–control study begins with the cases. Ideally, the sample of cases would include everyone who developed the disease under study, or a random selection from those cases. An immediate problem comes up, however: How do we know who has developed the disease and who has not? In cross-sectional and cohort studies the disease is systematically sought in all the study participants, but in case–control studies the cases must be sampled from patients in whom the disease has already been diagnosed and who are available for study. This sample may not be representative of all patients who develop the disease because those who are undiagnosed, misdiagnosed, unavailable for study, or dead are unlikely to be included (Figure 8.2).



■ **FIGURE 8.2** Some reasons that the cases in a case–control study may not be representative of all cases of the disease.

In general, sampling bias matters when the sample of cases is unrepresentative with respect to the risk factor being studied. Diseases that almost always require hospitalization and are straightforward to diagnose, such as hip fracture and traumatic amputation, can be sampled safely from diagnosed and accessible cases, at least in developed countries. On the other hand, conditions that may not come to medical attention are more difficult to study with case–control studies because of the selection that precedes diagnosis. For example, women seen in a gynecologic clinic with first-trimester spontaneous abortions would probably differ from the entire population of women experiencing spontaneous abortions, many of whom do not seek medical attention. Thus women with a prior history of infertility would be over-represented in a clinic-based sample, while those with poor access to prenatal care would be under-represented. If a predictor variable of interest is associated with gynecologic care in the population (such as past use of an intrauterine device [IUD]), sampling cases from the clinic could be an important source of bias. If, on the other hand, a predictor is unrelated to gynecologic care (such as blood type), there would be less likelihood of a clinic-based sample being unrepresentative.

Although it is important to think about these issues, the selection of cases is often limited to the accessible sources of subjects. The sample of cases may not be entirely representative, but it may be all that the investigator has to work with. The difficult decisions faced by an investigator designing a case–control study then relate to the more open-ended task of selecting appropriate controls. The general goal is to sample controls from the population who would have become a case in the study if they had developed the disease. Four strategies for sampling controls follow:

- **Clinic- or hospital-based controls.** One strategy to compensate for the possible selection bias caused by obtaining cases from a clinic or hospital is to select controls from the same facility or facilities. For example, in a study of past use of an IUD as a risk factor for spontaneous abortion, controls could be sampled from a population of women seeking care for other problems (e.g., vaginitis) at the same gynecologic clinic. Compared with a random sample of women from the same area, these controls would presumably better represent the population of women who, if they had a spontaneous abortion, would have come to the clinic and become a case.

  However, selection of an unrepresentative sample of controls to compensate for an unrepresentative sample of cases can be problematic. If the risk factor of interest causes a medical problem for which the controls seek care, the prevalence of the risk factor in the control group will be falsely high, diminishing or reversing the association between the risk factor and the outcome. If, for example, many women in the control group sought attention at the clinic for a medical condition associated with past use of an IUD (e.g., infertility from older models of IUDs), there would be an excess of former IUD users among the controls, reducing the size of the association between past IUD use and spontaneous abortion in the study.

  Because hospital- and clinic-based control subjects often have conditions that are associated with the risk factor(s) being studied, these types of controls can produce misleading findings. Thus it is essential to consider whether the convenience of using hospital- or clinic-based controls is worth the possible threat to the validity of the study.
- **Using population-based samples of cases and controls.** Because of the rapid increase in the use of disease registries in geographic populations and within health plans, population-based case–control studies are now possible for many diseases. Cases obtained from such registries are generally representative of the general population of patients in the area with the disease, thus simplifying the choice of a control group: It should be a representative sample of "non-cases" from the population covered by the registry. In Example 8.1, all residents of the town were registered with the local government, making selection of such a sample straightforward.

  When registries are available, population-based case–control studies are the most desirable design. As a disease registry approaches completeness and the population it covers approaches

stability (no migration in or out), a population-based case–control study approaches a case–control study that is nested within a cohort study or clinical trial (page 104) assuming that the controls can be identified and enrolled. Those latter tasks are relatively straightforward when the population has been enumerated and these records are available to investigators, as in the vitamin K and leukemia study described in Example 8.1. Lacking such registration records, a commonly used approach is random digit dialing of (landline) phone numbers with prefixes in the region covered by the registry. (When controls are selected this way, the cases who have no landline telephone need to be excluded.) With increasing numbers of households with mobile phones only, this approach has become problematic (12). Random-digit dialing including cell phone numbers is possible, but must be done carefully, immediately ending the call if the recipient is driving and avoiding calls for which the recipient might be charged (13).

It's important to recognize, however, that bias can be introduced any time subjects need to be contacted to obtain information because some subjects (say, those who do not speak English, or who are hard of hearing) may be less likely to be included. A similar problem can occur any time informed consent is needed.

- **Using two or more control groups.** Because selection of a control group can be so tricky, particularly when the cases may not be a representative sample of those with disease, it is sometimes advisable to use two or more control groups selected in different ways. The Public Health Service study of Reye's syndrome and medications (14), for example, used four types of controls: emergency room controls (seen in the same emergency room as the case), inpatient controls (admitted to the same hospital as the case), school controls (attending the same school or day care center as the case), and community controls (identified by random-digit dialing). The odds ratios for salicylate use in cases compared with each of these control groups were all at least 30 and highly statistically significant. The consistent finding of a strong association using control groups that would have different sampling biases strengthens the inference that there is a real association in the population.

  Unfortunately, few associations have odds ratios anywhere near that large, and the biases associated with different strategies for selecting controls may cause the results using different control groups to conflict with one another, thereby revealing the inherent fragility of the case–control design for the research question at hand. When this happens, the investigator should seek additional information (e.g., the chief complaint of clinic-based controls) to try to determine the magnitude of potential biases from each of the control groups (Chapter 9). In any case it is better to have inconsistent results and conclude that the answer is not known than to have just one control group and draw the wrong conclusion.
- **Matching.** Matching is a simple method of ensuring that cases and controls are comparable with respect to major factors that are related to the disease but not of interest to the investigator. So many risk factors and diseases are related to age and sex, for example, that the study results may be unconvincing unless the cases and controls are comparable with regard to these two variables. One approach to avoiding this problem is to choose controls that match the cases on these constitutional predictor variables. However, matching does have substantial disadvantages, particularly if modifiable predictors such as income or serum cholesterol level are matched. The reasons for this and the alternatives that are often preferable to matching are discussed in Chapter 9.

Differential Measurement Bias and How to Control It

The second major weakness of case–control studies is the risk of bias due to **measurement error**. This is caused by the retrospective approach to measuring the predictor variables: both cases and control may be asked to recall exposures that happened years before. Unfortunately, people's memories for past exposures are imperfect. If they are similarly imperfect in cases and

controls, the problem is called **nondifferential misclassification** of the exposure, which makes it more difficult to find associations. (In epidemiologic terms, the odds ratio is biased toward 1.) Of greater concern, however, being diagnosed with a disease may lead cases to remember or report their exposures differently from controls; this **differential misclassification** of exposure, called **recall bias**, has unpredictable effects on associations measured in a study.

For example, widespread publicity about the relationship between sun exposure and malignant melanoma might lead cases diagnosed with that cancer to recall their history of sun exposure differently from controls. Cockburn et al. (15) found some evidence of this in a clever study of twins discordant for melanoma: The matched odds ratio for sunbathing as a child was 2.2 (95% CI 1.0 to 4.7) when the twin with melanoma was asked which twin had sunbathed more as a child, but only 0.8 (0.4 to 1.8) when the co-twin without melanoma was asked the same question. However, for some other questions, such as which twin tanned or burned more easily, there was no evidence of recall bias.

Recall bias cannot occur in a cohort study because the subjects are asked about exposures before the disease has been diagnosed. A case–control study of malignant melanoma nested within a cohort with sun exposure data collected years earlier provided a direct test of recall bias: The investigators compared self-reported sun exposure in cases and controls both before and after the case was diagnosed with melanoma (16). The investigators found some inaccuracies in recollections of exposure in both cases and controls, but little evidence of recall bias (16). Thus, while it is important to consider the possibility of recall bias, it is not inevitable (17).

In addition to the strategies set out in Chapter 4 for controlling bias in measurements (standardizing the operational definitions of variables, choosing objective approaches, supplementing key variables with data from several sources, etc.), here are two specific strategies for avoiding bias in measuring exposures in case–control studies:

- **Use data recorded before the outcome occurred.** It may be possible, for example, to examine perinatal medical records in a case–control study of intramuscular vitamin K as a risk factor for cancer. This excellent strategy is limited to the extent that recorded information about the risk factor of interest is available and reliable. For example, information about vitamin K administration was often missing from medical records, and how that missing information was treated affected results of some studies of vitamin K and subsequent cancer risk (8).
- **Use blinding.** The general approach to blinding was discussed in Chapter 4, but there are some issues that are specific to designing interviews in case–control studies. In theory, both observers and study subjects could be blinded to the case–control status of each subject and to the risk factor being studied; thus, four types of blinding are possible (Table 8.1).

TABLE 8.1 APPROACHES TO BLINDING IN A CASE–CONTROL STUDY

PERSON BLINDED	BLINDING CASE–CONTROL STATUS	BLINDING RISK FACTOR MEASUREMENT
Subject	Possible if both cases and controls have diseases that could plausibly be related to the risk factor	Include "dummy" risk factors and be suspicious if they differ between cases and controls May not work if the risk factor for the disease has already been publicized
Observer	Possible if cases are not externally distinguishable from controls, but subtle signs and statements volunteered by the subjects may make it difficult	Possible if interviewer is not the investigator, but may be difficult to maintain

Ideally, neither the study subjects nor the observers should know which subjects are cases and which are controls. In practice, this is often difficult. The subjects know whether they are sick or well, so they can be blinded to case–control status only if controls are also ill with diseases that they believe might be related to the risk factors being studied. Efforts to blind interviewers are hampered by the obvious nature of some diseases (an interviewer can hardly help noticing if the subject is jaundiced or has had a laryngectomy), and by the clues that interviewers may discern in the subject's responses.

Blinding to specific risk factors being studied is usually easier than blinding to case–control status. Case–control studies are often the first step in investigating an illness, so there may not be just one risk factor of particular interest. Thus, the study subjects and the interviewer can be kept in the dark about the study hypotheses by including "dummy" questions about plausible risk factors not associated with the disease. For example, in a study of honey consumption as a risk factor for infant botulism, equally detailed questions about yogurt and bananas could be included in the interview. This type of blinding does not prevent differential bias, but it allows an estimate of whether it is a problem: If the cases report more exposure to honey but no increase in the other foods, then differential measurement bias is less likely. This strategy would not work if the association between eating honey and infant botulism had previously been widely publicized, or if some of the dummy risk factors turned out to be real ones.

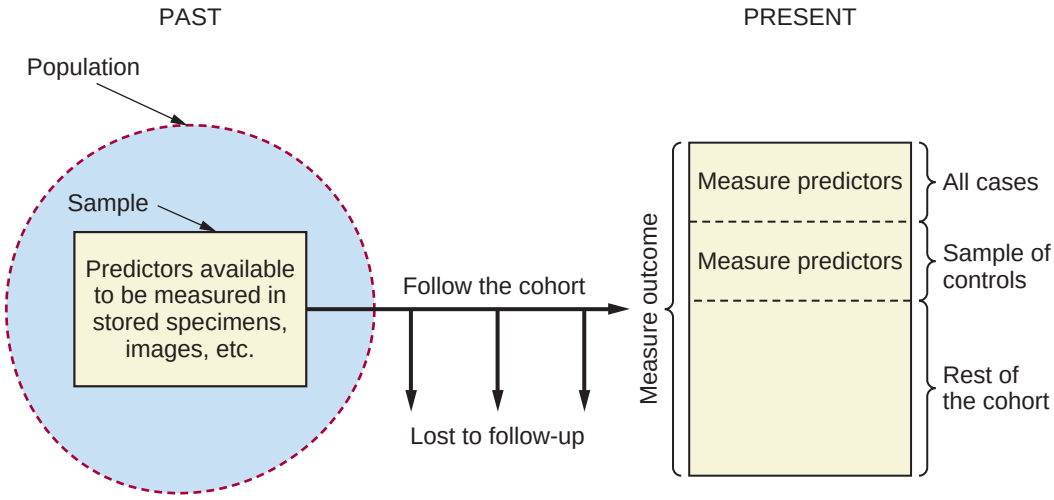
Blinding the observer to the case–control status of the study subject is a particularly good strategy for **laboratory measurements** such as blood tests and x-rays. Blinding under these circumstances is easy and should always be done, simply by having someone other than the individual who will make the measurement apply a coded identification label to each specimen (or patient). The importance of blinding was illustrated by 15 case–control studies comparing measurements of bone mass between hip fracture patients and controls; much larger differences were found in the studies that used unblinded measurements than in the blinded studies (18).

■ NESTED CASE–CONTROL, INCIDENCE-DENSITY NESTED CASE–CONTROL, AND CASE–COHORT STUDIES

A **nested case–control** design has a case–control study "nested" within a defined cohort (Figure 8.3). The cohort may already have been defined by the investigator as part of a formal cohort study, often including banking of specimens, images, and so on, to be analyzed in the future after outcomes occur. Alternatively, the investigator can design a nested case–control study *de novo*, in a cohort that is not already defined, in which case defining the cohort will be the first step.

The investigator begins by identifying a cohort of subjects at risk for the outcome that is large enough to yield sufficient numbers of cases to answer the research question, and that provides the ability to measure the exposure variable, either because specimens have been banked or medical records (or subjects) with exposure information are available. As described in Chapter 7, definition of the cohort will include the specific inclusion and exclusion criteria that define a population at risk. In addition, the **date of entry** into the cohort must be clear for each subject. This could be a fixed date (e.g., everyone meeting inclusion criteria who was enrolled in a health plan on January 1, 2008), or it could be a variable date on which a period at risk begins (e.g., the date of enrollment in a cohort study or the date of first myocardial infarction in a study of risk factors for recurrent myocardial infarction).

The investigator next describes the criteria that define the occurrence of the outcome of interest, which in all cases will be after the date of entry into the cohort and before the end of the defined follow-up period. If the outcome is rare, follow-up close to complete, and a single measurement of the exposure at baseline is sufficient, then it is simple. The investigator identifies all the individuals in the cohort who developed the outcome by the end of follow-up (the cases) and then selects a random sample of the subjects who were also part of the cohort but did not develop the outcome (the controls). The investigator then measures the predictor variables



■ **FIGURE 8.3** A *nested* case–control study can be either prospective or retrospective. For the retrospective version, the steps are to

- Identify a cohort from the population with previously stored specimens, images, and other data.
- Measure the outcome variable that distinguishes cases from controls.
- Measures predictor variables in specimens, images, and other data stored since the cohort was formed, as well as other variables, in all the cases and in a sample of the non-cases (controls).

for cases and controls, and compares levels of the risk factor in cases to the levels in the sample of controls. This is a simple nested case–control study (Example 8.2).

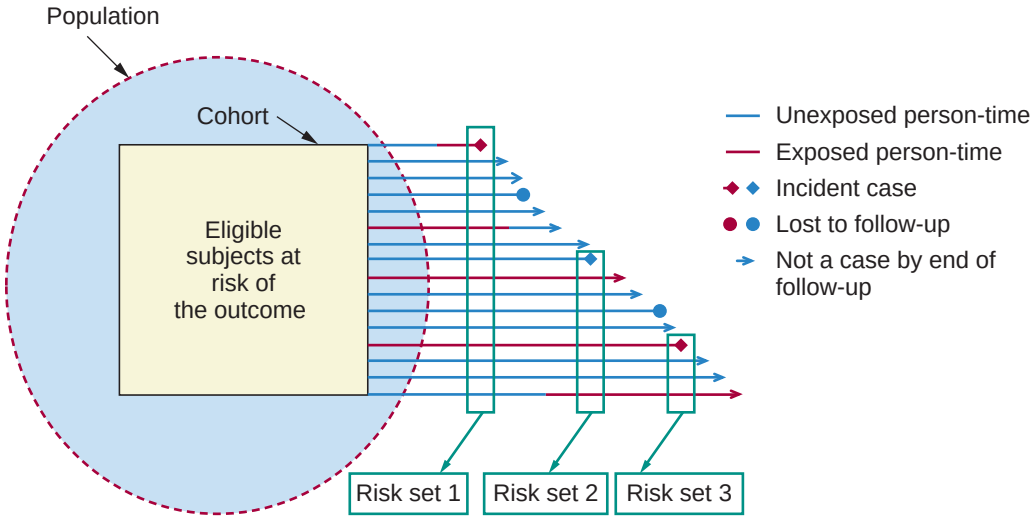
If follow-up is variable or **incomplete**, or the exposure of interest **varies over time**, a single measurement of exposure at entry into the cohort in the cases and a random sample of controls

EXAMPLE 8.2 Simple Nested Case–Control Design

To determine whether higher levels of sex hormones increased the risk of breast cancer, Cauley (19) and colleagues conducted a nested case–control study. The basic steps in performing this study were to:

1. **Identify a cohort.** The investigators used the Study of Osteoporotic Fractures (SOF) cohort. This was a good choice because serum samples of members of this cohort had been drawn by the same investigators during the baseline examination and put into frozen storage at –190°C with the expectation that just such a study would be designed.
2. **Identify cases at the end of follow-up.** Based on responses to follow-up questionnaires and review of death certificates, the investigators identified 97 subjects who had developed a first occurrence of breast cancer during 3.2 years of follow-up.
3. **Select controls.** The investigators selected a random sample of 244 women in the cohort who did not develop breast cancer during that follow-up period.
4. **Measure predictors.** Levels of sex hormones, including estradiol and testosterone, were measured in the samples of frozen serum from the baseline examination of cases and controls. The laboratory was blinded to whether the samples came from cases or controls.

Women who had high levels of either estradiol or testosterone had a threefold increase in the risk of a subsequent diagnosis of breast cancer compared with women who had very low levels of these hormones.



■ **FIGURE 8.4** An *incidence-density* nested case–control study can be either prospective or retrospective. For the prospective version, the steps are to:

- Define selection criteria and recruit a cohort from the population.
- Define the date of entry for each member of the cohort to align follow-up times.
- Store specimens, images, etc for later analysis.
- Follow the cohort to identify cases and the date they were diagnosed.
- Sample one or more controls for each case from "risk sets," defined as members of the cohort who have been followed for the same amount of time as the case and have not become a case, died, or been lost to follow-up at the time the case was diagnosed.
- Measure predictor variables in specimens, images, etc. stored since baseline, as well as other current variables, on cases and matched controls.

will not be sufficient. In that case it is better to design an ***incidence-density* nested case–control study** and sample the controls from **risk sets**, defined for each case as it occurs as the members of the cohort who were followed the same length of time as the case but had not yet become cases (Figure 8.4). As is the case for any other form of matching of controls to cases, this matching on follow-up time needs to be accounted for in the analysis.

For example, if entry in the cohort was a fixed date (e.g., January 1, 2008), the controls for a case diagnosed on July 1, 2009, would be sampled from among the subjects who had not yet developed the outcome as of July 1, 2009. If the date of entry into the cohort was variable, controls for a case diagnosed 18 months after entry would be sampled from among those who had not yet become a case after 18 months of follow-up. Depending on the research hypothesis of the investigator, values of the exposure at entry or at some point after entry could be compared between cases and controls.

This sampling according to risk sets introduces the complexity that the same subject may be selected as a control for a case that occurs early in follow-up and later become a case himself, perhaps after his value for his exposure variable changes. In effect, what this design does (with the help of appropriate statistical analysis) is sequentially consider chunks of person-time at risk, for each chunk using values of predictor variables to predict occurrence of cases in that chunk of person-time, with the boundaries of each chunk defined by the occurrence of the cases. This is called an **incidence-density** design (Example 8.3).

A **nested case–*cohort*** design is similar to the simple nested case–control design except that, instead of selecting controls who did not develop the outcome of interest, the investigator selects a random sample of all the members of the cohort, regardless of outcomes. A few subjects who are part of that random sample may have developed the outcome (the number is very small when the outcome is uncommon). An advantage of the case–cohort design is that a

single random sample of the cohort can provide the controls for several case–control studies of different outcomes. In addition, the random sample of the cohort provides information on the overall prevalence of risk factors in the cohort.

EXAMPLE 8.3 "Incidence-Density" Nested Case–Control Design

To investigate a possible association between the oral antidiabetes drug pioglitazone (Actos®) and bladder cancer, investigators from Montreal (20) performed a case–control study nested within the United Kingdom General Practice Research Database, which contains complete primary care medical records for more than 10 million people enrolled in more than 600 general practices in the UK. The steps were:

1. **Identify the cohort and time period at risk.** The investigators included adults with their first ever prescription for an oral antidiabetes drug between January 1, 1988, and December 31, 2009, who had been followed in the database for at least 1 year before that prescription and who were at least 40 years old at the time of that prescription. The date of this first antidiabetes drug prescription was the date of entry into the cohort. Participants were followed until a diagnosis of bladder cancer, death from any cause, end of registration with the general practice, or end of the study period on December 31, 2009, whichever came first. Subjects with a previous history of bladder cancer were excluded.
2. **Identify the cases, including dates of occurrence.** The investigators identified incident cases of bladder cancer using "Read codes" (a system for coding diagnoses validated in the general practice research database [21]). To account for the expectation that the effect of pioglitazone on cancer risk would not be expected to be immediate, they excluded cases occurring in the first year after cohort entry. They identified 376 remaining bladder cancer cases.
3. **Sample controls from "risk sets" matched to each case.** The investigators sampled up to 20 controls for each case, matched on year of birth, year of cohort entry, sex, and duration of follow-up, who had not been diagnosed with bladder cancer up to the date of diagnosis of the case. The total number of matched controls was 6,699 (average number of controls per case = 17.8).[1]
4. **Define and measure predictors.** The primary predictor of interest was receipt of a prescription of either pioglitazone or rosiglitazone, another antidiabetes drug in the same class as pioglitazone. The prescription needed to be at least 1 year before the date of diagnosis of the case in the risk set. Four exposure levels were defined: prescription for pioglitazone only, rosiglitazone only, both, or neither.

The authors (appropriately) used conditional logistic regression to analyze the data; this accounts for the matched nature of the data and, because of the risk-set sampling, allows estimation of adjusted rate ratios (22). They found adjusted rate ratios of 1.83 (95% CI 1.10 to 3.05) for exclusive pioglitazone use, 1.14 (95% CI 0.78 to 1.68) for exclusive rosiglitazone use, and 0.78 (95% CI 0.18 to 3.29) for use of both. (The wide confidence interval on the last group reflects a much smaller sample size [$N = 2$ cases and 56 controls]). They also found evidence of dose-response relationship between pioglitazone use and bladder cancer: The adjusted rate ratio for cumulative dose of 28 grams or more was 2.54 (1.05–6.14), P for dose-response trend = 0.03.

[1]We will point out in Chapter 9 that the gain in power from sampling more than four controls per case is slight, but in this case the additional cost was low because electronic data were already available. Even with 20 controls per case the nested case–control approach is much more computationally efficient than a retrospective cohort study.

Strengths

Nested case–control and case–cohort studies are especially useful for costly measurements on serum and other specimens or images that have been archived at the beginning of the study and preserved for later analysis. Making expensive measurements on all the cases and a sample of the controls is much less costly than making the measurements on the entire cohort.

This design preserves all the advantages of cohort studies that result from collecting predictor variables before the outcomes have happened. In addition, it avoids the potential biases of conventional case–control studies that cannot make measurements on fatal cases and that draw cases and controls from different populations.

Weaknesses

These designs share certain disadvantages of other observational designs: the possibilities that observed associations are due to the effect of unmeasured or imprecisely measured confounding variables and that baseline measurements may be affected by silent preclinical disease.

Other Considerations

Nested case–control and case–cohort designs have been used less often than they should be. An investigator planning large prospective studies should consider preserving biologic samples (e.g., banks of frozen sera) or storing images or records that are expensive to analyze for subsequent nested case–control analyses. She should ensure that the conditions of storage will preserve substances of interest for many years. It may also be useful to collect new samples or information during the follow-up period, which can also be used in the case–control comparisons.

CASE-CROSSOVER STUDIES

The case-crossover design is a variant of the case–control design that is useful for studying the short-term effects of intermittent exposures. As with ordinary case–control studies, these retrospective studies begin with a group of cases: people who have had the outcome of interest. However, unlike traditional case–control studies, in which the exposures of the cases are compared with exposures of a group of controls, in case-crossover studies each case serves as her own control. Exposures of the cases at the time (or right before) the outcome occurred are compared with exposures of those same cases at one or more other points in time.

For example, McEvoy et al. (23) studied cases who were injured in car crashes and reported owning or using a mobile phone. Using phone company records, they compared mobile phone usage in the 10 minutes before the crash with usage when the subjects were driving at the same time of day 24 hours, 72 hours, and 7 days before the crash. They found that mobile phone usage was more likely in the 10 minutes before a crash than in the comparison time periods, with an odds ratio of about 4. The analysis of a case-crossover study is like that of a matched case–control study, only the control exposures are exposures of the case at different time periods, rather than exposures of the matched controls. This is illustrated in Appendix 8A, scenario number 4. Case-crossover designs have been used in large populations to study time-varying exposures like levels of air pollution; associations have been found with myocardial infarction (24, 25), emergency room visits for respiratory disease (26), and even infant mortality (27).

CHOOSING AMONG OBSERVATIONAL DESIGNS

The pros and cons of the main observational designs presented in the last two chapters are summarized in Table 8.2. We have already described these issues in detail and will make only one final point here. Among all these designs, none is best and none is worst; each has its place and purpose, depending on the research question and the circumstances.

TABLE 8.2 ADVANTAGES AND DISADVANTAGES OF THE MAJOR OBSERVATIONAL DESIGNS

DESIGN	ADVANTAGES	DISADVANTAGES*
	Cross-sectional	
	Relatively short duration A good first step for a cohort study or clinical trial Yields prevalence of multiple predictors and outcomes	Does not establish sequence of events Not feasible for rare predictors or rare outcomes Does not yield incidence
	Cohort Designs	
All	Establishes sequence of events Multiple predictors and outcomes Number of outcome events grows over time Yields incidence, relative risk, excess risk	Often requires large sample sizes Less feasible for rare outcomes
Prospective cohort	More control over subject selection and measurements Avoids bias in measuring predictors	Follow-up can be lengthy Often expensive
Retrospective cohort	Follow-up is in the past Relatively inexpensive	Less control over subject selection and measurements
Multiple cohort	Useful when distinct cohorts have different or rare exposures	Bias and confounding from sampling distinct populations
	Case–Control	
	Useful for rare outcomes Short duration, small sample size Relatively inexpensive	Bias and confounding from sampling two populations Differential measurement bias Limited to one outcome variable Sequence of events may be unclear Does not yield prevalence, incidence, or excess risk unless nested within a cohort
	Hybrid Designs	
Nested case–control	Advantages of a retrospective cohort design, and less costly if measurement of predictors is expensive	Measurements of risk factors subject to bias if not previously measured or based on banked specimens or images stored previously; usually requires a preexisting defined cohort
Incidence-density nested case–control	Allows investigators to analyze risk relationships taking into account changes over time in risk factor levels and loss to follow-up	Requires measurements of risk factor levels and incidence of cases over time during follow-up; usually requires a preexisting defined cohort
Nested case–cohort	Same as nested case–control and can use a single control group for multiple case–control studies with different outcomes	Same as nested case–control
Case-crossover	Cases serve as their own controls, reducing random error and confounding	Requires that the exposure have only immediate, short-term effects

*All these observational designs have the disadvantage (compared with randomized trials) of being susceptible to the influence of confounding variables—see Chapter 9.

SUMMARY

1. In a **case–control study**, the prevalence of a risk factor in a sample of subjects who have the outcome of interest (**the cases**) is compared with the prevalence in a sample that does not (**the controls**). This design, in which people with and without the disease are sampled separately, is relatively **inexpensive** and uniquely **efficient** for studying **rare diseases**.
2. One problem with case–control studies is their susceptibility to **sampling bias**. Four approaches to reducing sampling bias are (a) to sample controls and cases in the **same** (admittedly unrepresentative) **way**; (b) to do a **population-based** study; (c) to use **several** control groups, sampled in different ways; and (d) to **match** the cases and controls.
3. The other major problem with case–control studies is their retrospective design, which makes them susceptible to **measurement bias** affecting cases and controls differentially. Such bias can be reduced by using **measurements of the predictor made prior to the outcome** and by **blinding** the subjects and observers.
4. The best way to **avoid both sampling and measurement bias** is to design a ***nested* case–control study** in which random samples of cases and controls are drawn from a larger cohort study at its conclusion. In addition to controlling both of these biases, expensive baseline measurements on serum, images, and so on, can be made at the end of the study on a relatively **small number of study subjects**.
5. The ***incidence-density* case–control design** allows investigators to analyze risk relationships, taking into account **changes over time** in **risk factor** levels and in the **availability of follow-up**.
6. The **nested case–*cohort*** design uses a random sample of the entire cohort in place of the non-cases; this can serve as a control group for studying **more than one outcome**, and provides direct information on the overall prevalence of risk factors in the cohort.
7. **Case-crossover studies** are a variation on the matched case–control design in which observations at two or more points in time allow each case to serve as her own control.

APPENDIX 8A
Calculating Measures of Association

1. **Cross-sectional study.** Reijneveld (28) did a cross-sectional study of maternal smoking as a risk factor for infant colic. Partial results are shown below:

TABLE 8A.1

	OUTCOME VARIABLE		
PREDICTOR VARIABLE	INFANT COLIC	NO INFANT COLIC	TOTAL
Mother smokes 15 to 50 cigarettes/day	15 *(a)*	167 *(b)*	182 *(a + b)*
Mother does not smoke	111 *(c)*	2,477 *(d)*	2,588 *(c + d)*
Total	126 *(a + c)*	2,644 *(b + d)*	2,770 *(a + b + c + d)*

Prevalence of colic with smoking mothers = $a/(a + b)$ = 15/182 = 8.2%.
Prevalence of colic with nonsmoking mothers = $c/(c + d)$ = 111/2,588 = 4.3%.
Prevalence of colic overall = $(a + c)/(a + b + c + d)$ = 126/2,770 = 4.5%.

$$\text{Relative prevalence}^2 = \frac{8.2\%}{4.3\%} = 1.9$$

$$\text{Excess prevalence}^2 = 8.2\% - 4.3\% = 3.9\%$$

In other words, colic was almost twice (1.9 times) as common, and occurred almost 4% more often, among children of smoking mothers.

2. **Case–control study.** The research question for Example 8.1was whether there is an association between intramuscular vitamin K and risk of childhood leukemia. The findings were that 69/107 leukemia cases and 63/107 controls had received vitamin K. A 2 × 2 table of these findings is as follows:

TABLE 8A.2

	OUTCOME VARIABLE: DIAGNOSIS	
PREDICTOR VARIABLE: MEDICATION HISTORY	CHILDHOOD LEUKEMIA	CONTROL
Intramuscular vitamin K	69(*a*)	63(*b*)
No intramuscular vitamin K	38(*c*)	44(*d*)
Total	107	107

$$\text{Relative risk} \approx \textit{odds ratio} = \frac{ad}{bc} = \frac{69 \times 44}{63 \times 38} = 1.27$$

[2]Relative prevalence and excess prevalence are the cross-sectional analogs of relative risk and excess risk.

Because the disease (leukemia in this instance) is rare, the odds ratio provides a good estimate of the relative risk. Thus, leukemia was about 1.3 times more likely after receipt of vitamin K, but this was not statistically significant.[3]

3. **Matched case–control study.**
(To illustrate the similarity between analysis of a matched case–control study and a case-crossover study, we will use the same example for both.) The research question is whether mobile telephone use increases the risk of car crashes among mobile telephone owners. A traditional matched case–control study might consider self-reported frequency of using a mobile telephone while driving as the risk factor. Then the cases would be people who had been in crashes and they could be compared with controls who had not been in crashes, matched by age, sex, and mobile telephone prefix to the cases. The cases and controls would then be asked whether they ever use a mobile telephone while driving. (To simplify, for this example, we dichotomize the exposure and consider people as either "users" or "nonusers" of mobile telephones while driving.) We then classify each case/control pair according to whether both are users, neither is a user, or the case was a user but not the control, or the control was a user but not the case. If we had 300 pairs, the results might look like this:

TABLE 8A.3

	CASES (WITH CRASH INJURIES)		
MATCHED CONTROLS	USER	NONUSER	TOTAL
User	110	40	150
Nonuser	90	60	150
Total	200	100	300

Table 8A.3 shows that there were 90 pairs where the case ever used a mobile phone while driving, but not the matched control, and 40 pairs where the matched control but not the case was a "user." Note that this 2 × 2 table is different from the 2 × 2 table from the unmatched vitamin K study in question 2, in which each cell in the table is the number of people in that cell. In the 2 × 2 table for a *matched* case–control study the number in each cell is the number of *pairs* of subjects in that cell; the total *N* in Table 8A.3 is therefore 600 (300 cases and 300 controls). The odds ratio for such a table is simply the ratio of the two types of discordant pairs; in the Table 8.A.3 the OR = 90/40 = 2.25. This implies that users of mobile phones had more than double the odds of being in a crash.

4. **Case-crossover study.** Now consider the case-crossover study of the same question. Data from the study by McEvoy et al. are shown below.

TABLE 8A.4

	CRASH TIME PERIOD		
SEVEN DAYS BEFORE CRASH	DRIVER USING PHONE	NOT USING	TOTAL
Driver using phone	5	6	11
Not using	27	288	315
Total	32	294	326

[3]The authors actually did a multivariate, matched analysis, as was appropriate for the matched design, but in this case the simple, unmatched odds ratio was almost the same as the one reported in the study.

For the case-crossover study, each cell in the table is a number of subjects, not a number of pairs, but *each cell represents two time periods* for that one subject: the time period just before the crash and a comparison time period 7 days before. Therefore, the 5 in the upper left cell means there were 5 drivers involved in crashes who were using a mobile phone just before they crashed, and also using a mobile phone during the comparison period 7 days before, while the 27 just below the 5 indicates that there were 27 drivers involved in crashes who were using a phone just before crashing, but *not* using a phone during the comparison period 7 days before. Similarly, there were 6 drivers involved in crashes who were not using their phone at the time of the crash, but were using them in the comparison time period 7 days before. The odds ratio is the ratio of the numbers of discordant time periods, in this example $27/6 = 4.5$, meaning that driving during time periods of mobile phone use is associated with 4.5-fold higher odds of a crash than driving during time periods when not using a mobile phone.

APPENDIX 8B
Why the Odds Ratio Can Be Used as an Estimate for Relative Risk in a Case–Control Study

The data in a case–control study represent two samples: The cases are drawn from a population of people who have the disease and the controls from a population of people who do not have the disease. The predictor variable (risk factor) is measured, and the results can be summarized in a 2 × 2 table like the following one:

	Cases	Controls
Risk factor present	a	b
Risk factor absent	c	d

If this 2 × 2 table represented data from a cohort study, then the incidence of the disease in those with the risk factor would be $a/(a + b)$ and the relative risk would be simply $[a/(a + b)]/[c/(c + d)]$. However, it is not appropriate to compute either incidence or relative risk in this way in a case–control study because the two samples are not drawn from the population in the same proportions. Usually, there are roughly equal numbers of cases and controls in the study samples but many fewer cases than controls in the population. Instead, relative risk in a case–control study can be approximated by the odds ratio, computed as the cross-product of the 2 × 2 table, ad/cb.

This extremely useful fact is difficult to grasp intuitively but easy to demonstrate algebraically. Consider the situation for the full population, represented by a', b', c', and d'.

	Disease	No Disease
Risk factor present	a'	b'
Risk factor absent	c'	d'

Here it is appropriate to calculate the risk of disease among people with the risk factor as $a'/(a' + b')$, the risk among those without the risk factor as $c'/(c' + d')$, and the relative risk as $[a'/(a' + b')]/[c'/(c' + d')]$. We have already discussed the fact that $a'/(a' + b')$ is not equal to $a/(a + b)$. However, if the disease is relatively uncommon in both those with and without the risk factor (as most are), then a' is much smaller than b', and c' is much smaller than d'. This means that $a'/(a' + b')$ is closely approximated by a'/b' and that $c'/(c' + d')$ is closely approximated by c'/d'. Therefore, the relative risk of the population can be approximated as follows:

$$\frac{a'/(a' + b')}{c'/(c' + d')} \approx \frac{a'/b'}{c'/d'}$$

The latter term is the odds ratio of the population (literally, the ratio of the odds of disease in those with the risk factor, a'/b', to the odds of disease in those without the risk factor, c'/d'). This can be rearranged as the cross-product:

$$\left(\frac{a'}{c'}\right)\left(\frac{d'}{c'}\right) = \left(\frac{a'}{c'}\right)\left(\frac{d'}{b'}\right)$$

However, a'/c' in the population equals a/c in the sample if the cases are representative of all cases in the population (i.e., have the same prevalence of the risk factor). Similarly, b'/d' equals b/d if the controls are representative.

Therefore, the population parameters in this last term can be replaced by the sample parameters, and we are left with the fact that the odds ratio observed in the sample, ad/bc, is a close approximation of the relative risk in the population, $[a'/(a' + b')]/[c'/(c' + d')]$, provided that the disease is rare.

REFERENCES

1. Herbst AL, Ulfelder H, Poskanzer DC. Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. *N Engl J Med* 1971;284(15):878–881.
2. Beal SM, Finch CF. An overview of retrospective case–control studies investigating the relationship between prone sleeping position and SIDS. *J Paediatr Child Health* 1991;27(6):334–339.
3. Mitchell EA, Hutchison L, Stewart AW. The continuing decline in SIDS mortality. *Arch Dis Child* 2007;92(7):625–626.
4. Golding J, Greenwood R, Birmingham K, Mott M. Childhood cancer, intramuscular vitamin K, and pethidine given during labour. *BMJ* 1992;305(6849):341–346.
5. Golding J, Paterson M, Kinlen LJ. Factors associated with childhood cancer in a national cohort study. *Br J Cancer* 1990;62(2):304–308.
6. von Kries R, Gobel U, Hachmeister A, et al. Vitamin K and childhood cancer: a population based case–control study in Lower Saxony, Germany. *BMJ* 1996;313(7051):199–203.
7. Fear NT, Roman E, Ansell P, et al. Vitamin K and childhood cancer: a report from the United Kingdom Childhood Cancer Study. *Br J Cancer* 2003;89(7):1228–1231.
8. Roman E, Fear NT, Ansell P, et al. Vitamin K and childhood cancer: analysis of individual patient data from six case-control studies. *Br J Cancer* 2002;86(1):63–69.
9. Kochen M, McCurdy S. Circumcision and the risk of cancer of the penis. A life-table analysis. *Am J Dis Child* 1980;134(5):484–486.
10. O'Brien KL, Selanikio JD, Hecdivert C, et al. Epidemic of pediatric deaths from acute renal failure caused by diethylene glycol poisoning. Acute Renal Failure Investigation Team. *JAMA* 1998;279(15):1175–1180.
11. Fatal poisoning among young children from diethylene glycol-contaminated acetaminophen - Nigeria, 2008–2009. *MMWR Morb Mortal Wkly Rep* 2009;58(48):1345–1347.
12. Puumala SE, Spector LG, Robison LL, et al. Comparability and representativeness of control groups in a case–control study of infant leukemia: a report from the Children's Oncology Group. *Am J Epidemiol* 2009;170(3):379–387.
13. Voigt LF, Schwartz SM, Doody DR, et al. Feasibility of including cellular telephone numbers in random digit dialing for epidemiologic case–control studies. *Am J Epidemiol* 2011;173(1):118–126.
14. Hurwitz ES, Barrett MJ, Bregman D, et al. Public Health Service study of Reye's syndrome and medications. Report of the main study. *JAMA* 1987;257(14):1905–1911.
15. Cockburn M, Hamilton A, Mack T. Recall bias in self-reported melanoma risk factors. *Am J Epidemiol* 2001; 153(10):1021–1026.
16. Parr CL, Hjartaker A, Laake P, et al. Recall bias in melanoma risk factors and measurement error effects: a nested case-control study within the Norwegian Women and Cancer Study. *Am J Epidemiol* 2009;169(3):257–266.
17. Gefeller O. Invited commentary: Recall bias in melanoma—much ado about almost nothing? *Am J Epidemiol* 2009;169(3):267–270; discussion 71–72.
18. Cummings SR. Are patients with hip fractures more osteoporotic? Review of the evidence. *Am J Med* 1985;78(3):487–494.
19. Cauley JA, Lucas FL, Kuller LH, et al. Elevated serum estradiol and testosterone concentrations are associated with a high risk for breast cancer. Study of Osteoporotic Fractures Research Group. *Ann Intern Med* 1999; 130(4 Pt 1):270–277.
20. Azoulay L, Yin H, Filion KB, et al. The use of pioglitazone and the risk of bladder cancer in people with type 2 diabetes: nested case–control study. *BMJ* 2012;344:e3645.
21. Hassey A, Gerrett D, Wilson A. A survey of validity and utility of electronic patient records in a general practice. *BMJ* 2001;322(7299):1401–1405.
22. Essebag V, Platt RW, Abrahamowicz M, et al. Comparison of nested case-control and survival analysis methodologies for analysis of time-dependent exposure. *BMC Med Res Methodol* 2005;5(1):5.
23. McEvoy SP, Stevenson MR, McCartt AT, et al. Role of mobile phones in motor vehicle crashes resulting in hospital attendance: a case-crossover study. *BMJ* 2005;331(7514):428.
24. Bhaskaran K, Hajat S, Armstrong B, et al. The effects of hourly differences in air pollution on the risk of myocardial infarction: case crossover analysis of the MINAP database. *BMJ* 2011;343:d5531.

25. Nuvolone D, Balzi D, Chini M, et al. Short-term association between ambient air pollution and risk of hospitalization for acute myocardial infarction: results of the cardiovascular risk and air pollution in Tuscany (RISCAT) study. *Am J Epidemiol* 2011;174(1):63–71.
26. Tramuto F, Cusimano R, Cerame G, et al. Urban air pollution and emergency room admissions for respiratory symptoms: a case-crossover study in Palermo, Italy. *Environ Health* 2011;10:31.
27. Scheers H, Mwalili SM, Faes C, et al. Does air pollution trigger infant mortality in Western Europe? A case-crossover study. *Environ Health Perspect* 2011;119(7):1017–1022.
28. Reijneveld SA, Brugman E, Hirasing RA. Infantile colic: maternal smoking as potential risk factor. *Arch Dis Child* 2000;83(4):302–303.

CHAPTER 9

Enhancing Causal Inference in Observational Studies

Thomas B. Newman, Warren S. Browner, and Stephen B. Hulley

Most observational studies are designed to suggest that a predictor may be a cause of an outcome, for example, that eating broccoli may reduce the risk of colon cancer. (Exceptions are studies of diagnostic and prognostic tests, discussed in Chapter 12.) Causal associations between a predictor and an outcome are important because they can provide insights into the underlying biology of a disease, identify ways to reduce or prevent its occurrence, and even suggest potential treatments.

However, not every association that is found in an observational study represents cause–effect. Indeed, there are four other general explanations for an association between a predictor and an outcome in an observational study (Table 9.1). Two of these, **chance** and **bias**, create spurious associations between the predictor and the outcome in the study sample that do not exist in the population. Two others, **effect–cause** and **confounding**, create real associations in the population, but these associations are not causal in the direction of interest. Establishing that cause–effect is the most likely explanation for an association requires demonstrating that these other explanations are unlikely.

We typically quantify the causal effect of a predictor variable on an outcome using a measure of association, such as a risk ratio or odds ratio. For example, suppose that a study reveals that coffee drinking has a risk ratio of 2.0 for myocardial infarction (MI). One possibility—presumably the one that the investigator found most interesting—is that drinking coffee doubles the risk of MI. Before reaching this conclusion, however, the four rival explanations must be considered and dismissed.

With chance and bias, coffee drinking was associated with a doubled risk of MI in the study, but that association is not actually present in the population. Thus, chance and bias are explanations for spurious (i.e., not real) associations in a study.

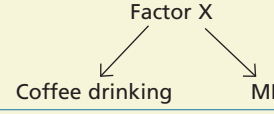
The other two alternatives—effect–cause and confounding—are true biological phenomena, which means that coffee drinkers in the population really do have twice the risk of MI. However, that increased risk is not due to a cause–effect relationship. In one situation, the association is due to effect–cause: having an MI causes people to drink more coffee. (This is just cause and effect in reverse.) The final possibility, confounding, occurs when a third factor, such as personality type, causes both coffee drinking and MI.

In the remainder of the chapter, we will discuss strategies for estimating and minimizing the likelihood of these four alternative explanations for finding an association in an observational study. These strategies can be used while designing a study or when analyzing its results. While this book emphasizes research design, understanding the analytic options can influence the choice of design, so both topics will be considered in this chapter.

SPURIOUS ASSOCIATIONS DUE TO CHANCE

Suppose that in reality there is no association between coffee drinking and MI among members of a population, 45% of whom drink coffee. If we were to select 20 cases with MI and 20 controls, we would expect that about 9 people in each group (45% of 20) would drink coffee.

TABLE 9.1 THE FIVE EXPLANATIONS FOR AN OBSERVED DOUBLING OF THE RISK OF MI ASSOCIATED WITH COFFEE DRINKING

EXPLANATION	TYPE OF ASSOCIATION	WHAT'S REALLY GOING ON IN THE POPULATION?	CAUSAL MODEL
1. Chance (random error)	Spurious	Coffee drinking and MI are not related.	—
2. Bias (systematic error)	Spurious	Coffee drinking and MI are not related.	—
3. Effect–cause	Real	MI is a cause of coffee drinking.	MI → Coffee drinking
4. Confounding	Real	A third factor causes both coffee drinking and MI.	
5. Cause–effect	Real	Coffee drinking is a cause of MI.	Coffee drinking → MI

However, *by chance alone,* we might enroll 12 coffee drinkers among the 20 MI cases, but only 6 in the 20 controls. If that happened, we would observe a spurious association between coffee consumption and MI in our study.

Chance is sometimes called **random error**, because it has no underlying explanation. When an association due to random error is statistically significant, it's known as a **type I error** (Chapter 5).

Strategies for reducing random error are available in both the design and analysis phases of research (Table 9.2). *Design strategies*, such as increasing the **precision of measurements** and increasing the **sample size**, are discussed in Chapters 4 and 6, respectively. The *analysis strategy* of calculating ***P* values** and **confidence intervals** helps the investigator quantify the magnitude of the observed association in comparison with what might have occurred by chance

TABLE 9.2 STRENGTHENING THE INFERENCE THAT AN ASSOCIATION IS DUE TO CAUSE–EFFECT BY REDUCING AND EVALUATING THE LIKELIHOOD OF SPURIOUS ASSOCIATIONS

TYPE OF SPURIOUS ASSOCIATION	DESIGN PHASE (HOW TO PREVENT THE RIVAL EXPLANATION)	ANALYSIS PHASE (HOW TO EVALUATE THE RIVAL EXPLANATION)
Chance (due to random error)	Increase sample size and other strategies to increase precision (Chapters 4 and 6)	Calculate *P* values and confidence intervals and interpret them in the context of prior evidence (Chapter 5)
Bias (due to systematic error)	Carefully consider the potential consequences of each difference between the research question and the study plan (Figure 9.1); alter the study plan if necessary	Check consistency with other studies (especially those using different designs)
	Collect additional data that will allow assessment of the extent of possible biases	Analyze additional data to see if potential biases have actually occurred
	Do not use variables affected by the predictor of interest as inclusion criteria or matching variables	Do not control for variables affected by your predictor variable

alone. For example, a *P* value of 0.10 indicates that chance alone could cause a difference at least as large as the investigators observed about 10% of the time. Even more useful than *P* values, confidence intervals show the possible values for statistics describing an association that fall within the range of random error estimated in the study.

■ SPURIOUS ASSOCIATIONS DUE TO BIAS

Many kinds of bias—systematic error—have been identified, and dealing with some of them is a major topic of this book. Along with the specific strategies described in Chapters 3, 4, 7 and 8, we now add a general approach to reducing the likelihood of bias.

Minimizing Bias

As was discussed in Chapter 1, there are almost always differences between the original research question and the one that is actually answered by the study. Those differences reflect the compromises that were made for the study to be feasible, as well as mistakes in the design or execution of the study. Bias occurs when those differences cause the answer provided by the study to differ from the right answer to the research question. Strategies for minimizing bias are available in both the design and analysis phases of research (Table 9.2).

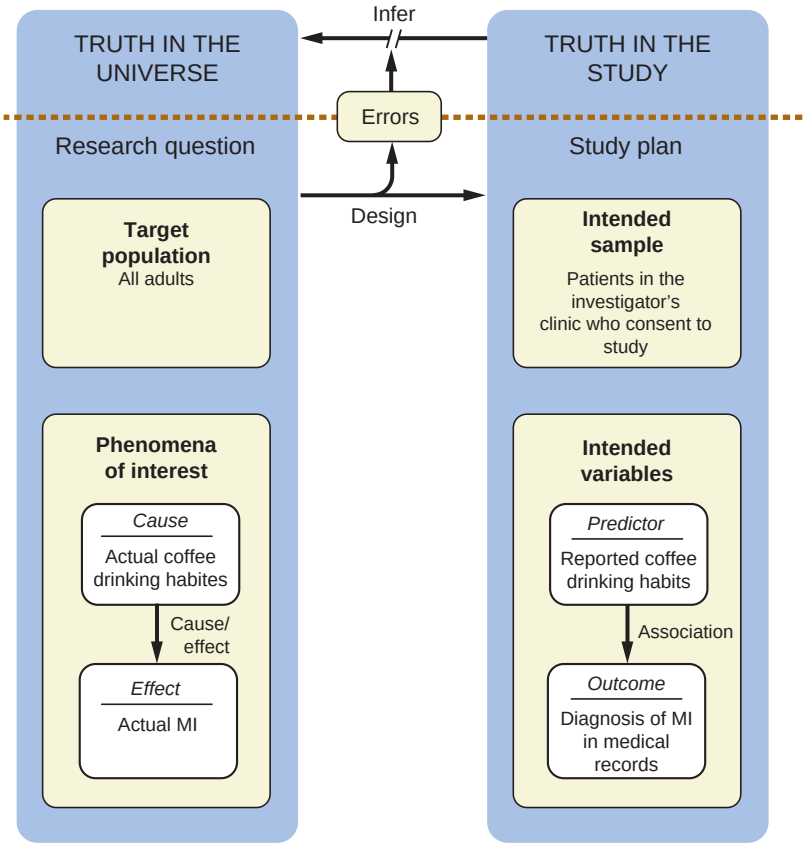
- **Design phase.** Begin by writing the research question next to the study plan, as in Figure 9.1. Then think through the following three concerns as they pertain to the research question:
  1. Do the **samples** of study subjects (e.g., cases and controls, or exposed and unexposed subjects) represent the population(s) of interest?
  2. Do the **measurements** of the **predictor variables** represent the predictors of interest?
  3. Do the **measurements** of the **outcome variables** represent the outcomes of interest?

  For each question answered "No" or "Maybe not," consider whether the bias applies similarly to one or both groups studied (e.g., cases and controls, or exposed and unexposed subjects) and whether the bias is likely to be large enough to affect the answer to the research question.

  To illustrate this with our coffee and MI example, consider a case–control study in which the control subjects are sampled from patients hospitalized for diseases other than MI. If many of these patients have chronic illnesses that led them to reduce their coffee intake, the sample of controls will not represent the target population from which the MI cases arose: There will be a shortage of coffee drinkers. And if esophageal spasm, which can be exacerbated by coffee, is misdiagnosed as MI, a spurious association between coffee and MI could be found because the measured outcome (diagnosis of MI) did not accurately represent the outcome of interest (actual MI).

  The next step is to think about possible strategies for preventing each potential bias, such as selecting more than one control group in a case–control study (Chapter 8) or the strategies for reducing measurement bias described in Chapter 4. In each case, judgments are required about the likelihood of bias and how easily it could be prevented with changes in the study plan. If the bias is easily preventable, revise the study plan and ask the three questions again. If the bias is not easily preventable, decide whether the study is still worth doing by judging the likelihood of the potential bias and the degree to which it will distort the association you are trying to estimate.

  Potential biases may either be unavoidable or costly to prevent, or it may be uncertain to what extent they will be a problem. In either case, the investigator should consider designing the study to collect additional data that will allow an assessment of the seriousness of the biases. For example, if the investigator is concerned that the cases in a study of pancreatic cancer may over-report recent exposures to toxic chemicals (perhaps because these individuals are searching desperately for an explanation for why they have pancreatic cancer),



■ **FIGURE 9.1** Minimizing bias by carefully considering differences between the research question and the study plan.

they could also be asked about exposures (such as coffee drinking!) that previous studies have shown have no effect on the risk of pancreatic cancer. If the investigator is concerned that a questionnaire does not accurately capture coffee drinking (perhaps because of poorly worded questions), she could assign a blinded interviewer to question a subset of the cases and controls to determine the agreement with their questionnaire responses. Similarly, if she is concerned that rather than causing MI, coffee increases survival among MI patients (which could lead to coffee drinkers being over-represented in a sample of MI survivors), the investigator could identify MI patients who died and interview their surviving spouses about their previous coffee-drinking habits.

- **Analysis phase.** Once the data have been collected, the goal shifts from minimizing bias to assessing its likely severity. The first step is to analyze data that have been collected for that purpose. For example, an investigator anticipating imperfect memory of coffee-drinking habits may have included questions about how sure the cases and controls are of their answers. The association between coffee drinking and MI could then be examined after stratifying on certainty about coffee intake, to see whether the association is stronger among those more certain of their exposure history.

The investigator can also look at the results of other studies. If the conclusions are consistent, the association is less likely to be due to bias. This is especially true if the other studies have used different designs and are therefore unlikely to share the same biases. However, in many situations the potential biases turn out not to be a major problem. The decision on how

vigorously to pursue additional information and how best to discuss these issues in reporting the study are matters of judgment for which it is helpful to seek advice from colleagues.

REAL ASSOCIATIONS OTHER THAN CAUSE–EFFECT

In addition to chance and bias, the two types of associations that are real but do not represent cause–effect must be considered (Table 9.3).

Effect–Cause

One possibility is that the cart has come before the horse—the outcome has caused the predictor. Effect–cause is often a problem in cross-sectional and case–control studies: Does a sedentary lifestyle cause obesity, or vice versa? Effect–cause can also be a problem in case-crossover studies. For example, in the study of mobile phone use and motor vehicle accidents described in Chapter 8 (1), a car crash could cause the driver to make a mobile phone call reporting the crash, rather than the crash having been caused by an inattentive driver. To address this possibility, the investigators asked drivers about phone use before and after the crash, and verified the responses using phone records.

Effect–cause is less commonly a problem in cohort studies of disease causation because risk factor measurements can be made among subjects who do not yet have the disease. Even in cohort studies, however, effect–cause is possible if the disease has a long latent period and those with subclinical disease cannot be identified at baseline. For example, Type II diabetes is associated with subsequent risk of pancreatic cancer. Some of this association may well be effect–cause, because pancreatic cancer could affect the pancreatic islet cells that secrete insulin, thus causing diabetes. Consistent with effect–cause, the risk of pancreatic cancer is highest just after diabetes is diagnosed (2). The association diminishes with the duration of diabetes, but some association persists even 4 years or more after the onset of diabetes (2–4) suggesting that at least some of the relationship may be cause–effect.

This example illustrates a general approach to ruling out effect–cause: looking for a diminution in the association with increasing time between the presumed cause and its effect. A second approach is to assess the biologic plausibility of effect–cause versus cause–effect. In this example effect-cause was plausible because pancreatic cancer could damage the pancreas, but the observation that having diabetes for more than 10 years is associated with an increased risk of a variety of other cancers as well as pancreatic cancer (4) increases the biologic plausibility of diabetes causing pancreatic cancer, rather than being only one of its effects.

TABLE 9.3 STRENGTHENING THE INFERENCE THAT AN ASSOCIATION HAS A CAUSE–EFFECT BASIS: RULING OUT OTHER REAL ASSOCIATIONS

TYPE OF REAL ASSOCIATION	DESIGN PHASE (HOW TO PREVENT THE RIVAL EXPLANATION)	ANALYSIS PHASE (HOW TO EVALUATE THE RIVAL EXPLANATION)
Effect–cause (the outcome is actually the cause of the predictor)	Do a longitudinal study to discover which came first Obtain data on the historic sequence of the variables (Ultimate solution: do a randomized trial)	Consider biologic plausibility Compare the strength of the association immediately after the exposure to the predictor with the strength later on Consider findings of other studies with different designs
Confounding (another variable causes both the predictor and outcome)	See Table 9.4	See Table 9.5

Confounding

The other rival explanation in Table 9.3 is confounding, which occurs when a third factor is a real cause of the outcome and the predictor of interest is associated with, but not a cause of, this third factor. For example, if certain personality traits cause people to drink more coffee and also to be at higher risk of MI, these personality traits will confound the association between coffee and MI. If this is the actual explanation, then the association between coffee and MI does not represent cause–effect, although it is perfectly real: Coffee drinking is an innocent bystander in terms of causation.

In order to be a confounder, a variable must be associated with the predictor of interest and also be a cause of the outcome. Confounding can be even more complicated, and sometimes, yet another factor is involved. For example, work environment could cause people to drink coffee and to smoke cigarettes, which is a risk factor for MI. Appendix 9A gives a numeric example of how differences in cigarette smoking could lead to an apparent association between coffee drinking and MI.

What if coffee drinking caused smoking and smoking caused MI? In that case, smoking is called a **mediator** of the (causal) association between coffee drinking and MI, not a confounder. In general, it is best to avoid controlling for factors that lie along the causal path between a predictor and an outcome.

Aside from bias, confounding is often the only likely alternative explanation to cause–effect and the most important one to try to rule out. It is also the most challenging; much of the rest of this chapter is devoted to strategies for coping with confounders. It is worth noting, however, that all of these strategies involve judgments, and that no amount of epidemiologic or statistical sophistication can substitute for understanding the underlying biology.

■ COPING WITH CONFOUNDERS IN THE DESIGN PHASE

Most strategies for coping with confounding variables require that an investigator measure them, so it is helpful to begin by listing the variables (like age and sex) that may be associated with the predictor variable and also cause the outcome. The investigator must then choose among design and analysis strategies for controlling the influence of these potential confounding variables.

The first two design phase strategies (Table 9.4), **specification** and **matching**, involve changes in the sampling scheme. Cases and controls (in a case–control study) or exposed and unexposed subjects (in a cohort study) can be sampled in such a way that they have comparable values of the confounding variable. This removes the confounder as an explanation for any association that is observed between predictor and outcome. A third design phase strategy, using **opportunistic study designs**, is only applicable to selected research questions for which the right conditions exist. However, when applicable, these designs can resemble randomized trials in their ability to reduce or eliminate confounding not only by measured variables, but by unmeasured variables as well.

Specification

The simplest strategy is to design inclusion criteria that **specify** a value of the potential confounding variable and exclude everyone with a different value. For example, the investigator studying coffee and MI could specify that only nonsmokers be included in the study. If an association were then observed between coffee and MI, it obviously could not be due to smoking.

Specification is an effective strategy, but, as with all restrictions in the sampling scheme, it has disadvantages. First, even if coffee does not cause MIs in nonsmokers, it may cause them in smokers. This phenomenon—an effect of coffee on MI that is different in smokers from that in nonsmokers—is called **effect modification** (also known as an **interaction**); see Appendix 9A.Thus, specification limits the generalizability of information available from a study, in this instance

TABLE 9.4 DESIGN PHASE STRATEGIES FOR COPING WITH CONFOUNDERS

STRATEGY	ADVANTAGES	DISADVANTAGES
Specification	• Easily understood • Focuses the sample of subjects for the research question at hand	• Limits generalizability and sample size
Matching	• Can eliminate influence of strong constitutional confounders like age and sex • Can eliminate the influence of confounders that are difficult to measure • Can increase power by balancing the number of cases and controls in each stratum • May be a sampling convenience, making it easier to select the controls in a case–control study	• May be time-consuming and expensive; may be less efficient than increasing the number of subjects • Decision to match must be made at outset of study and has an irreversible effect on analysis • Requires early decision about which variables are predictors and which are confounders • Eliminates the option of studying matched variables as predictors or as intervening variables • Requires matched analysis • Creates the danger of overmatching (i.e., matching on a factor that is not a confounder, thereby reducing power) • Only feasible for case–control and multiple-cohort studies
"Opportunistic" study designs	• Can provide great strength of causal inference • May be a lower cost and elegant alternative to a randomized trial	• Only possible in select circumstances where predictor variable is randomly or virtually randomly assigned, or instrumental variable exists

compromising our ability to generalize to smokers. A second disadvantage is that if smoking is highly prevalent among the patients available for the study, the investigator may not be able to recruit a large enough sample of nonsmokers. These problems can become serious if specification is used to control too many confounders or to control them too narrowly. Sample size and generalizability would be major problems if a study were restricted to lower-income, nonsmoking, 70- to 74-year-old men.

Matching

In a case–control study, **matching** can be used to prevent confounding by selecting cases and controls who have the same (matching) values of the confounding variable(s). Matching and specification both prevent confounding by allowing comparison only of cases and controls who share similar levels of the confounder. Matching differs from specification, however, in preserving generalizability, because subjects at all levels of the confounder can be studied.

Matching is usually done individually (**pair-wise matching**). To control for smoking in a study of coffee drinking as a predictor of MI, for example, each case (a subject with an MI) would be individually matched to one or more controls who smoked roughly the same amount as the case (e.g., 10 to 20 cigarettes/day). The coffee drinking of each case would then be compared with the coffee drinking of the matched control(s).

An alternative approach to pair-wise matching is to match in groups (**frequency matching**). For each level of smoking, the cases with that amount of smoking are counted, and an appropriate number of controls with the same level of smoking are selected. If the study called for two controls per case and there were 20 cases who smoked 10 to 20 cigarettes/day, the investigators would select 40 controls who smoked this amount, matched as a group to the 20 cases.

Matching is most commonly used in **case–control studies**, but it can also be used with **multiple-cohort designs**. For example, to investigate the effects of service in the 1990–1991 Gulf War on subsequent fertility in male veterans, Maconochie et al. (5) compared men deployed to the Gulf region during the war with men who were not deployed, but were frequency-matched by service, age, fitness to be deployed, and so on. They found a slightly higher risk of reported infertility (OR ~1.5) and a longer time to conception in the Gulf War veterans.

Advantages to Matching (Table 9.4)

- Matching is an effective way to **prevent confounding by constitutional factors** like age, sex, and race that are strong determinants of outcome, not susceptible to intervention, and unlikely to be intermediaries on a causal path.
- Matching can be used to **control confounders that cannot be measured** and controlled in any other way. For example, matching siblings (or, better yet, twins) with one another can control for a whole range of genetic and familial factors that would be impossible to measure. Matching for clinical center in a multicenter study can control for unspecified differences among the populations or staff at geographically dispersed centers.
- Matching may **increase the precision** of comparisons between groups (and therefore the power of the study to find a real association) by balancing the *number* of cases and controls at each level of the confounder. This may be important if the available number of cases is limited or if the cost of studying the subjects is high. However, the effect of matching on precision is modest and not always favorable (see "overmatching," p. 125). In general, the desire to enhance precision is a less important reason to match than the need to control confounding.
- Finally, matching may be used primarily as a **sampling convenience**, to narrow down an otherwise impossibly large number of potential controls. For example, in a study of marijuana use as a risk factor for testicular germ cell tumors, investigators asked cases (men with testicular tumors) to suggest friends of similar age without tumors to be in the control group (6). This convenience, however, also runs the risk of overmatching.

Disadvantages to Matching (Table 9.4)

- Matching requires additional **time and expense** to identify a match for each subject. In case–control studies, for example, the more matching criteria there are, the larger the pool of controls that must be searched to match each case. The possible increase in statistical power from matching must therefore be weighed against the increase in power that might be obtained by enrolling more cases.
- When matching is used as a sampling strategy, the decision to match must be made at the beginning of the study. It is therefore **irreversible**. This precludes further analysis of the effect of the matched variables on the outcome. It also can create a serious error if the matching variable is not a constitutional variable like age or sex, but an intermediary in the causal path between the predictor and outcome. For example, if an investigator wishing to investigate the effects of alcohol intake on risk of MI matched on serum high-density lipoprotein (HDL) levels, she would miss any beneficial effects of alcohol that are mediated through an increase in HDL. Although the same error can occur with the analysis phase strategies, matching builds the error into the study in a way that cannot be undone; with the analysis phase strategies the error can be avoided by altering the analysis.
- Correct analysis of pair-matched data requires special analytic techniques (**matched analyses**) that compare each subject only with her match, and not with other subjects who have differing levels of confounders. This means cases for whom a match cannot be found cannot be included. In the study of marijuana use and germ cell tumors, 39 of the 187 cases did not provide a friend control (6). The authors had to exclude these 39 cases from the

matched analysis. The use of unmatched analytic techniques on matched data can lead to incorrect results (generally biased toward no effect) because the assumption that the groups are sampled independently is violated.

- A final disadvantage of matching is the possibility of **overmatching**, which occurs when the matching variable is associated with the predictor but turns out not to be a confounder because it is not associated with the outcome. Overmatching can reduce the power of a case–control study, because the matched analysis discards matched case–control sets with the same level of exposure (Appendix 8A.3). In the marijuana and germ cell tumor study, for example, use of friend controls may have reduced the power by increasing the concordance in exposures between cases and their matched controls: Friends might tend to have similar patterns of marijuana use.

Opportunistic Studies

Occasionally, there are opportunities to control for confounding variables in the design phase, even without measuring them; we call these "opportunistic" designs because they utilize unusual opportunities for controlling confounding. One example, useful when studying the immediate effects of short-term exposures, is the **case-crossover** study (Chapter 8)—all potential confounding variables that are constant over time (e.g., sex, race, social class, genetic factors) are controlled because each subject is compared only with herself in a different time period.

Another opportunistic design involves a **natural experiment**, in which subjects are either exposed or not exposed to a particular risk factor through a process that, in effect, acts randomly (7). For example, Lofgren et al. (8) studied the effects of discontinuity of in-hospital care by taking advantage of the fact that patients admitted after 5:00 PM to their institution were alternately assigned to senior residents who either maintained care of the patients or transferred them to another team the following morning. They found that patients whose care was transferred had 38% more laboratory tests ordered ($P = 0.01$) and 2-day longer median length of stay ($P = 0.06$) than those kept on the same team. Similarly, Bell and Redelmeier (9) studied the effects of nursing staffing by comparing outcomes for patients with selected diagnoses who were admitted on weekends to those admitted on weekdays. They found higher mortality from all three conditions they predicted would be affected by reduced weekend staffing ratios, but no increase in mortality for patients hospitalized for other conditions.

As genetic differences in susceptibility to an exposure are elucidated, a strategy called **Mendelian randomization** (10) becomes an option. This strategy works because, for common genetic polymorphisms, the allele a person receives is determined at random within families, and not linked to most confounding variables. For example, some farmers who dip sheep in insecticides (to kill ticks, lice, etc.) have health complaints, such as headache and fatigue, that might be due to that occupational insecticide exposure. Investigators (11) took advantage of a polymorphism in the paraoxonase-1 gene, which leads to enzymes with differing ability to hydrolyze the organophosphate insecticide (diazinonoxon) used in sheep dip. They found that exposed farmers with health complaints were more likely to have alleles associated with reduced paraoxonase-1 activity than similarly exposed but asymptomatic farmers. This finding provided strong evidence of a causal relationship between exposure to sheep dip and health problems.

Natural experiments and Mendelian randomization are examples of a more general approach to enhancing causal inference in observational studies, the use of **instrumental variables**. These are variables associated with the predictor of interest, but not independently associated with the outcome. Whether someone is admitted on a weekend, for example, is associated with staffing levels, but was thought not to be otherwise associated with mortality risk (for the diagnoses studied), so admission on a weekend can be considered an instrumental variable. Similarly, activity of the paraoxonase-1 enzyme is associated with possible toxicity due to dipping sheep, but not otherwise associated with ill health. Other examples of instrumental variables are draft

lottery numbers to investigate delayed effects on mortality of military service during the Vietnam War era (12); and whether long-term survival for early-stage kidney cancer depends on how far someone lives from a urologist who does partial nephrectomies versus one who only does radical nephrectomies (13).

COPING WITH CONFOUNDERS IN THE ANALYSIS PHASE

The Design phase strategies specification and matching require deciding at the outset of the study which variables are confounders, and the investigators cannot subsequently estimate the effects of those confounders on an outcome. By contrast, analysis phase strategies keep the investigator's options open, so that she can change her mind about which variables to control for at the time of analysis.

Sometimes there are several predictor variables, each of which may act as a confounder to the others. For example, although coffee drinking, smoking, male sex, and personality type are associated with MI, they are also associated with each other. The goal is to determine which of these predictor variables are independently associated with MI and which are associated with MI only because they are associated with other (causal) risk factors. In this section, we discuss analytic methods for assessing the **independent** contribution of predictor variables in observational studies. These methods are summarized in Table 9.5.[1]

Stratification

Like specification and matching, **stratification** ensures that only cases and controls (or exposed and unexposed subjects) with similar levels of a potential confounding variable are compared. It involves segregating the subjects into strata (**subgroups**) according to the level of a potential confounder and then examining the relation between the predictor and outcome separately in each stratum. Stratification is illustrated in Appendix 9A. By considering smokers and nonsmokers separately ("stratifying on smoking"), the confounding effects of smoking can be removed.

Appendix 9A also illustrates **effect modification**, in which stratification reveals that the association between predictor and outcome varies with (is modified by) the level of a third factor. Effect modification introduces additional complexity, because a single measure of association no longer can summarize the relationship between predictor and outcome. By chance alone, the estimates of association in different strata will rarely be precisely the same, and it is only when the estimates vary markedly that the findings suggest effect modification. Clinically significant effect modification is uncommon, and before concluding that it is present it is necessary to assess its statistical significance, and, especially if many subgroups have been examined (increasing the likelihood of at least one being statistically significant due to chance), to see if it can be replicated in another population. Biologic plausibility, or the lack thereof, may also contribute to the interpretation. The issue of effect modification also arises for subgroup analyses of clinical trials (Chapter 11), and for meta-analyses when homogeneity (similarity) of studies is being considered (Chapter 13).

Stratification has the advantage of **flexibility**: by performing several stratified analyses, the investigator can decide which variables appear to be confounders and ignore the remainder. This can be done by combining knowledge about the likely directions of causal relationships with analyses determining whether the results of stratified analyses substantially differ from those of unstratified analyses (see Appendix 9A). Stratification also has the advantage of being reversible: No choices need be made at the beginning of the study that might later be regretted.

[1]Similar questions arise in studies of diagnostic tests (Chapter 12), but in those situations the goal is not to determine a causal effect, but to determine whether the test being studied adds substantial predictive power to information already available at the time it was done.

TABLE 9.5 ANALYSIS PHASE STRATEGIES FOR COPING WITH CONFOUNDERS

STRATEGY	ADVANTAGES	DISADVANTAGES

Stratification	• Easily understood
	• Flexible and reversible; can choose which variables to stratify upon after data collection
		• Number of strata limited by sample size needed for each stratum
		  • Few covariables can be considered
		  • Few strata per covariable leads to incomplete control of confounding
		• Relevant covariables must have been measured

Statistical adjustment	• Multiple confounders can be controlled simultaneously
	• Information in continuous variables can be fully used
	• Flexible and reversible
		• Model may not fit:
		  • Incomplete control of confounding (if model does not fit confounder-outcome relationship)
		  • Inaccurate estimates of strength of effect (if model does not fit predictor–outcome relationship)
		• Results may be hard to understand. (Many people do not readily comprehend the meaning of a regression coefficient.)
		• Relevant covariables must have been measured

Propensity scores	• Multiple confounders can be controlled simultaneously • Information in continuous variables can be fully used • Enhances power to control for confounding when more people receive the treatment than get the outcome • If a stratified or matched analysis is used, does not require model assumptions • Flexible and reversible • Lack of overlap of propensity scores can highlight subgroups in whom control of confounding is difficult or impossible	• Results may be hard to understand • Relevant covariables must have been measured • Can only be done for exposed and unexposed subjects with overlapping propensity scores, reducing sample size

The principal disadvantage of stratified analysis is the limited number of variables that can be controlled simultaneously. For example, possible confounders in the coffee and MI study might include age, personality type, systolic blood pressure, serum cholesterol, and cigarette smoking. To stratify on these five variables with only three strata for each would require $3^5 = 243$ strata! With this many strata there will be some strata with no cases or no controls, and these strata cannot be used.

To maintain a sufficient number of subjects in each stratum, a variable is often divided into broader strata. When the strata are too broad, however, the confounder may not be adequately controlled. For example, if the preceding study stratified age using only two strata (e.g., <50 and ≥50 years), some residual confounding would still be possible if within each age stratum the subjects drinking the most coffee were older and therefore at higher risk of MI.

Adjustment

Several statistical techniques are available to adjust for confounders. These techniques *model* the nature of the associations among the variables to isolate the effects of predictor variables and confounders. For example, a study of the effect of lead levels on the intelligence quotient (IQ)

in children might examine parental education as a potential confounder. Statistical adjustment might model the relation between parents' years of schooling and the child's IQ as a straight line, in which each year of parent education is associated with a fixed increase in child IQ. The IQs of children with different lead levels could then be adjusted to remove the effect of parental education using the approach described in Appendix 9B.

Often, an investigator wants to adjust simultaneously for several potential confounders—such as age, sex, race, and education. This requires using multivariate adjustment techniques, such as multivariable linear or logistic regression, or Cox proportional hazards analysis. These techniques have another advantage: They enable the use of all the information in continuous variables. It is easy, for example, to adjust for a parent's education level in 1-year intervals, rather than stratifying into just a few categories. In addition, **interaction terms** can be used to model effect modification among the variables.

There are, however, disadvantages of multivariate adjustment. Most important, the model may not fit. Computerized statistical packages have made these models so accessible that the investigator may not stop to consider whether their use is appropriate for the predictor and outcome variables in the study.[3] Taking the example in Appendix 9B, the investigator should examine whether the relation between the parents' years of schooling and the child's IQ is actually linear. If the pattern is very different (e.g., the slope of the line becomes steeper with increasing education) then attempts to adjust IQ for parental education using a linear model will be imperfect and the estimate of the independent effect of lead will be incorrect.

Second, the resulting statistics are often difficult to understand. This is particularly a problem if transformations of variables (e.g., parental education squared) or interaction terms are used. Investigators should spend the necessary time with a statistician (or take the necessary courses) to make sure they can explain the meaning of coefficients or other highly derived statistics they plan to report. As a safety precaution, it is a good idea always to start with simple, stratified analyses, and to seek help understanding what is going on if more complicated analyses yield substantially different results.

Propensity Scores

Propensity scores can be particularly useful for observational studies of treatment efficacy to control **confounding by indication**—the problem that patients for whom a treatment is indicated (and prescribed) are often at higher risk, or otherwise different, from those who do not get the treatment. Recall that in order to be a confounder, a variable must be associated with both the predictor and outcome. Instead of adjusting for all factors that predict *outcome*, use of propensity scores involves creating a multivariate model to predict receipt of the *treatment*. Each subject can then be assigned a predicted probability of treatment—a "propensity score." This single score can be used as the only confounding variable in a stratified or mutlivariable analysis.

Alternatively, subjects who did and did not receive the treatment can be matched by propensity score, and outcomes compared between matched pairs. Unlike use of matching as a design-phase (sampling) strategy, propensity matching resembles other analysis phase strategies in being reversible. However, matched propensity analyses fail for subjects who cannot be matched because their propensity scores are close to 0 or 1. While this reduces sample size, it may be an advantage because in these unmatchable subjects the propensity score analysis has identified a lack of comparability between groups and inability to control for confounding that might not have been apparent with other methods of multivariable analysis.

[3]One of our biostatistician colleagues has quipped that trying to design a user-friendly, intuitive statistical software package is like trying to design a car so that a child can reach the pedals.

EXAMPLE 9.1 Propensity Analysis

Gum et al. (14) prospectively studied 6,174 consecutive adults undergoing stress echocardiography, 2,310 of whom (37%) were taking aspirin and 276 of whom died in the 3.1-year follow-up period. In unadjusted analyses, aspirin use was not associated with mortality (4.5% in both groups). However, when 1,351 patients who had received aspirin were matched to 1,351 patients with the same propensity to receive aspirin but who did not, mortality was 47% lower in those treated ($P = 0.002$).

Analyses using propensity scores have several advantages. The number of potential confounding variables that can be modeled as predictors of an intervention is usually greater than the number of variables that can be modeled as predictors of an outcome, because the number of people treated is generally much greater than the number who develop the outcome (2,310 compared with 276 in the Example 9.1). Another reason that more confounders can be included is that there is no danger of "overfitting" the propensity model—interaction terms, quadratic terms, and multiple indicator variables can all be included (15). Finally, investigators are usually more confident in identifying the determinants of treatment than the determinants of outcome, because the treatment decisions were made by clinicians based on a limited number of patient characteristics.

Of course, like other multivariate techniques, use of propensity scores still requires that potential confounding variables be identified and measured. A limitation of this technique is that it does not provide information about the relationship between any of the confounding variables and outcome—the only result is for the predictor (usually, a treatment) that was modeled. However, because this is an analysis phase strategy, it does not preclude doing more traditional multivariate analyses as well, and both types of analysis are usually done.

■ OTHER PITFALLS IN QUANTIFYING CAUSAL EFFECTS

Conditioning on a Shared Effect

The bias caused by **conditioning on a shared effect** is kind of tricky, and it is sometimes skipped in introductory textbooks because most explanations of it use abstract diagrams and notation. By contrast, we will first give a few examples of how it might occur, and then try to explain what the name means.

Consider a study of people who have lost at least 15 pounds in the previous year. An investigator finds that the subjects who have been dieting have a lower risk of cancer than those who have not been dieting. Do you think dieting prevented cancer in these subjects?

If you stop and think, you'll probably answer no, because cancer also causes weight loss. You can imagine that if someone loses weight for no apparent reason it is much more likely to signify a cancer than if someone loses weight while dieting. *Among people who have lost weight*, if the weight loss was not caused by dieting, it is more likely to have been caused by something more ominous. The investigators created an inverse association between dieting and cancer by conditioning on (restricting attention to) a shared effect (weight loss, which is caused by both dieting and cancer).

Here's another example. Among low birth weight babies, those whose mothers smoked during pregnancy have lower infant mortality than those whose mothers did not smoke (16). Should we encourage more mothers to smoke during pregnancy? Definitely not! The reason for this observation is that smoking causes low birth weight, but so do other things, especially prematurity. So *among low birth weight babies*, if the low birth weight was not caused by smoking, it is more likely to have been caused by prematurity. The investigators created an inverse

association between smoking and prematurity (and its associated mortality risk) by conditioning on (restricting attention to) a shared effect (low birth weight, which is caused by both smoking and prematurity).

Now the phrase "conditioning on a shared effect" makes sense. **Conditioning** is an epidemiologic term that means looking at associations between predictor and outcome variables "conditioned on" (i.e., at specified levels of) some attribute. A **shared effect** refers to an attribute (like losing weight, or being a low birth weight baby) that has several causes. Bias due to conditioning on a shared effect can occur if the investigator treats something *caused* by the risk factor being studied as an inclusion criterion, a matching variable, or a possible confounding variable.

Underestimation of Causal Effects

To this point, our emphasis has been on evaluating the likelihood of alternative explanations for an association, in order to avoid concluding that an association is real and causal when it is not. However, another type of error is also possible—*underestimation* of causal effects. Chance, bias, and confounding can also be reasons why a real association might be missed or underestimated.

We discussed **chance** as a reason for missing an association in Chapter 5, when we reviewed type II errors and the need to make sure the sample size will provide adequate **power** to find real associations. After a study has been completed, however, the power calculation is no longer a good way to quantify uncertainty due to random error. At this stage a study's hypothetical power to detect an effect of a specified size is less relevant than the actual findings, expressed as the observed estimate of association (e.g., risk ratio) and its 95% **confidence interval** (17).

Bias can also distort estimates of association toward no effect. In Chapter 8, the need for blinding in ascertaining risk factor status among cases and controls was to avoid **differential measurement bias**, for example, differences between the cases and controls in the way questions were asked or answers interpreted that might lead observers to get the answers they desire. Because observers might desire results in either direction, differential measurement bias can bias results to either overestimate or underestimate causal effects. Non-differential bias, on the other hand, will generally lead to underestimation of associations.

Confounding can also lead to attenuation of real associations. For example, suppose coffee drinking actually protected against MI, but was more common in smokers. If smoking were not controlled for, the beneficial effects of coffee might be missed—coffee drinkers might appear to have the same risk of MI as those who did not drink coffee, when their higher prevalence of smoking should have caused their risk to be higher. This type of confounding, in which the effects of a beneficial factor are hidden by its association with a cause of the outcome, is sometimes called **suppression** (18). It is a common problem for observational studies of treatments, because treatments are often most indicated in those at higher risk of a bad outcome. The result, noted earlier, is that a beneficial treatment can appear to be useless (as aspirin did in Example 9.1) or even harmful until the confounding by indication is controlled.

■ CHOOSING A STRATEGY

What general guidelines can be offered for deciding whether to cope with confounders during the design or analysis phases, and how best to do it? The use of **specification** to control confounding is most appropriate for situations in which the investigator is chiefly interested in specific subgroups of the population; this is really just a special form of the general process of establishing criteria for selecting the study subjects (Chapter 3). However, for studies in which causal inference is the goal, there's the additional caution to avoid inclusion criteria that could be caused by predictor variables you wish to study (i.e., conditioning on a shared effect).

An important decision to make in the design phase of the study is whether to **match**. Matching is most appropriate for case–control studies and fixed constitutional factors such as age,

race, and sex. Matching may also be helpful when the sample size is small compared with the number of strata necessary to control for known confounders, and when the confounders are more easily matched than measured. However, because matching can permanently compromise the investigator's ability to observe real associations, it should be used sparingly, particularly for variables that may be in the causal chain. In many situations the analysis phase strategies (stratification, adjustment, and propensity scores) are just as good for controlling confounding, and have the advantage of being **reversible**—they allow the investigator to add or subtract covariates to explore different causal models.

Although not available for all research questions, it is always worth considering the possibility of an **opportunistic** study design. If you don't stop and consider (and ask your colleagues about) these studies, you might miss a great opportunity to do one.

The final decision to **stratify**, **adjust**, or use **propensity scores** need not be made until after the data are collected; in many cases the investigator may wish to do all of the above. However, it is important during study design to consider which factors may later be used for adjustment, in order to know which variables to measure. In addition, because different analysis phase strategies for controlling confounding do not always yield the same results, it is best to specify a primary analysis plan in advance. This may help investigators resist the temptation of selecting the strategy that provides the most desired results.

Evidence Favoring Causality

The approach to enhancing causal inference has largely been a negative one thus far—how to rule out the four rival explanations in Table 9.1. A complementary strategy is to seek characteristics of associations that provide positive evidence for causality, of which the most important are the consistency and strength of the association, the presence of a dose–response relation, and biologic plausibility.

When the results are **consistent** in studies of various designs, it is less likely that chance or bias is the cause of an association. Real associations that represent effect–cause or confounding, however, will also be consistently observed. For example, if cigarette smokers drink more coffee and have more MIs in the population, studies will consistently observe an association between coffee drinking and MI.

The **strength** of the association is also important. For one thing, stronger associations give more significant *P* values, making chance a less likely explanation. Stronger associations also provide better evidence for causality by reducing the likelihood of confounding. Associations due to confounding are indirect (i.e., via the confounder) and therefore are generally weaker than direct cause–effect associations. This is illustrated in Appendix 9A: The strong associations between coffee and smoking (odds ratio = 16) and between smoking and MI (odds ratio = 4) led to a much weaker association between coffee and MI (odds ratio = 2.25).

A **dose–response** relation provides positive evidence for causality. The association between cigarette smoking and lung cancer is an example: Moderate smokers have higher rates of cancer than do nonsmokers, and heavy smokers have even higher rates. Whenever possible, predictor variables should be measured continuously or in several categories, so that any dose–response relation that is present can be observed. Once again, however, a dose–response relation can be observed with effect–cause associations or with confounding.

Finally, **biologic plausibility** is an important consideration for drawing causal inference—if a causal mechanism that makes sense biologically can be proposed, evidence for causality is enhanced, whereas associations that do not make sense given our current understanding of biology are less likely to represent cause–effect. For example, in the study of marijuana use as a risk factor for germ cell tumors, use of marijuana less than once a day was associated with lower risk than no use (6). It is hard to explain this biologically.

It is important not to overemphasize biologic plausibility, however. Investigators seem to be able to come up with a plausible mechanism for virtually any association and some associations

originally dismissed as biologically implausible, such as a bacterial etiology for peptic ulcer disease, have turned out to be real.

SUMMARY

1. The design of **observational studies** should anticipate the need to interpret **associations**. The inference that the association represents a **cause–effect** relationship (often the goal of the study) is strengthened by strategies that reduce the likelihood of the **four rival explanations—chance**, **bias**, **effect–cause**, and **confounding**.
2. The role of **chance (random error)** can be minimized by designing a study with **adequate sample size and precision** to assure low **type I and type II error** rates. Once the study is completed, the effect of random error can be judged from the width of the **95% confidence interval** and the consistency of the results with **previous evidence**.
3. **Bias (systematic error)** arises from differences between the population and phenomena addressed by the research question and the actual subjects and measurements in the study. Bias can be minimized by basing design decisions on a **judgment** as to whether these differences will lead to a wrong answer to the research question.
4. **Effect–cause** is made less likely by designing a study that permits assessment of **temporal sequence**, and by considering **biologic plausibility**.
5. **Confounding**, which may be present when a third variable is associated with the predictor of interest and is a cause of the outcome, is made less likely by the following strategies, most of which require potential confounders to be anticipated and measured:
   a. **Specification** or **matching** in the **design phase**, which alters the sampling strategy to ensure that only groups with similar levels of the confounder are compared. These strategies **should be used judiciously** because they can **irreversibly** limit the information available from the study.
   b. **Analysis phase** strategies that accomplish the same goal and preserve options for investigating causal paths:
      - **Stratification**, which in addition to controlling for confounding can reveal **effect modification** ("**interaction**"), a different magnitude of predictor–outcome association at different levels of a third variable.
      - **Adjustment**, which can permit the impact of many predictor variables to be controlled simultaneously.
      - **Propensity scores**, which enhance the power for addressing **confounding by indication** in observational studies of treatment efficacy.
6. Investigators should be on the lookout for **opportunistic** observational designs, including **natural experiments**, **Mendelian randomization**, and other **instrumental variable** designs, that offer a strength of causal inference that can approach that of a randomized clinical trial.
7. Investigators should avoid **conditioning on shared effects** in the design phase by not selecting subjects based on covariates that might be caused by the predictor, and in the analysis phase by not controlling for these covariates.
8. Causal inference can be enhanced by positive evidence, notably the **consistency and strength of the association**, the presence of a **dose–response** relation, and **biologic plausibility**.

APPENDIX 9A
Hypothetical Example of Confounding and Effect Modification

The entries in these tables are numbers of subjects in this hypothetical case–control study

Panel 1. If we look at the entire group of study subjects, there appears to be an association between coffee drinking and MI:

	Smokers and Nonsmokers Combined	
	MI	No MI
Coffee	90	60
No coffee	60	90

Odds ratios (OR) for MI associated with coffee in smokers and nonsmokers combined $= \dfrac{90 \times 90}{60 \times 60} = 2.25$

Panel 2. However, this could be due to **confounding**, as shown by the tables stratified on smoking which show that coffee drinking is not associated with MI in either smokers or nonsmokers:

	Smokers	
	MI	No MI
Coffee	80	40
No coffee	20	10

	Nonsmokers	
	MI	No MI
Coffee	10	20
No coffee	40	80

Odds ratios for MI associated with coffee:

$$\text{OR in smokers} = \frac{80 \times 10}{20 \times 40} = 1 \qquad \text{OR in nonsmokers} = \frac{10 \times 80}{40 \times 20} = 1$$

Smoking is a confounder because it is strongly associated with coffee drinking (below, left panel) and with MI (below, right panel): These tables were obtained by rearranging numbers in Panel 2.

	MI and No MI Combined	
	Coffee	No Coffee
Smokers	120	30
Nonsmokers	30	120

	Coffee and No Coffee Combined	
	MI	No MI
Smokers	100	50
Nonsmokers	50	100

Odds ratio for coffee drinking associated with smoking $= \dfrac{120 \times 120}{30 \times 30} = 16$

Odds ratio for MI associated with smoking $= \dfrac{100 \times 100}{50 \times 50} = 4$

Panel 3. The association between coffee drinking and MI in Panel 1 could also represent **effect modification**, if stratification on smoking revealed that the association between coffee drinking and MI differs in smokers and nonsmokers. In the table below, the OR of 2.25 for the association between coffee drinking and MI in smokers and nonsmokers combined is due entirely to a strong association in smokers. When effect modification is present, the odds ratios in different strata are different, and must be reported separately:

	Smokers	
	MI	No MI
Coffee	50	15
No coffee	10	33

	Nonsmokers	
	MI	No MI
Coffee	40	45
No coffee	50	57

Odds ratios for MI associated with coffee:

$$\text{OR in smokers} = \frac{50 \times 33}{15 \times 10} = 11 \qquad \text{OR in nonsmokers} = \frac{40 \times 57}{45 \times 50} = 1$$

Bottom Line: The overall association between coffee drinking and MI in Panel 1 could be hiding the presence of confounding by smoking, which would be revealed by stratification on smoking (Panel 2). Or it could be hiding the presence of effect modification by smoking, which would also be revealed by stratification on smoking (Panel 3). It could also represent cause-effect, which would be supported (though not proven) if stratification on smoking did not alter the association between coffee drinking and MI. Finally (and most realistically), if could be a result of some mixture of all of the above.

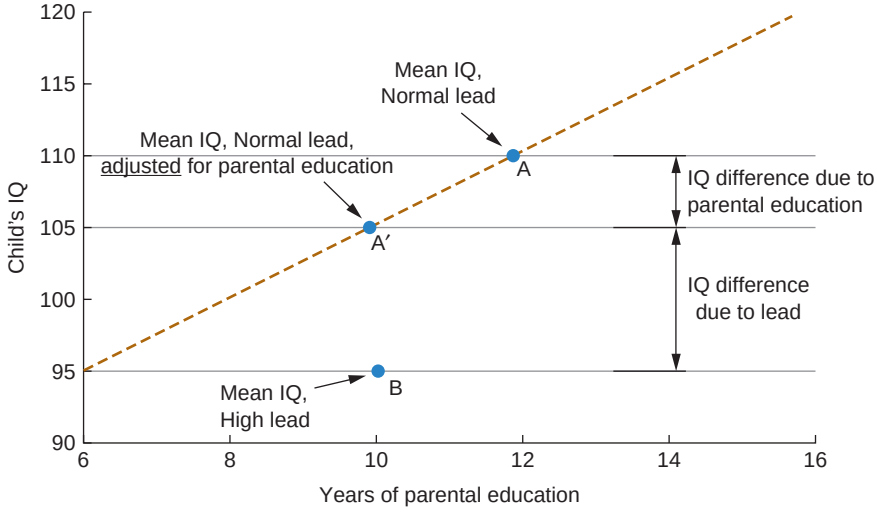
APPENDIX 9B
A Simplified Example of Adjustment

Suppose that a study finds two major predictors of the intelligence quotient (IQ) of children: the parental education level and the child's blood lead level. Consider the following hypothetical data on children with normal and high lead levels:

	Average Years of Parental Education	Average IQ of Child
High lead level	10.0	95
Normal lead level	12.0	110

Note that the parental education level is also associated with the child's blood lead level. The question is, "Is the difference in IQ between children with normal and high lead levels more than can be accounted for on the basis of the difference in parental education?" To answer this question we look at how much difference in IQ the difference in parental education levels would be expected to produce. We do this by plotting parental educational level versus IQ in the children with normal lead levels (Figure 9.2).[4]

The diagonal dashed line in Figure 9.2 shows the relationship between the child's IQ and parental education in children with normal lead levels; there is an increase in the child's IQ of 5 points for each 2 years of parental education. Therefore, we can adjust the IQ of the normal lead group to account for the difference in mean parental education by sliding down the line from



■ **FIGURE 9.2** Hypothetical graph of child's IQ as a linear function (*dashed line*) of years of parental education.

[4]This description of analysis of covariance (ANCOVA) is simplified. Actually, parental education is plotted against the child's IQ in both the normal and high lead groups, and the single slope that fits both plots the best is used. The model for this form of adjustment therefore assumes linear relationships between education and IQ in both groups, and that the slopes of the lines in the two groups are the same.

point *A* to point *A′*. (Because the group with normal lead levels had 2 more years of parental education on the average, we adjust their IQs downward by 5 points to make them comparable in mean parental education to the high lead group.) This still leaves a 10-point difference in IQ between points *A* and *B*, suggesting that lead has an independent effect on IQ of this magnitude. Therefore, of the 15-point difference in IQ of children with low and high lead levels, 5 points can be accounted for by their parents' different education levels and the remaining 10 are attributable to the lead exposure.

REFERENCES

1. McEvoy SP, Stevenson MR, McCartt AT, et al. Role of mobile phones in motor vehicle crashes resulting in hospital attendance: a case-crossover study. *BMJ* 2005;331(7514):428.
2. Magruder JT, Elahi D, Andersen DK. Diabetes and pancreatic cancer: chicken or egg? *Pancreas* 2011;40(3):339–351.
3. Huxley R, Ansary-Moghaddam A, Berrington de Gonzalez A, et al. Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies. *Br J Cancer* 2005;92(11):2076–2083.
4. Bosetti C, Rosato V, Polesel J, et al. Diabetes mellitus and cancer risk in a network of case-control studies. *Nutr Cancer* 2012;64(5):643–651.
5. Maconochie N, Doyle P, Carson C. Infertility among male UK veterans of the 1990-1 Gulf war: reproductive cohort study. *BMJ* 2004;329(7459):196–201.
6. Trabert B, Sigurdson AJ, Sweeney AM, et al. Marijuana use and testicular germ cell tumors. *Cancer* 2011; 117(4):848–853.
7. Newman TB, Kohn M. *Evidence-based diagnosis*. New York: Cambridge University Press, 2009. Chapter 10.
8. Lofgren RP, Gottlieb D, Williams RA, et al. Post-call transfer of resident responsibility: its effect on patient care [see comments]. *J Gen Intern Med* 1990;5(6):501–505.
9. Bell CM, Redelmeier DA. Mortality among patients admitted to hospitals on weekends as compared with weekdays. *N Engl J Med* 2001;345(9):663–668.
10. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;32(1):1–22.
11. Cherry N, Mackness M, Durrington P, et al. Paraoxonase (PON1) polymorphisms in farmers attributing ill health to sheep dip. *Lancet* 2002;359(9308):763–764.
12. Hearst N, Newman TB, Hulley SB. Delayed effects of the military draft on mortality. A randomized natural experiment. *N Engl J Med* 1986;314(10):620–624.
13. Tan HJ, Norton EC, Ye Z, et al. Long-term survival following partial vs radical nephrectomy among older patients with early-stage kidney cancer. *JAMA* 2012;307(15):1629–1635.
14. Gum PA, Thamilarasan M, Watanabe J, et al. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: A propensity analysis. *JAMA* 2001;286(10):1187–1194.
15. Klungel OH, Martens EP, Psaty BM, et al. Methods to assess intended effects of drug treatment in observational studies are reviewed. *J Clin Epidemiol* 2004;57(12):1223–1231.
16. Hernandez-Diaz S, Schisterman EF, Hernan MA. The birth weight "paradox" uncovered? *Am J Epidemiol* 2006;164(11):1115–1120.
17. Bacchetti P. Current sample size conventions: flaws, harms, and alternatives. *BMC Med* 2010;8:17.
18. Katz MH. *Multivariable analysis: a practical guide for clinicians*, 2nd ed. Cambridge, UK; New York: Cambridge University Press, 2006.

CHAPTER 10

Designing a Randomized Blinded Trial

Steven R. Cummings, Deborah Grady, and Stephen B. Hulley

In a clinical trial, the investigator applies an **intervention** and observes the effect on one or more **outcomes**. The major advantage of a trial over an observational study is the ability to **demonstrate causality**. **Randomly assigning** the intervention minimizes the influence of confounding variables, and **blinding** its administration minimizes the possibility that the apparent effects of the intervention are due to differential use of other treatments in the intervention and control groups or to biased ascertainment or adjudication of the outcome.

However, a clinical trial is generally **expensive** and **time-consuming**, addresses a **narrow question**, and sometimes exposes participants to **potential harm**. For these reasons, trials are best reserved for relatively mature research questions, when observational studies and other lines of evidence suggest that an intervention might be effective and safe but stronger evidence is required before it can be approved or recommended. Not every research question is amenable to the clinical trial design—it is not feasible to study whether drug treatment of high-LDL cholesterol in children will prevent heart attacks many decades later and it is not ethical to randomize people to smoke real or sham cigarettes to determine the effect on lung cancer. But clinical trial evidence on the efficacy and safety of clinical interventions should be obtained whenever possible.

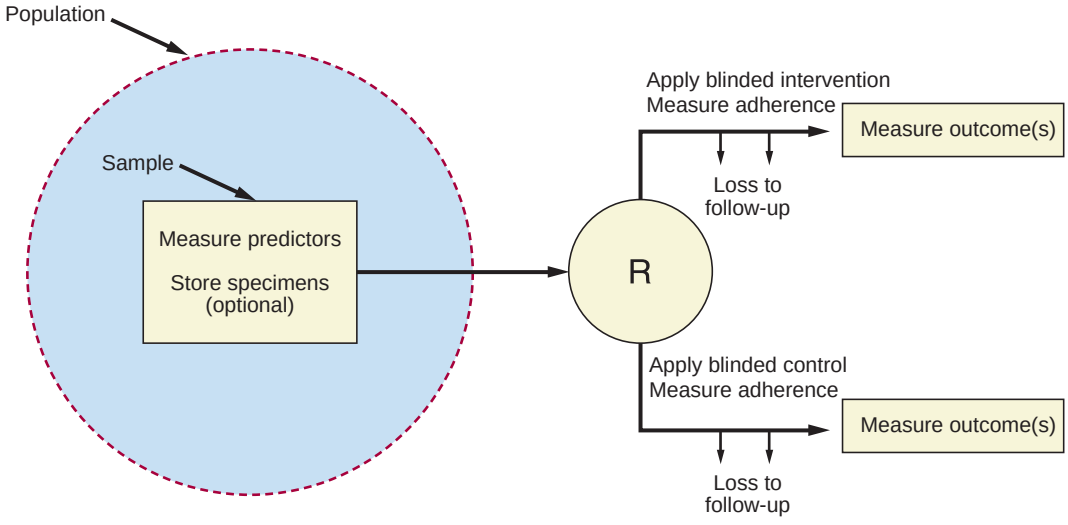
This chapter focuses on **designing** the classic **randomized blinded trial**: choosing the **intervention and control conditions**, defining **outcomes and adverse effects**, selecting **participants**, measuring **baseline and outcome variables**, and evaluating approaches to **randomizing** and **blinding**. In the next chapter we will cover alternative trial designs, and implementation and analysis issues.

SELECTING THE INTERVENTION AND CONTROL CONDITIONS

The classic randomized trial is a **parallel**, **between-group design** that includes a group that receives an intervention to be tested, and a control group that receives either no active treatment (preferably a placebo) or a comparison treatment. The investigator applies the intervention and control, follows both groups over time, and compares the outcome between the intervention and control groups (Figure 10.1).

Choice of Intervention

The choice of intervention is the critical first step in designing a clinical trial. Investigators should consider several issues as they design the intervention, including the dosage, duration, and frequency of the intervention that best balances efficacy and safety. It is also important to consider the feasibility of blinding, whether to treat with one or a combination of interventions, acceptability to participants, and generalizability to the way the treatment will be used in practice. If important decisions are uncertain, such as which dose best balances efficacy and



■ **FIGURE 10.1** In a randomized blinded trial, the steps are to

- Select a sample of subjects from a population suitable to receive the intervention.
- Measure the predictor variables and, if appropriate, the baseline level of the outcome variable.
- Consider the option of storing serum, images, and so on, for later analysis.
- Randomly assign the blinded intervention and control condition (e.g., placebo).
- Follow the cohort over time, minimizing loss to follow-up and assessing compliance with the intervention and control.
- Measure the outcome variables.

safety, it is generally best to postpone major or costly trials until preliminary studies have been completed to help resolve the issue.

The best balance between **efficacy** and **safety** depends on the intervention and the condition being studied. On the one hand, efficacy is generally the paramount consideration in designing interventions to treat illnesses that cause severe symptoms or death. Therefore, it may be best to choose the highest tolerable dose for treatment of metastatic cancer. On the other hand, safety should be the primary criterion for designing interventions to treat symptomatic conditions that rarely result in disease progression or death. **Preventive** therapy for healthy people should meet stringent tests of safety: If it is effective, the treatment will prevent the condition in a few persons, but everyone treated will be at risk of the adverse effects of the therapy. In this case, it is generally best to choose the dose that maximizes efficacy with a very low risk of side effects. If the best dose is not certain based on prior animal and human research findings, there may be a need for additional trials that compare the effects of multiple doses on intermediate markers or clinical outcomes (see phase II trials, Chapter 11).

Sometimes an investigator may decide to compare **several doses** or levels of intensity with a single control group. For example, at the time the Multiple Outcomes of Raloxifene Evaluation Trial was designed, it was not clear which dose of raloxifene (60 or 120 mg) was best, so the trial tested two doses for preventing vertebral fractures (1). This is sometimes a reasonable strategy, but it has costs: a larger and more expensive trial, and the complexity of dealing with multiple hypotheses (Chapter 5).

For some treatments the dose is adjusted to optimize the effect for each individual patient. In these instances, it may be best to design an intervention so that the dose of **active drug is titrated** to achieve a clinical outcome such as reduction in the hepatitis C viral load. To maintain blinding, corresponding changes should be made (by someone not otherwise involved in the trial) in the "dose" of placebo for a randomly selected or matched participant in the placebo group.

Trials to test **single interventions** are generally much easier to plan and implement than those testing combinations of treatments. However, many medical conditions, such as HIV infection or congestive heart failure, are treated with **combinations** of drugs or therapies. The most important disadvantage of testing combinations of treatments is that the result cannot provide clear conclusions about any one element of the interventions. In one of the Women's Health Initiative trials, for example, postmenopausal women were treated with estrogen plus progestin therapy or placebo. The intervention increased the risk of several outcomes, including breast cancer; however, it was unclear whether the effect was due to the estrogen or the progestin (2). In general, it is preferable to design trials that have only one major difference between any two study groups.

The investigator should consider how receptive participants will be to the proposed intervention, and whether it can be blinded. Another consideration is how well the intervention can be incorporated in practice. **Simple interventions** are generally better than complicated ones (patients are more likely to take a pill once a day than subcutaneous injections two or three times a day). Complicated interventions with qualitative aspects, such as multifaceted counseling about changing behavior, may not be feasible to incorporate in general practice because they are difficult to replicate, time-consuming, and costly. Such interventions are less likely to have public health impact, even if a trial proves that they are effective.

Choice of Control

The best control group receives **no active treatment** in a way that can be **blinded**, which for medications generally requires a **placebo** that is indistinguishable from active treatment. This strategy compensates for any placebo effect of the active intervention (i.e., through suggestion or expectation) so that any outcome difference between study groups can be ascribed to a specific effect of the intervention.

The cleanest comparison between the intervention and control groups occurs when there are no **co-interventions**—medications, therapies, or behaviors (other than the study intervention) that alter the risk of developing the outcome of interest. For example, in a randomized trial evaluating a yoga intervention compared to usual care to prevent diabetes, study staff may urge participants to exercise and to lose weight. These are potentially effective co-interventions that may reduce the risk of developing diabetes. If participants in both groups use effective co-interventions, the rate of outcomes will be decreased, power will be reduced, and the sample size will need to be larger or the trial longer. If use of effective co-interventions differs between the intervention and control groups, the outcome will be biased. In the absence of effective blinding, the protocol must include plans to obtain data to allow statistical adjustment for differences between the groups in the rate of use of such co-interventions during the trial. However, measuring co-interventions may be difficult, and adjusting for such postrandomization differences should be viewed as a secondary or explanatory analysis because it violates the intention-to-treat principle (Chapter 11).

Often it is not possible to withhold treatments other than the study intervention. For example, in a trial of a new drug to reduce the risk of myocardial infarction in persons with known coronary heart disease (CHD), the investigators cannot ethically prohibit or discourage participants from taking medical treatments that are indicated for persons with known CHD, such as aspirin, statins, and beta-blockers. One solution is to give **standard care drugs** to all participants in the trial; although this approach reduces the overall event rate and therefore increases the required sample size, it tests the most relevant clinical question: whether the new intervention improves the outcome when given in addition to standard care.

When the treatment to be studied is a new drug that is believed to be a good alternative to standard care, one option is to design a **non-inferiority** or **equivalence trial** in which the new treatment is compared with the one that is already proven to be effective (see Chapter 11).

CHOOSING OUTCOME MEASUREMENTS

The definition of the specific outcomes of the trial influences many other design components, as well as the cost and feasibility of the trial. Trials should usually include several outcomes to increase the richness of the results and possibilities for secondary analyses. However, one of these should be designated as the ***primary* outcome** that reflects the main question, allows calculation of the sample size, and sets the priority for efforts to implement the study.

Clinical outcomes provide the best evidence about whether and how to use treatments or preventive interventions. However, for outcomes that are uncommon, such as the occurrence of cancer, trials must generally be large, long, and expensive. As noted in Chapter 6, outcomes measured as continuous variables, such as quality of life, can generally be studied with fewer participants and than dichotomous outcomes. However the most important clinical outcome is sometimes unavoidably dichotomous, such as recurrence of cancer.

Intermediate markers, such as bone density, are measurements that are related to the clinical outcome. Trials that use intermediate outcomes can further our understanding of pathophysiology and provide information for choosing the best dose or frequency of treatment in trials with clinical outcomes. The clinical relevance of trials with intermediate outcomes depends on how accurately changes in these markers, especially changes that occur due to treatment, represent changes in the risk of clinical outcomes. Intermediate markers can be considered **surrogate markers** for the clinical outcome to the extent that treatment-induced changes in the marker consistently predict how treatment changes the clinical outcome (3). Generally, a good surrogate marker measures changes in an intermediate factor in the main pathway that determines the clinical outcome.

HIV viral load is a good surrogate marker because treatments that reduce the viral load consistently reduce morbidity and mortality in patients with HIV infection. In contrast, bone mineral density (BMD) is a poor surrogate marker (3). It reflects the amount of mineral in a section of bone, but treatments that improve BMD sometimes have little or no effect on fracture risk, and the magnitude of increase in BMD is not consistently related to how much the treatment reduces fracture risk (4). The best evidence that a biological marker is a good surrogate comes from randomized trials of the clinical outcome (fractures) that also measure change in the potential surrogate marker (BMD) in all participants. If the marker is a good surrogate, then statistical adjustment for changes in the marker will account for much of the effect of treatment on the outcome (3).

Number of Outcome Variables

It is often desirable to have **several outcome variables** that measure different aspects of the phenomena of interest. In the Heart and Estrogen/Progestin Replacement Study (HERS), coronary heart disease events were chosen as the primary endpoint. Coronary revascularization, hospitalization for unstable angina or congestive heart failure, stroke, transient ischemic attack, venous thromboembolic events, and all-cause mortality were also ascertained and adjudicated to provide a more detailed description of the cardiovascular effects of hormone therapy (5). However, a **single primary outcome** (CHD events) was designated for the purpose of planning the sample size and duration of the study and to avoid the problems of interpreting tests of multiple hypotheses (Chapter 5).

Composite Outcomes

Some trials define outcomes that are composed of a number of different events or measures. For example, many trials of interventions to reduce the risk of coronary heart disease include several specific coronary events in the outcome, such as myocardial infarction, coronary death, and coronary revascularization procedures. This may be reasonable if each of these outcomes is clinically important, the treatment works by similar mechanisms for each condition, and the

intervention is expected to reduce the risk of each event. In addition, a composite outcome generally provides greater power than a single outcome because there will be more events. However, composite outcomes that include events that are not as clinically meaningful or occur much more commonly than others in the composite can result in misleading findings. For example, if hospitalization for evaluation of chest pain is added to the composite coronary outcome, this event will dominate the composite if such hospitalizations occur much more commonly than myocardial infarction, coronary death, or revascularization. An intervention that alters the composite may then be reported to reduce the risk of "coronary events," when in reality it only reduces the risk of hospitalization for chest pain.

Composite outcomes must be carefully designed. If treatment produces only a small effect on one outcome, especially if that outcome is relatively common, it may add little statistical power or even increase the sample size required to detect an effect. For example, if stroke is added to a composite "cardiovascular outcome," the intervention might reduce the risk of coronary events, have no impact, or even increase the risk of stroke and therefore be found to have no effect on the composite cardiovascular outcome.

Adverse Effects

The investigator should include outcome measures that will detect the occurrence of **adverse effects** that may result from the intervention. Revealing whether the beneficial effects of an intervention outweigh the adverse ones is a major goal of most clinical trials, even those that test apparently innocuous treatments like a health education program. Adverse effects may range from relatively minor symptoms, such as a mild transient rash, to serious and fatal complications. The rate of occurrence, effect of treatment, and sample size requirements for detecting adverse effects is generally different from those required for detecting benefits. Unfortunately, rare side effects will usually be impossible to detect even in large trials and are only discovered (if at all) by large observational studies or case reports after an intervention is in widespread clinical use.

In the early stages of testing a new treatment when potential adverse effects are unclear, investigators should ask broad, open-ended questions about all types of potential adverse effects. In large trials, assessment and coding of all potential adverse events can be very expensive and time-consuming, often with a low yield of important results. Investigators should consider strategies for minimizing this burden while preserving an adequate assessment of potential harms of the intervention. For example, in very large trials, common and minor events, such as upper respiratory infections and gastrointestinal upset, might be recorded in a subset of the participants. It may not be necessary to record adverse effects that are not serious if previous studies have found no differences in the incidence of minor symptoms. In addition to these open-ended questions, specific queries should be designed to discover important adverse events that are expected because of previous research or clinical experience. For example, because myositis is a reported side effect of treatment with statins, the signs and symptoms of myositis should be queried in any trial of a new statin.

Adverse effects that are reported as symptoms or clinical terms must be coded and categorized for analysis. MedDRA (www.ich.org/products/meddra.html) and SNOMED (https://www.nlm.nih.gov/research/umls/) are commonly used dictionaries of terms that are grouped in several ways, as symptoms, specific diagnoses, and according to organ system. For example, an adverse event recorded as "fever and cough" and an adverse event recorded as "bronchitis," will be grouped with other conditions, like pneumonia, as a "respiratory infection" and, at a higher level, as an adverse effect in the respiratory system. These classification schemes provide a good general summary of adverse effects and are reasonably accurate for diseases that are specifically diagnosed, such as fractures. However, they may miss important adverse events that are described by several terms if these terms are not grouped together. For example, in a trial of denosumab for prevention of osteoporotic fractures, MedDRA coded cases

of cellulitis separately from cases of erysipelas (two names for the same type of infection). When combined, 12 serious cases of cellulitis occurred with denosumab versus 1 with placebo (P <0.001) (6).

Adverse effects are also generally classified by severity. **Serious adverse events (SAEs)** are defined as death or life-threatening events, events requiring or extending hospitalization, disability or permanent damage, birth defects, and other important medical events that may require medical or surgical intervention to prevent one of the other outcomes (www.fda.gov/Safety/MedWatch/HowToReport/ucm053087.htm). Serious adverse events generally must be promptly reported to institutional review boards and to the sponsor of the trial.

When data from a trial is used to apply for regulatory approval of a new drug, the trial design must satisfy regulatory expectations for reporting adverse events (http://www.fda.gov/Drugs/InformationOnDrugs/ucm135151.htm). Certain disease areas, such as cancer, have established methods for classifying adverse events (http://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm).

■ SELECTING THE PARTICIPANTS

Chapter 3 discussed how to specify entry criteria defining a target population that is appropriate to the research question and an accessible population that is practical to study, how to design an efficient and scientific approach to selecting participants, and how to recruit them. Here we cover issues that are especially relevant to clinical trials.

Define Entry Criteria

In a clinical trial, inclusion and exclusion criteria have the goal of identifying a population in which it is feasible, ethical, and relevant to study the impact of the intervention on outcomes. **Inclusion criteria** should produce a sufficient number of participants who have a high enough rate of the primary outcome to achieve adequate power to find an important effect of the intervention on the outcome. On the other hand, criteria should also maximize the generalizability of findings from the trial and the ease of recruitment. For example, if the outcome of interest is an uncommon event, such as breast cancer, it is usually necessary to recruit high-risk participants in order to reduce the sample size and follow-up time to feasible levels. On the other hand, narrowing the inclusion criteria to high-risk women limits the generalizability of the results and makes it more difficult to recruit enough participants into the trial.

To plan the right **sample size**, the investigator must have reliable estimates of the rate of the primary outcome in people who might be enrolled. These estimates can be based on data from vital statistics, longitudinal observational studies, or rates observed in the untreated group in trials with inclusion criteria similar to those in the planned trial. For example, expected rates of pancreatic cancer in adults can be estimated from cancer registry data. The investigator should keep in mind, however, that screening and healthy volunteer effects generally mean that event rates among those who qualify and agree to enter clinical trials are lower than in the general population; it may be preferable to obtain rates of pancreatic cancer from the placebo group of other trials with similar inclusion criteria.

Including persons with a **high risk** of the outcome can decrease the number of participants needed for the trial. If risk factors for the outcome have been established, then the selection criteria can be designed to include participants who have a minimum estimated risk of the outcome of interest. The Raloxifene Use for The Heart trial, designed to test the effect of raloxifene for prevention of cardiovascular disease (CVD) and breast cancer, enrolled women who were at increased risk of CVD based on a combination of risk factors (7). Another way to increase the rate of events is to limit enrollment to people who already have the disease. The Heart and Estrogen/Progestin Replacement Study included 2,763 women who already had coronary heart disease (CHD) to test whether estrogen plus progestin reduced the risk of new CHD events (5).

This approach was much less costly than the Women's Health Initiative trial of the same research question in women without CHD, which required about 17,000 participants (8).

Although **probability samples of general populations** confer advantages in observational studies, this type of sampling is generally not feasible or necessary for randomized trials. Inclusion of participants with diverse characteristics will increase the confidence that the results of a trial apply broadly. However, unless there are biological or genetic differences between populations that influence the effect of a treatment, it is generally true that results of a trial done in a convenience sample (e.g., women with CHD who respond to advertisements) will be similar to results obtained in probability samples of eligible people (all women with CHD). Occasionally, the efficacy of treatment depends on characteristics of the subjects; this is termed **effect modification or interaction** (see Chapter 11) For example, some osteoporosis treatments substantially reduce the risk of fracture in women with very low bone density (T-scores below –2.5) with little or no effect in women with higher bone density ($P = 0.02$ for interaction) (9, 10). In this case, including only women with very low bone density in a trial may increase the effect size and reduce the sample size for a trial of similar treatments.

Stratification of participants by a characteristic, such as racial group, allows investigators to enroll a desired number of participants with a characteristic that may have an influence on the effect of the treatment or its generalizability. Recruitment to a stratum can be closed when the goal for participants with that characteristic has been reached. However, since most trials are not designed with sufficient sample size to test for heterogeneity in the effects of the intervention among such subgroups, this strategy may be of limited practical value.

Exclusion criteria should be parsimonious because unnecessary exclusions may make it more difficult to recruit the necessary number of participants, diminish the generalizability of the results, and increase the complexity and cost of recruitment. There are five main reasons for excluding people from a clinical trial (Table 10.1).

TABLE 10.1 REASONS FOR EXCLUDING PEOPLE FROM A CLINICAL TRIAL

REASON	EXAMPLE: A TRIAL OF RALOXIFENE (A SELECTIVE ESTROGEN RECEPTOR MODULATOR) VERSUS PLACEBO TO PREVENT HEART DISEASE
1. A study treatment may be harmful.	
• Unacceptable risk of harm if assigned to active treatment	Prior venous thromboembolic event (raloxifene increases risk of venous thromboembolic events)
• Unacceptable risk of harm if assigned to control	Recent estrogen receptor–positive breast cancer (treatment with a selective estrogen receptor modulator is effective, and a standard of care)
2. Active treatment is unlikely to be effective.	
• At low risk for the outcome	Teenaged women with very low risk for coronary heart disease
• Has a type of disease that is not likely to respond to treatment	Patient with valvular heart disease, which is not likely to respond to the hypothesized anti-atherogenic effects of raloxifene
• Taking a treatment that is likely to interfere with the intervention	Taking estrogen therapy (which competes with raloxifene)
3. Unlikely to adhere to the intervention.	Poor adherence during the run-in period (Chapter 11).
4. Unlikely to complete follow-up.	Plans to move before trial ends and won't be available for final outcome measures Short life expectancy because of a serious illness
5. Practical problems with participating in the protocol.	Impaired mental state that prevents accurate answers to questions

Potential participants should be excluded if the treatment or control is **unsafe**. The active treatment may be unsafe in people who are susceptible to known or suspected adverse effects of the active treatment. For example, myocardial infarction is a rare adverse effect of treatment with sildenafil (Viagra), so trials of this drug to treat painful vasospasm in patients with Raynaud's disease should exclude patients who have CHD (11). Conversely, being assigned to the inactive group or to placebo may be unsafe for some participants. For example, in women with vertebral fractures, bisphosphonates are known to reduce the risk of subsequent fractures, making it unacceptable to enter them in a placebo-controlled trial of a new treatment for osteoporosis unless bisphosphonates are provided for all participants. Persons in whom the active treatment is unlikely to be effective should be excluded, as well as those who are unlikely to be adherent to the intervention or unlikely to complete follow-up. Occasionally, practical problems such as impaired mental status that makes it difficult to follow instructions justify exclusion. Investigators should carefully weigh potential exclusion criteria that apply to many people (e.g., diabetes or upper age limits) as these may have a large impact on the feasibility and costs of recruitment and the generalizability of results.

Design an Adequate Sample Size and Plan the Recruitment Accordingly

Trials with too few participants to detect important effects are wasteful, unethical, and may produce misleading conclusions (12). Estimating the sample size is therefore one of the most important early parts of planning a trial (Chapter 6), and should take into account the fact that outcome rates in clinical trials are commonly lower than estimated due to healthy volunteer biases. In addition, recruitment for a trial is often more difficult than recruitment for an observational study because participants have to be willing to be randomized, often to a placebo or "experimental" drug. For these reasons, the investigator should plan for a generous sample from a large accessible population, and enough time and money to enroll the desired sample size when (as often happens) the barriers to doing so turn out to be greater than expected.

■ MEASURING BASELINE VARIABLES

To facilitate contacting participants who are lost to follow-up, it is important to record the names, phone numbers, addresses, and e-mail addresses of two or three friends or relatives who will always know how to reach the participant. If permissible, it is also valuable to record Social Security numbers or other national ID numbers. These can be used to determine the vital status of participants (through the National Death Index) or to detect key outcomes using health records (e.g., health insurance systems). However, this "protected health information" must be kept confidential and should not accompany data that are sent to a coordinating center or sponsoring institution.

Describe the Participants

Investigators should collect information on risk factors or potential risk factors for the outcome and on participant characteristics that may affect the efficacy of the intervention. These measurements also provide a means for checking on the comparability of the randomized study groups at baseline and provide information to assess the generalizability of the findings. The goal is to make sure that differences in baseline characteristics do not exceed what might be expected from the play of chance, suggesting a technical error or bias in carrying out the randomization. In small trials that are prone to sizeable maldistributions of baseline characteristics across randomized groups by chance alone, measurement of important predictors of the outcome permits statistical adjustment of the randomized comparison to reduce the influence of these chance maldistributions. Measuring predictors of the outcome also allows the investigator to examine whether the intervention has different effects in **subgroups** classified by baseline variables (**effect modification**, see Chapter 11).

Measure Baseline Value of the Outcome Variable

If outcomes include change in a variable, the outcome variable must be measured at the beginning of the study in the same way that it will be measured at the end. In studies that have a continuous outcome variable (effects of cognitive behavioral therapy on depression scores) the best measure is generally a change in the outcome over the course of the study. This approach usually minimizes the variability in the outcome between study participants and offers more power than simply comparing values at the end of the trial. In studies that have a dichotomous outcome (incidence of CHD, for example) it may be important to demonstrate by history and electrocardiogram that the disease is not present at the outset. It may also be useful to measure secondary outcome variables, and outcomes of planned ancillary studies, at baseline.

Be Parsimonious

Having pointed out multiple uses for baseline measurements, we should stress that the design of a clinical trial does not require that *any* be measured, because randomization minimizes the problem of confounding by factors that are present at the outset. Making a lot of measurements adds expense and complexity. In a randomized trial that has a limited budget, time and money are usually better spent on things that are vital to the integrity of the trial, such as the adequacy of the sample size, the success of randomization and blinding, and the completeness of adherence and follow-up. Yusuf et al. have promoted the use of large trials with very few measurements (13).

Bank Specimens

Storing images, **sera**, **DNA**, etc. at baseline will allow subsequent measurement of changes caused by the treatment, markers that predict the outcome, and factors such as genotype that might identify people who respond well or poorly to the treatment. Stored specimens can also be a rich resource to study other research questions not directly related to the main outcome.

RANDOMIZING AND BLINDING

The fourth step in Figure 10.1 is to randomly assign the participants to two groups. In the simplest design, one group receives an active treatment intervention and the other receives a placebo. **Random assignment** assures that age, sex, and other prognostic baseline characteristics that could confound an observed association (even those that are unknown or unmeasured) will be distributed equally, except for chance variation, among the randomized groups at baseline. **Blinding** is important to maintain comparability of the study groups during the trial and to assure unbiased outcome ascertainment.

Randomization

Because randomization is the cornerstone of a clinical trial, it is important that it be done correctly. The two most important features are that the procedure truly **allocates treatments randomly** and that the assignments are **tamperproof** so that neither intentional nor unintentional factors can influence the randomization.

It is important that the participant complete the baseline data collection, be found eligible for inclusion, and give consent to enter the study before randomization. He is then randomly assigned by computerized algorithm or by applying a set of random numbers. Once a list of the random order of assignment to study groups is generated, it must be applied to participants in strict sequence as they enter the trial.

It is essential to design the random assignment procedure so that members of the research team cannot influence the allocation. For example, for trials done at one site, random treatment

assignments can be placed in advance in a set of sealed envelopes by someone who will not be involved in opening the envelopes. Each envelope must be numbered (so that all can be accounted for at the end of the study), opaque (to prevent transillumination by a strong light), and otherwise tamperproof. When a participant is randomized, his name and the number of the next unopened envelope are first recorded *in the presence of a second staff member* and both staff sign the envelope; *then* the envelope is opened and the treatment group contained therein assigned to the participant and recorded on a log.

Multicenter trials typically use a separate tamperproof randomization facility that the trial staff contact when an eligible participant is ready to be randomized. The staff member provides the name and study ID of the new participant. This information is recorded and the treatment group is then randomly assigned based on a computer program that provides a treatment assignment number linked to the interventions. Treatment can also be randomly assigned by computer programs at a single research site as long as these programs are tamperproof. Rigorous precautions to prevent tampering with randomization are needed because investigators sometimes find themselves under pressure to influence the randomization process (e.g., for an individual who seems particularly suitable for an active treatment group in a placebo-controlled trial).

Consider Special Randomization Techniques

The preferred approach is typically simple randomization of individual participants to each intervention group. Trials of small to moderate size will have a small gain in power if special randomization procedures are used to balance the number of participants in each group (blocked randomization) and the distribution of baseline variables known to predict the outcome (stratified blocked randomization).

Blocked randomization is a commonly used technique to ensure that the number of participants is equally distributed among the study groups. Randomization is done in "blocks" of predetermined size. For example, if the block size is six, randomization proceeds normally within each block of six until three persons are randomized to one of the groups, after which participants are automatically assigned to the other group until the block of six is completed. This means that in a study of 30 participants exactly 15 will be assigned to each group, and in a study of 33 participants, the disproportion could be no greater than 18:15. Blocked randomization with a fixed block size is less suitable for nonblinded studies because the treatment assignment of the participants at the end of each block could be anticipated and manipulated. This problem can be minimized by varying the size of the blocks randomly (ranging, for example, from blocks of four to eight) according to a schedule that is not known to the investigator.

Stratified blocked randomization ensures that an important predictor of the outcome is more evenly distributed between the study groups than chance alone would dictate. In a trial of the effect of a drug to prevent fractures, having a prior vertebral fracture is such a strong predictor of outcome that it may be best to ensure that similar numbers of people who have vertebral fractures are assigned to each group. This can be achieved by carrying out blocked randomization separately by "strata"—those with and without vertebral fractures. Stratified blocked randomization can slightly enhance the power of a small trial by reducing the variation in outcome due to chance disproportions in important baseline predictors. It is of little benefit in large trials (more than 1,000 participants) because random assignment ensures nearly even distribution of baseline variables.

An important limitation of stratified blocked randomization is the small number of baseline variables, not more than two or three, that can be balanced by this technique. A technique for addressing this limitation is **adaptive randomization**, which uses a "biased coin" to alter the probability of assigning each new participant so that, for example, someone with a high risk score based on any number of baseline prognostic variables would be slightly more likely to join the study group that is at lower overall risk based on all participants randomized to that point. Disadvantages of this technique include the difficulty of explaining the likelihood of

assignment to study groups to potential participants during informed consent and the complexity of implementation, with an interactive computerized system that recomputes the biased coin probabilities with each randomization.

Usually, the best decision is to assign equal numbers to each study group, as this maximizes power for any given total sample size., However,the attenuation in power of even a 2:1 disproportion is quite modest (14)), and **unequal allocation** of participants to treatment and control groups may sometimes be appropriate (15):

- Increasing the ratio of active to control treatment can make the trial more **attractive** to potential participants, such as those with HIV infection who would like the greater chance of receiving active treatment if they enroll;
- Decreasing the ratio of active to control participants can make the trial **affordable** when the intervention is very expensive (as in the Women's Health Initiative low-fat diet trial (16)).
- Increasing the proportion assigned to the group serving as a control for several active treatment groups will **increase the power** of each comparison by increasing the precision of the control group estimate (as in the Coronary Drug Project trial (17)).

Randomization of matched pairs is a strategy for balancing baseline confounding variables that requires selecting pairs of participants who are matched on important characteristics like age and sex, then randomly assigning one member of each pair to each study group. A drawback of randomizing matched pairs is that it complicates recruitment and randomization, requiring that an eligible participant wait for randomization until a suitable match has been identified. In addition, matching is generally not necessary in large trials in which random assignment balances the groups on baseline variables. However, an attractive version of this design can be used when the circumstances permit a contrast of treatment and control effects in two parts of the same individual. In the Diabetic Retinopathy Study, for example, each participant had one eye randomly assigned to photocoagulation treatment while the other served as a control (18).

Blinding

Whenever possible, the investigator should design the interventions in such a fashion that the study participants, staff who have contact with them, persons making measurements, and those who ascertain and adjudicate outcomes do not know the study group assignment. When it is not possible to blind all of these individuals, it is highly desirable to blind as many as possible (always, for example, blinding personnel making outcome measurements). In a randomized trial, **blinding** *is as important as* **randomization**. Randomization minimizes the influence of confounding variables at the time of randomization, but it has no impact on differences that develop between the groups during follow-up (Table 10.2). Blinding minimizes post-randomization sources of bias, such as co-interventions and biased outcome ascertainment and adjudication.

The use of blinding to prevent bias caused by **co-interventions**—medications, therapies, or behaviors other than the study intervention that change the risk of developing the outcome of interest—has been discussed (p. 139). The second important purpose of blinding is to **minimize biased ascertainment and adjudication of outcomes**. In an unblinded trial, the investigator may be tempted to look more carefully for outcomes in the untreated group or to diagnose the outcome more frequently. For example, in an unblinded trial of statin therapy, the investigators may be more likely to ask participants in the active treatment group about muscle pain or tenderness and to order tests to make the diagnosis of myositis. Blinding of subjects is particularly important when outcomes are based on self-reported symptoms.

After a possible outcome event has been ascertained, it may require adjudication. For example, if the outcome of the trial is myocardial infarction, investigators typically collect data on symptoms, EKG findings, and cardiac enzymes. Experts blinded to treatment group then use these data and specific definitions to adjudicate whether or not a myocardial infarction has

TABLE 10.2 IN A RANDOMIZED BLINDED TRIAL, RANDOMIZATION MINIMIZES CONFOUNDING AT BASELINE AND BLINDING MINIMIZES CO-INTERVENTIONS AND BIASED OUTCOME ASCERTAINMENT AND ADJUDICATION

EXPLANATION FOR ASSOCIATION		STRATEGY TO RULE OUT RIVAL EXPLANATION
1. Chance		Same as in observational studies (Table 9.2)
2. Bias		Same as in observational studies (Table 9.2)
3. Effect—Cause		(Not a possible explanation in a trial)
4. Confounding	Prerandomization confounding variables	**Randomization**
	Postrandomization confounding variables (co-interventions)	**Blinding**
5. Cause—Effect		

occurred. Results of the Canadian Cooperative Multiple Sclerosis trial illustrate the importance of blinding for unbiased outcome adjudication (19). Persons with multiple sclerosis were randomly assigned to combined plasma exchange, cyclophosphamide and prednisone, or to sham plasma exchange and placebo medications. At the end of the trial, the severity of multiple sclerosis was assessed using a structured examination by neurologists blinded to treatment assignment and again by neurologists who were unblinded. Therapy was not effective based on the assessment of the blinded neurologists, but was statistically significantly effective based on the assessment of the unblinded neurologists. The unblinded neurologists were not purposefully trying to bias the outcome of the trial, but there is a strong human desire to see patients improve after treatment, especially if the treatment is painful or potentially harmful. Blinding minimizes such biased outcome adjudication.

Blinded assessment of outcome may be less important if the outcome of the trial is a "hard" outcome such as death, or automated measurements about which there is little or no opportunity for biased assessment. Most other outcomes, such as cause of death, disease diagnosis, physical measurements, questionnaire scales, and self-reported conditions, are susceptible to biased ascertainment and adjudication.

After a trial is over, it may be a good idea to assess whether the participants and investigators were unblinded by asking them to guess which treatment the participant was assigned to. If a higher-than-expected proportion guesses correctly, the published discussion of the findings should include an assessment of the potential biases that partial unblinding may have caused.

What to Do When Blinding Is Impossible

In some cases blinding is difficult or impossible, either for technical or ethical reasons. For example, it is difficult to blind participants if they are assigned to an educational, dietary, or exercise intervention. Surgical interventions often cannot be blinded because it may be unethical to perform sham surgery in the control group. However, surgery is always associated with some risk, so it is very important to determine if the procedure is truly effective. For example, a recent randomized trial found that arthroscopic debridement of the cartilage of the knee was no more effective than arthroscopy with sham debridement for relieving osteoarthritic knee pain (20). In this case, the risk to participants in the control group might be outweighed if the results of the trial prevented thousands of patients from undergoing an ineffective procedure.

If the intervention cannot be blinded, the investigator should at least limit potential co-interventions as much as possible, and assure that individuals who ascertain and adjudicate the outcomes are blinded. For example, an investigator testing the effect of yoga for relief of menopausal hot flashes could instruct both yoga and control participants to refrain from starting new medications, relaxation activities, or other treatments for hot flushes until the trial has ended. Also, study staff who collect information on the severity of hot flashes could be different from those who provide the yoga training.

SUMMARY

1. A **randomized blinded trial**, properly designed and carried out, can provide the most **definitive causal inference** as a basis for practice guidelines based on **evidence-based medicine**.
2. The **choice and dose of intervention** is a difficult decision that balances judgments about **efficacy** and **safety**; other considerations include **relevance** to clinical practice, suitability for **blinding**, and whether to use a **combination of drugs**.
3. When possible, the **comparison group** should be a **placebo control** that allows participants, investigators, and study staff to be **blinded**.
4. **Clinically relevant outcomes** such as pain, quality of life, occurrence of cancer, and death are the most meaningful outcomes of trials. **Intermediary outcomes**, such as HIV viral load, are valid **surrogate markers** for clinical outcomes to the degree that treatment-induced changes in the marker predict changes in the clinical outcome.
5. Measuring **more than one outcome** variable is usually helpful, but combining them into **composite outcomes** requires careful consideration; a **single primary outcome** should be specified to test the main hypothesis.
6. All clinical trials should include measures of potential **adverse effects** of the intervention, both **targeted** and (in moderation) **open-ended** measures with procedures to assure that **serious adverse events (SAEs)** are promptly reported to IRBs and sponsors.
7. The criteria for **selecting study participants** should identify those who are likely to experience the **most benefit and the least harm** from treatment, and to **adhere to treatment** and **follow-up** protocols. Choosing participants at **high risk** of the outcome can **decrease sample size**, but may make recruitment more difficult and decrease the generalizability of the findings.
8. **Baseline variables** should be measured parsimoniously to **describe** participant characteristics, measure **risk factors** for and **baseline values of the outcome**, and enable later examination of disparate intervention effects in various subgroups (**effect modification**). Consider **storing baseline** serum, genetic material, images, etc. for later analyses.
9. **Randomization**, which minimizes the influence of baseline **confounding** variables, should be tamperproof; **matched pair** randomization is an excellent design when feasible, and in small trials **stratified blocked randomization** can reduce chance maldistributions of key predictors.
10. **Blinding** the intervention is **as important as randomization** and serves to control **co-interventions** and biased **outcome ascertainment** and **adjudication**.

REFERENCES

1. Ettinger B, Black DM, Mitlak BH, et al. Reduction of vertebral fracture risk in postmenopausal women with osteoporosis treated with raloxifene: results from a 3-year randomized clinical trial. Multiple Outcomes of Raloxifene Evaluation (MORE) investigators. *JAMA* 1999;282:637–645.
2. The Women's Health Initiative Study Group. Design of the women's health initiative clinical trial and observational study. *Control Clin Trials* 1998;19:61–109.
3. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 1989;8:431–440.
4. Cummings SR, Karpf DB, Harris F, et al. Improvement in spine bone density and reduction in risk of vertebral fractures during treatment with antiresorptive drugs. *Am J Med* 2002;112:281–289.

5. Hulley S, Grady D, Bush T, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA* 1998;280:605–613.
6. Cummings SR, San Martin J, McClung MR, et al. Denosumab for prevention of fractures in postmenopausal women with osteoporosis. *N Engl J Med* 2009;361(8):756–765.
7. Mosca L, Barrett-Connor E, Wenger NK, et al. Design and methods of the Raloxifene Use for The Heart (RUTH) Study. *Am J Cardiol* 2001;88:392–395.
8. Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. *JAMA* 2002;288:321–333.
9. McClung M, Boonen S, Torring O, et al. Effect of denosumab treatment on the risk of fractures in subgroups of women with postmenopausal osteoporosis. *J Bone Miner Res* 2011;27:211–218.
10. Cummings SR, Black DM, Thompson DE, et al. Effect of alendronate on risk of fracture in women with low bone density but without vertebral fractures: results from the fracture intervention trial. *JAMA* 1998;280:2077–2082.
11. Fries R, Shariat K, von Wilmowsky H, et al. Sildenafil in the treatment of Raynaud's phenomenon resistant to vasodilatory therapy. *Circulation* 2005;112:2980–2985.
12. Freiman JA, Chalmers TC, Smith H Jr, et al. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. *N Engl J Med* 1978;299:690–694.
13. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984;3:409–420.
14. Friedman LM, Furberg C, DeMets DL. *Fundamentals of clinical trials*, 4th ed. New York: Springer, 2010.
15. Avins AL. Can unequal be more fair? Ethics, subject allocation, and randomised clinical trials. *J Med Ethics* 1998;24:401–408.
16. Prentice RL, Caan B, Chlebowski RT, et al. Low-fat dietary pattern and risk of invasive breast cancer: the women's health initiative randomized controlled dietary modification trial. *JAMA* 2006;295:629–642.
17. CDP Research Group. The coronary drug project. Initial findings leading to modifications of its research protocol. *JAMA* 1970;214:1303–1313.
18. Diabetic Retinopathy Study Research Group. Preliminary report on effects of photocoagulation therapy. *Am J Ophthalmol* 1976;81:383–396.
19. Noseworthy JH, O'Brien P, Erickson BJ, et al. The Mayo-Clinic Canadian cooperative trial of sulfasalazine in active multiple sclerosis. *Neurology* 1998;51:1342–1352.
20. Moseley JB, O'Malley K, Petersen NJ, et al. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *N Engl J Med* 2002;347:81–88.

CHAPTER 11

Alternative Clinical Trial Designs and Implementation Issues

Deborah Grady, Steven R. Cummings, and Stephen B. Hulley

In the last chapter, we discussed the classic randomized, blinded, parallel group trial: how to select and blind the intervention and control conditions, randomly assign the interventions, choose outcomes, deal with adverse events, select participants, and measure baseline and outcome variables.

In this chapter, we describe alternative **randomized and non-randomized** between-group **trial designs**, as well as **within-group** designs, **cross-over** studies, and **pilot studies**. We then address the **conduct of clinical trials**, including **adherence to the intervention and follow-up**, and **ascertaining and adjudicating outcomes**. We conclude with a discussion of statistical issues such as **interim monitoring** for stopping the trial early, **intention to treat** and **per-protocol** analyses, and the use of **subgroup analysis** to discover effect modification.

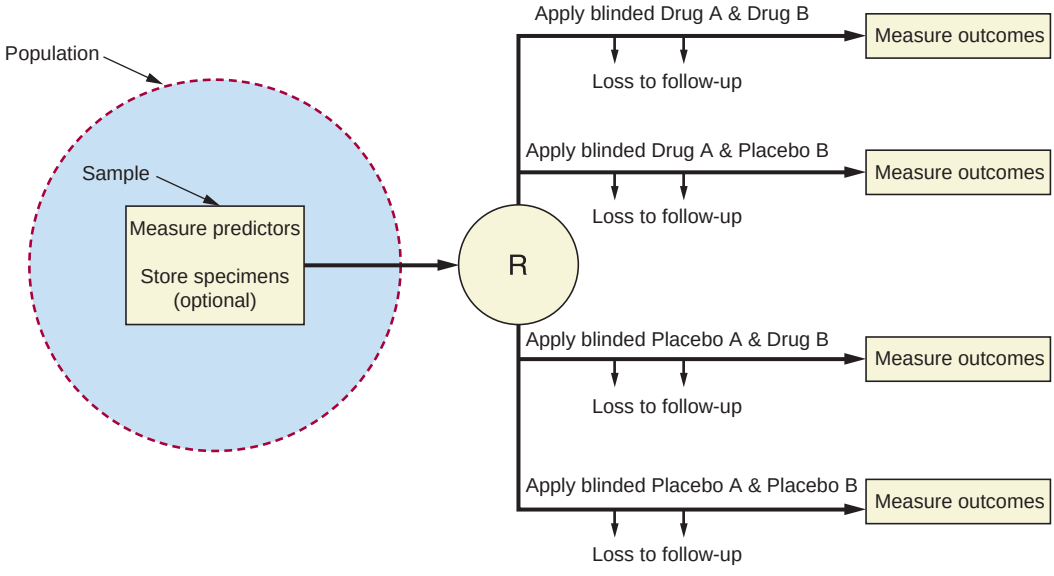
■ ALTERNATIVE RANDOMIZED DESIGNS

There are a number of variations on the classic parallel group randomized trial that may be useful when the circumstances are right.

Factorial Design

The **factorial design** aims to answer two (or more) separate research questions in a single trial (Figure 11.1). A good example is the Women's Health Study, which was designed to test the effect of low-dose aspirin and of vitamin E on the risk for cardiovascular events among healthy women (1). The participants were randomly assigned to four groups, and two hypotheses were tested by comparing two halves of the study cohort. First, the rate of cardiovascular events in women on aspirin was compared with women on aspirin placebo (disregarding the fact that half of each of these groups received vitamin E); then the rate of cardiovascular events in those on vitamin E was compared with all those on vitamin E placebo (now disregarding the fact that half of each of these groups received aspirin). The investigators have two complete trials for the price of one.

A limitation is the possibility of **effect modification** (interaction): if the effect of aspirin on risk for cardiovascular disease is different in women treated with vitamin E than in those not treated with vitamin E, effect modification is present and the effect of aspirin would have to be calculated separately in these two groups. This would reduce the power of these comparisons, because only half of the participants would be included in each analysis. Factorial designs can actually be used to *study* effect modification, but trials designed for this purpose are more complicated and difficult to implement, larger sample sizes are required, and the results can be hard to interpret. Other limitations of the factorial design are that the same study population must be appropriate for each intervention, multiple treatments may interfere with recruitment and adherence, and analyses are more complex. That said, the factorial design can be very **efficient**. For example, the Women's Health Initiative randomized trial was able to test the effect of three



■ **FIGURE 11.1** In a factorial randomized trial, the steps are to:

- Select a sample of participants from a population suitable for receiving the intervention.
- Measure the predictor variables and (if appropriate) the baseline level of the outcome variable.
- Consider the option of storing serum, images, and so on, for later analysis.
- Randomly assign two (or more) active interventions and their controls to four (or more) groups.
- Follow the cohorts over time, minimizing loss to follow-up and assessing adherence to the intervention and control conditions.
- Measure the outcome variables.
- Analyze the results, first comparing the two intervention A groups (combined) to the combined placebo A groups and then comparing the two intervention B groups (combined) to the combined placebo B groups.

interventions (postmenopausal hormone therapy, low-fat diet, and calcium plus vitamin D) on a number of outcomes (2).

Cluster Randomization

Cluster randomization requires that the investigator randomly assign naturally occurring groups or clusters of participants to the interventions, rather than individuals. A good example is a trial that enrolled players on 120 college baseball teams, randomly allocated half of the teams to an intervention to encourage cessation of spit-tobacco use, and observed a significantly lower rate of spit-tobacco use among players on the teams that received the intervention compared to control teams (3). Applying the intervention to groups of people may be more feasible and cost effective than treating individuals one at a time, and it may better address research questions about the effects of public health programs in the population. Some interventions, such as a low-fat diet, are difficult to implement in only one member of a family. When participants in a natural group are randomized individually, those who receive the intervention are likely to discuss or share the intervention with family members, colleagues, team members, or acquaintances who have been assigned to the control group.

In the cluster randomization design, the units of randomization and analysis are groups, not individuals. Therefore, the effective sample size is smaller than the number of individual participants and power is diminished. The effective sample size depends on the correlation of the effect of the intervention among participants in the clusters and is somewhere between the

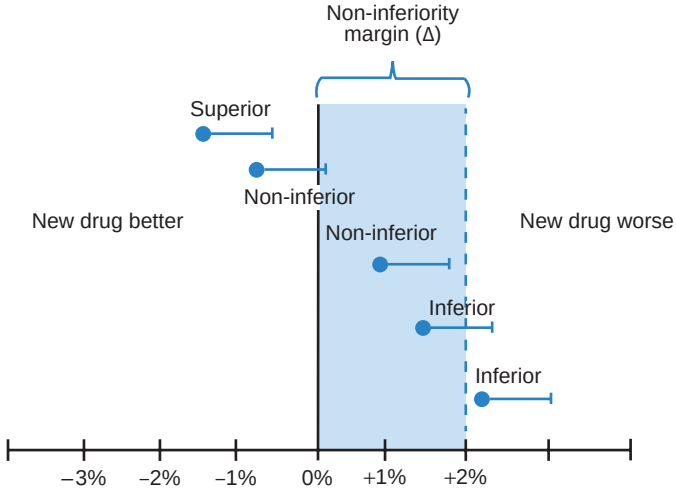
number of clusters and the number of participants (4). Other drawbacks are that sample size estimation and data analysis are more complicated in cluster randomization designs than for individual randomization (4).

Active Control Trials: Equivalence and Non-Inferiority

An **active control trial** is one in which the control group receives an active treatment. This design may be optimal when there is a known effective treatment or "standard of care" for a condition. This type of trial is sometimes called a **comparative effectiveness trial** because two treatments are compared.

In some cases, the aim of an active control trial is to show that a new treatment is **superior** to an established treatment. In this situation, the design and methods are similar to a placebo-controlled trial. In most cases, however, investigators want to establish that a new therapy that has some advantages over an established therapy (easier to use, less invasive, safer) has *similar* efficacy. In this case, an **equivalence** or **non-inferiority** trial is more appropriate.

The **statistical methods** for equivalence or non-inferiority trials are different than for trials designed to show that one treatment is better than another. In a trial designed to show that a treatment is superior, the standard analysis uses tests of statistical significance to accept or reject the null hypothesis that there is no difference between groups. In a trial designed to show that a new treatment is equivalent to the standard treatment, on the other hand, the ideal goal would be to *accept* the null hypothesis of no difference. But proving that there is *no* difference between treatments (not even a tiny one) would require an infinite sample size. So the practical solution is to design the sample size and analysis plan using a confidence interval (CI) approach—considering where the CI for the effect of the new treatment compared to the standard treatment lies with respect to a prespecified delta ("Δ"), the unacceptable difference in efficacy between the two treatments (5, 6). Equivalence or non-inferiority is considered established at the level of significance specified by the CI if the CI around the difference in efficacy of the new compared to the established treatment does not include Δ (Figure 11.2). This is a



■ **FIGURE 11.2** Possible outcomes in a non-inferiority trial comparing a new drug to warfarin as treatment to reduce stroke risk among patients with atrial fibrillation, with the non-inferiority margin (delta) set at +2%. The one-sided 95% confidence intervals around the difference in stroke rate between warfarin and the new drug are shown illustrating the outcomes of superiority, inferiority, and non-inferiority.

two-tailed consideration in the case of an equivalence trial (i.e., the new treatment is neither worse *nor* better than the standard treatment). However, it is uncommon for investigators to be interested in whether a new treatment is *both* no better *and* no worse than an established treatment. Most often, investigators are especially interested in showing that a new treatment with other advantages is not inferior to the standard treatment. The one-tailed nature of the non-inferiority trial design also has the advantage of permitting either a smaller sample size or a smaller alpha; the latter is usually preferred (e.g., 0.025 rather than 0.05), to be conservative.

One of the most difficult issues in designing a non-inferiority trial is establishing the **non-inferiority margin (Δ)**—the loss of efficacy of the new treatment that would be unacceptable (7). This decision is based on both statistical and clinical considerations of the potential efficacy and advantages of the new treatment, and requires expert judgment (8) (see Appendix 11A for an example of how this is done). Non-inferiority trials generally need to be larger than placebo-controlled trials because the acceptable difference between the new and established treatment is usually smaller than the expected difference between a new treatment and placebo.

It is important to note that non-inferiority may not mean that both the established and new treatments are effective—they could be equivalently ineffective or harmful. To ensure that a new treatment evaluated in a non-inferiority trial is more effective than placebo, there should be strong prior evidence supporting the efficacy of the established treatment. This also means that the design of the non-inferiority trial should be as similar as possible to trials that have established the efficacy of the standard treatment, including selection criteria, dose of the established treatment, adherence to the standard treatment, length of follow-up, loss to follow-up, and so on (6, 7). Any problem that reduces the efficacy of the standard treatment (enrolling participants unlikely to benefit, non-adherence to treatment, loss to follow-up) will make it more likely that the new therapy will be found to be non-inferior—simply because the efficacy of the standard treatment has been reduced. A new, less effective treatment may appear to be non-inferior when, in reality, the findings represent a poorly done study.

In summary, non-inferiority and equivalence trials are particularly worthwhile if a new treatment has important advantages such as lower cost, ease of use, or safety. It is difficult to justify large trials to test a new "me-too" drug with none of these advantages. Importantly, non-inferiority and equivalence trials can produce the misleading conclusion that two treatments are equivalent if the trial is poorly conducted.

Adaptive Designs

Clinical trials are generally conducted according to a protocol that does not change during the conduct of the study. However, for some types of treatments and conditions, it is possible to monitor results from the trial as it progresses and **change the design** of the trial **based on interim analyses** of the results (9). For example, consider a trial of several doses of a new treatment for non-ulcer dyspepsia. The initial design may plan to enroll 50 participants to a placebo group and 50 to each of three doses for 12 weeks of treatment over an enrollment period lasting 1 year. Review of the results after the first 10 participants in each group have completed 4 weeks of treatment might reveal that there is a trend toward relief of dyspepsia only in the highest dose group. It may be more efficient to stop assigning participants to the two lower doses and continue randomizing only to the highest dose and the placebo. Other facets of a trial that could be changed based on interim results include increasing or decreasing the **sample size** or **duration** of the trial if interim results indicate that the effect size or rate of outcomes differ from the original assumptions.

Adaptive designs are feasible only for treatments that produce outcomes that are measured and analyzed early enough in the course of the trial to make design changes in the later stages of the trial possible. To prevent bias, rules for how the design may be changed should be established before the trial begins, and the interim analyses and consideration of change in design should be done by an independent data and safety monitoring board that reviews unblinded

data. Multiple interim analyses will increase the probability of finding a favorable result that is due to chance variation, and the increased chance of a type I error must be considered in the analysis of the results.

In addition to being more complex to conduct and analyze, adaptive designs require that informed consent include the range of possible changes in the study design, and it is difficult to estimate the cost of an adaptive trial and the specific resources needed to complete it. Despite these precautions and limitations, adaptive designs are efficient and may be valuable, especially during the development of a new treatment; they can allow earlier identification of the best dose and duration of treatment, and ensure that a high proportion of participants receive the optimal treatment.

NONRANDOMIZED DESIGNS

Nonrandomized Between-Group Designs

Trials that compare groups that have not been randomized are far less effective than randomized trials in controlling for confounding variables. For example in a trial of the effects of coronary artery bypass surgery compared to percutaneous angioplasty, if clinicians are allowed to decide which patients undergo the procedures rather than using random allocation, patients chosen for surgery are likely to be different than those chosen for angioplasty. Analytic methods can adjust for baseline factors that are unequal in the two study groups, but this strategy does not deal with the problem of unmeasured confounding. When the findings of randomized and nonrandomized studies of the same research question are compared, the apparent benefits of intervention are often greater in the nonrandomized studies, even after adjusting statistically for differences in baseline variables (10). The problem of confounding in nonrandomized clinical studies can be serious and not fully removed by statistical adjustment (11).

Sometimes participants are allocated to study groups by a **pseudorandom** mechanism. For example, every participant with an even hospital record number may be assigned to the treatment group. Such designs may offer logistic advantages, but the predictability of the study group assignment permits the investigator or the study staff to tamper with it by manipulating the sequence or eligibility of new participants.

Participants are sometimes assigned to study groups by the investigator according to certain specific criteria. For example, patients with diabetes may be allocated to receive either insulin four times a day or long-acting insulin once a day according to their willingness to accept four daily injections. The problem with this design is that those willing to take four injections per day might differ from those who are unwilling (for example, being more compliant with other health advice), and this might be the cause of any observed difference in the outcomes of the two treatment programs.

Nonrandomized designs are sometimes chosen in the mistaken belief that they are more ethical than randomization because they allow the participant or clinician to choose the intervention. In fact, studies are only ethical if they have a reasonable likelihood of producing the correct answer to the research question, and randomized studies are more likely to lead to a conclusive and correct result than nonrandomized designs. Moreover, the ethical basis for any trial is the uncertainty as to whether the intervention will be beneficial or harmful. This uncertainty, termed **equipoise**, means that an evidence-based choice of interventions is not possible and this justifies random assignment.

Within-Group Designs

Designs that do not include a separate control group can be useful options for some types of questions. In a **time series design**, measurements are made before and after each participant receives the intervention (Figure 11.3). Therefore, each participant serves as his own control to evaluate the effect of treatment. This means that individual characteristics such as age, sex,

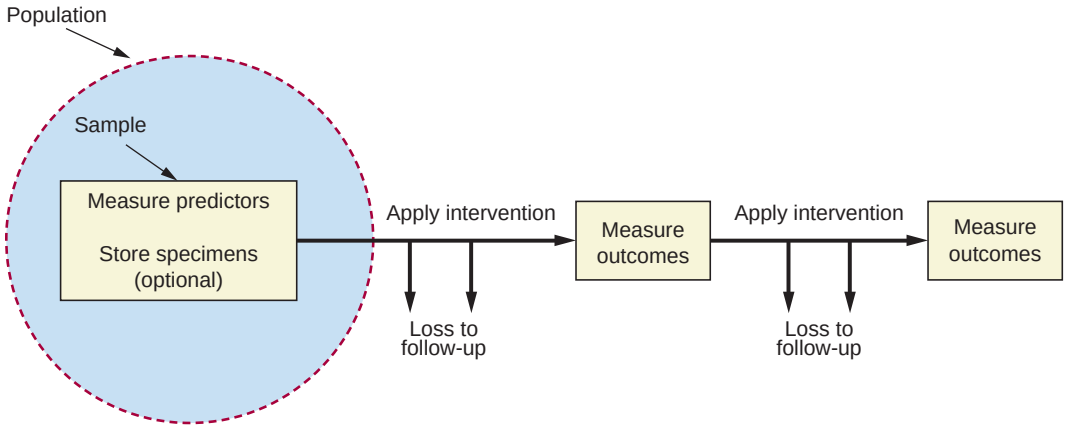


FIGURE 11.3 In a time series trial, the steps are to:

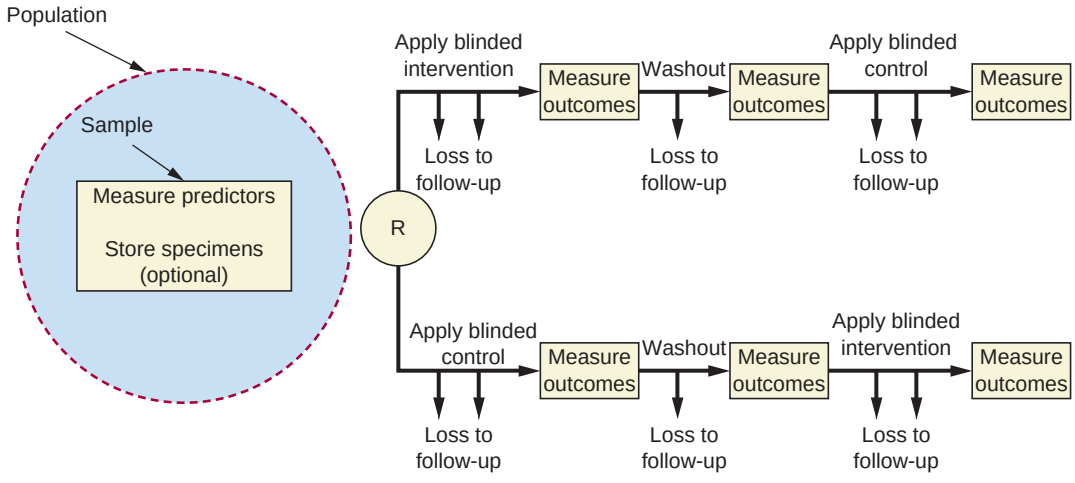
- Select a sample of participants from a population suitable for receiving the intervention.
- Measure the predictor variables and (if appropriate) the baseline level of the outcome variable.
- Consider the option of storing serum, images, and so on for later analysis.
- Apply the intervention to the whole cohort.
- Follow the cohort over time, minimizing loss to follow-up and assessing adherence to the intervention.
- Measure the outcome variables.
- Remove the intervention, continue the follow-up and measure the outcome variable again, then re-initiate the intervention, and so on.

and genetic factors are not merely balanced (as they are in between-group studies), but actually eliminated as confounding variables.

The major disadvantage of within-group designs is the lack of a *concurrent* control group. The apparent efficacy of the intervention might be due to **learning effects** (participants do better on follow-up cognitive function tests because they learned from the baseline test), **regression to the mean** (participants who were selected for the trial because they had high blood pressure at baseline are found to have lower blood pressure at follow-up simply due to random variation in blood pressure), or **secular trends** (upper respiratory infections are less frequent at follow-up because the flu season ended before follow-up was completed). Within-group designs sometimes use a strategy of repeatedly starting and stopping the treatment. If repeated onset and offset of the intervention produces corresponding patterns in the outcome, this is strong support that these changes are due to the treatment. This approach is only useful when the outcome variable responds rapidly and reversibly to the intervention. The design has a clinical application in "**N-of-one**" **trials** in which an individual patient can alternate between active and inactive versions of a drug (using identical-appearing placebo prepared by the local pharmacy) to detect his particular response to the treatment (12).

Crossover Designs

The **crossover design** has features of both within- and between-group designs (Figure 11.4). Half of the participants are randomly assigned to start with the control period and then switch to active treatment; the other half begins with the active treatment and then switches to control. This approach permits between-group, as well as within-group, analyses. The advantages are substantial: it minimizes the potential for confounding because each participant serves as his own control and the paired analysis increases the statistical power of the trial so that it needs fewer participants. However, the disadvantages are also substantial: a doubling of the duration of the study, the added expense required to measure the outcome at the beginning and end of each crossover period, and the added complexity of analysis and interpretation created by



■ **FIGURE 11.4** In a crossover randomized trial, the steps are to:

- Select a sample of participants from a population suitable for receiving the intervention.
- Measure the predictor variables and (if appropriate) the baseline level of the outcome variable.
- Randomly assign the blinded intervention and control condition.
- Follow the cohorts over time, minimizing loss to follow-up and assessing compliance with the interventions and control conditions.
- Measure the outcome variables.
- Discontinue the intervention and control condition and provide a washout period to reduce carryover effect, if appropriate.
- Apply intervention to former control group and control condition to former intervention group and measure outcomes after following cohorts over time.

potential **carryover effects**. A carryover effect is the residual influence of the intervention on the outcome during the period after it has been stopped—blood pressure not returning to baseline levels for months after a course of diuretic treatment, for example. To reduce the carryover effect, the investigator can introduce an untreated "**washout**" period between treatments with the hope that the outcome variable will return to baseline before starting the next intervention, but it is difficult to know whether all carryover effects have been eliminated. In general, crossover studies are a good choice when the number of study participants is limited and the outcome responds rapidly and reversibly to an intervention.

A variation on the crossover design may be appropriate when the intervention to be studied cannot be blinded and the intervention is believed by participants to be much more desirable than the control (such as a new noninvasive procedure). In this situation, where it may be very difficult to find eligible participants who are willing to be randomized, an excellent approach may be randomization to immediate intervention versus a **wait-list** (delayed) **control**. Another situation in which a wait-list control may be appropriate is when a community, school, government, or similar entity has decided that all members of a group should receive an intervention, despite limited evidence of efficacy. In this situation, randomization to not receive the intervention may be considered unethical, while randomization to delayed intervention may be acceptable.

The wait-list design provides an opportunity for a randomized comparison between the immediate intervention and wait-list control groups. In addition, the two intervention periods (immediate intervention in one group and delayed intervention in the other) can be pooled to increase power for a within-group comparison before and after the intervention. For example, in a trial in which women with symptomatic fibroids are randomized to a new treatment that is less invasive than hysterectomy (uterine artery embolization) versus wait-list,

the wait-list control would receive no treatment during the initial period; then be offered uterine artery embolization at the beginning of the next period. Subsequently, within-group measurements of changes in fibroid symptom score can be pooled among all of the participants who received the intervention.

This design has the advantage of making enrollment much more feasible in a trial where the intervention is highly desirable, and of allowing a randomized comparison in situations where all eligible participants will eventually receive an intervention. However, the outcome must occur in a short period of time (or the wait period becomes prohibitively long). In addition, providing the intervention to the control group at the end of the trial prolongs the length of follow-up and can be expensive.

Trials for Regulatory Approval of New Interventions

Many trials are done to test the effectiveness and safety of new treatments that might be considered for approval for marketing by the U.S. Food and Drug Administration (FDA) or another regulatory body. Trials are also done to determine whether drugs that have FDA approval for one condition might be approved for the treatment or prevention of other conditions. The design and conduct of these trials is generally the same as for other trials, but regulatory requirements must be considered.

The FDA publishes general and specific guidelines on how such trials should be conducted (search for "FDA" on the Web). It would be wise for investigators and staff conducting trials with the goal of obtaining FDA approval of a new medication or device to seek specific training on general guidelines, called **Good Clinical Practice** (Chapter 17). In addition, the FDA provides specific guidelines for studies of certain outcomes. For example, studies designed to obtain FDA approval of treatments for hot flashes in menopausal women must currently include participants with at least seven hot flashes per day or 50 per week. FDA guidelines are regularly updated and similar guidelines are available from international regulatory agencies.

Trials for regulatory approval of new treatments are generally described by phase. This system refers to an orderly progression in the testing of a new treatment, from experiments in animals, human cell cultures or tissues (**preclinical**) and initial unblinded, uncontrolled treatment of a few human volunteers to test safety (**phase I**), to small randomized or time series trials that test the effect of a range of doses on adverse effects and biomarkers or clinical outcomes (**phase II**), to randomized trials large enough to test the hypothesis that the treatment improves the targeted condition (such as blood pressure) or reduces the risk of disease (such as stroke) with acceptable safety (**phase III**) (Table 11.1). The FDA usually defines the endpoints for phase III trials that are required to obtain approval to market the new drug. **Phase IV** refers to large studies that may be randomized trials, but are often large observational studies that are conducted after a drug is approved. These studies are often designed to assess the rate of serious

TABLE 11.1 STAGES IN TESTING NEW THERAPIES

Preclinical	Studies in cell cultures, tissues, and animals
Phase I	Unblinded, uncontrolled studies in a few volunteers to test safety
Phase II	Relatively small randomized or time series trials to test tolerability and different intensity or dose of the intervention on biomarkers or clinical outcomes
Phase III	Relatively large randomized blinded trials to test conclusively the effect of the therapy on clinical outcomes and adverse events
Phase IV	Large trials or observational studies conducted after the therapy has been approved by the FDA to assess the rate of uncommon serious side effects and evaluate additional therapeutic uses

side effects when the drug is used in large populations or to test additional uses of the drug that might be approved by the FDA. Sometimes, phase IV studies do not have a clear scientific goal, but are performed to introduce physicians and patients to new drugs.

Pilot Studies

Designing and conducting a successful clinical trial requires extensive information on the type, dose, and duration of the intervention; the likely effect of the intervention on the outcome; potential adverse effects; the feasibility of recruiting, randomizing, and maintaining participants in the trial; and likely costs. Often, the only way to obtain some of this information is to conduct a good pilot study.

Pilot studies vary from a brief test of feasibility in a small number of participants to a long trial in hundreds of participants (in preparation for a major multicenter multi-year investment). Pilot studies should be as carefully planned as the main trial, with clear objectives and methods. Many pilot studies are focused primarily on determining the **feasibility**, **time required**, and **cost** of recruiting adequate numbers of eligible participants, and discovering if they are willing to accept randomization and can comply with the intervention. Pilot studies may also be designed to demonstrate that planned **measurements**, data collection **instruments**, and **data management** systems are feasible and efficient. For pilot studies done primarily to test feasibility, a control group is generally not included.

An important goal of many pilot studies is to define the optimal **intervention**—the frequency, intensity, and duration of the intervention that will result in minimal toxicity and maximal effectiveness.

Pilot studies are sometimes used to provide estimates of parameters needed to estimate **sample size**. Sound estimates of the rate of the outcome or mean outcome measure in the placebo group, the effect of the intervention on the main outcome (**effect size**), and the statistical **variability** of this outcome are crucial to planning the sample size. In most cases, it's best to obtain these estimates from published full-scale studies of similar interventions in similar participants. In the absence of such data, using estimates from a pilot study may be helpful, but the sample size for pilot studies is usually so small that the calculated effect size and variance are unstable, with very wide confidence intervals.

Many trials fall short of estimated power not because the effect of the intervention is less than anticipated, but because the rate of dichotomous **outcome events** in the placebo group is much lower than expected. This likely occurs because persons who fit the enrollment criteria for a clinical trial and agree to be randomized are healthier than the general population with the condition of interest. Therefore, it is crucial to determine the rate of the outcome in the placebo group, which may be done by evaluating the placebo group of prior trials with similar participants, or by randomizing participants to placebo in a pilot study.

A pilot study should have a short but **complete protocol** (approved by the institutional review board), data collection forms, and analysis plans. Variables should include the typical baseline measures, predictors, and outcomes included in a full-scale trial, but also estimates of the number of participants available or accessible for recruitment, the number who are contacted or respond using different sources or recruitment techniques, the number and proportion eligible for the trial, those who are eligible but refuse (or say they would refuse) randomization, the time and cost of recruitment and randomization, and estimates of adherence to the intervention and other aspects of the protocol, including study visits. It is usually helpful to "debrief" both participants and staff after the pilot study to obtain their views on how the trial methods could be improved.

A good pilot study requires substantial time and can be costly, but markedly improves the chance of funding for a major clinical trial and the likelihood that the trial will be successfully completed.

CONDUCTING A CLINICAL TRIAL

Follow-Up and Adherence to the Protocol

If a substantial number of study participants do not receive the study intervention, do not adhere to the protocol, or are lost to follow-up, the results of the trial can be underpowered or biased. Strategies for **maximizing follow-up and adherence** are outlined in Table 11.2.

The effect of the intervention (and the power of the trial) is reduced to the degree that participants do not receive it. The investigator should try to choose a study drug or intervention that is easy to apply or take and is well-tolerated. Adherence is likely to be poor if a behavioral intervention requires hours of practice by participants. Drugs that can be taken in a single daily dose are the easiest to remember and therefore preferable. The protocol should include provisions that will enhance adherence, such as instructing participants to take the pill at a standard point in the morning routine, giving them pill containers labeled with the day of the week, or sending reminders to their cell phones.

There is also a need to consider how best to **measure adherence** to the **intervention**, using such approaches as self-report, pill counts, pill containers with computer chips that record when the container is opened, and serum or urinary metabolite levels. This information can

TABLE 11.2 MAXIMIZING FOLLOW-UP AND ADHERENCE TO THE PROTOCOL

PRINCIPLE	EXAMPLE
Choose participants who are likely to be adherent to the intervention and protocol	Require completion of two or more visits before randomization
	Exclude those who are non-adherent in a pre-randomization run-in period
	Exclude those who are likely to move or be noncompliant
Make the intervention simple	Use a single tablet once a day if possible
Make study visits convenient and enjoyable	Schedule visits often enough to maintain close contact but not frequently enough to be tiresome
	Schedule visits in the evening or on weekends, or collect information by phone or e-mail
	Have adequate and well-organized staff to prevent waiting
	Provide reimbursement for travel and parking
	Establish good interpersonal relationships with participants
Make study measurements painless, useful, and interesting	Choose noninvasive, informative tests that are otherwise costly or unavailable
	Provide test results of interest to participants and appropriate counseling or referrals
Encourage participants to continue in the trial	Never discontinue follow-up for protocol violations, adverse events, or stopping the intervention
	Send participants birthday and holiday cards
	Send newsletters and e-mail messages
	Emphasize the scientific importance of adherence and follow-up
Find participants who are lost to follow-up	Pursue contacts of participants
	Use a tracking service

identify participants who are not complying, so that approaches to improving adherence can be instituted and the investigator can interpret the findings of the study appropriately.

Adherence to **study visits** and measurements can be enhanced by discussing what is involved in the study before consent is obtained, by scheduling the visits at a time that is convenient and with enough staff to prevent waiting, by calling or e-mailing the participant the day before each visit, and by reimbursing travel, parking, and other out-of-pocket costs.

Failure to **follow-up** trial participants and measure the outcome of interest can result in biased results, diminished credibility of the findings, and decreased statistical power. For example, a trial of nasal calcitonin spray to reduce the risk of osteoporotic fractures reported that treatment reduced fracture risk by 36% (13). However, about 60% of those randomized were lost to follow-up, and it was not known if fractures had occurred in these participants. Because the overall number of fractures was small, even a few fractures in the participants lost to follow-up could have altered the findings of the trial. This uncertainty diminished the credibility of the study findings (14).

Even if participants violate the protocol or discontinue the trial intervention, they should be followed so that their outcomes can be used in **intention-to-treat analyses** (see "Analyzing the Results" in this chapter). In many trials, participants who violate the protocol by enrolling in another trial, missing study visits, or discontinuing the study intervention are discontinued from follow-up; this can result in biased or uninterpretable results. Consider, for example, a drug that causes a symptomatic side effect that results in more frequent discontinuation of the study medication in those on active treatment compared to those on placebo. If participants who discontinue study medication are not continued in follow-up, this can bias the findings if the side effect is associated with the main outcome or with a serious adverse event (SAE).

Strategies for achieving complete **follow-up** are similar to those discussed for cohort studies (Chapter 7). At the outset of the study, participants should be informed of the importance of follow-up and investigators should record the name, address, e-mail address, and telephone number of one or two family members or close acquaintances who will always know where the participant is. In addition to enhancing the investigator's ability to assess vital status, the ability to contact participants by phone or e-mail may give him access to proxy outcome measures from those who refuse to come for a visit at the end. The Heart and Estrogen/Progestin Replacement Study (HERS) trial used all of these strategies: 89% of the women returned for the final clinic visit after an average of 4 years of follow-up, another 8% had a final telephone contact for outcome ascertainment, and information on vital status was determined for every one of the remaining participants by using registered letters, contacts with close relatives, and a tracking service (15).

The design of the trial should make it as easy as possible for participants to adhere to the intervention and complete all follow-up visits and measurements. Lengthy and stressful visits can deter some participants from attending. Participants are more likely to return for visits that involve noninvasive tests, such as computed tomography scans, than for invasive tests such as coronary angiography. Collecting follow-up information by phone or electronic means may improve adherence for participants who find visits difficult. On the other hand, participants may lose interest in a trial if there are not some social or interpersonal rewards for participation. Participants may tire of study visits that are scheduled monthly, and they may lose interest if visits only occur annually. Follow-up is improved by making the trial experience positive and enjoyable for participants: designing trial measurements and procedures to be painless and interesting; performing tests that would not otherwise be available; providing results of tests to participants (unless they are specialized research tests that are not yet established for clinical practice); sending newsletters, text messages, or e-mail notes of appreciation; hosting social media sites; sending holiday and birthday cards; giving inexpensive gifts; and developing strong interpersonal relationships with enthusiastic and friendly staff.

Two design aspects that are specific to trials may improve adherence and follow-up: screening visits before randomization and a run-in period. Asking participants to attend one or two **screening visits** before randomization may exclude participants who find that they cannot

complete such visits. The trick here is to set the hurdles for entry into the trial high enough to exclude those who will later be non-adherent, but not high enough to exclude participants who will turn out to have satisfactory adherence.

A **run-in period** may be useful for increasing the proportion of study participants who adhere to the intervention and follow-up procedures. During the baseline period, all participants are placed on placebo. A specified time later (usually a few weeks), only those who have complied with the intervention (e.g., taken at least 80% of the assigned placebo) are randomized. Excluding non-adherent participants before randomization in this fashion may increase the power of the study and permit a better estimate of the full effects of intervention. However, a run-in period delays entry into the trial, the proportion of participants excluded is generally small, and participants randomized to the active drug may notice a change in their medication following randomization, contributing to unblinding. It is also not clear that a placebo run-in is more effective in increasing adherence than the requirement that participants complete one or more screening visits before randomization. In the absence of a specific reason to suspect that adherence in the study will be poor, it is probably not necessary to include a run-in period in the trial design.

A variant of the **placebo run-in** design is the use of the active drug rather than the placebo for the run-in period. In addition to increasing adherence among those who enroll, an **active run-in** can select participants who tolerate and respond to the intervention; the absence of adverse effects, or the presence of a desired effect of treatment on a biomarker associated with the outcome, can be used as criteria for randomization. For example, in a placebo-controlled trial testing the effect of nitroglycerin on bone mass, the investigators used a 1-week active run-in period and excluded women who stopped nitroglycerin due to headache (16). This design maximized power by increasing the proportion of the intervention group that tolerated the drug and were likely to be adherent. However, the findings of trials using this strategy may not be generalizable to those excluded.

Using an active run-in may also result in underestimation of the rate of adverse effects. A trial of the effect of carvedilol on mortality in 1,094 patients with congestive heart failure used a 2-week active run-in period. During the run-in, 17 people had worsening congestive heart failure and 7 died (17). These people were not randomized in the trial, and these adverse effects of drug treatment were not included as outcomes.

Ascertaining and Adjudicating Outcomes

Data to ascertain that an outcome has occurred can come from many sources: self-report, standardized questionnaires, administrative or clinical records, laboratory or imaging tests, special measurements, and so on. Most self-reported outcomes, such as history of stroke or a participant report of quitting smoking, are not 100% accurate. Self-reported outcomes that are important to the trial should be confirmed if possible. Occurrence of disease, such as a stroke, is generally adjudicated by:

1. Creating clear criteria for the outcome (e.g., a new, persistent neurologic deficit with corresponding lesion on computed tomography or magnetic resonance imaging scan);
2. Collecting the clinical documents needed to make the assessment (e.g., discharge summaries and radiology reports);
3. Having blinded experts review each potential case and judge whether the criteria for the diagnosis have been met.

The adjudication is often done by two experts working independently, then resolving discordant cases by discussion between the two or by a third expert. However, involving multiple experts in adjudication can be expensive, and for straightforward outcomes in smaller studies it may be sufficiently accurate to have a single investigator carry out the adjudication. The important thing is that anyone involved in collecting the information and adjudicating the cases be blinded to the treatment assignment.

Monitoring Clinical Trials

Investigators must assure that participants are not exposed to a harmful intervention, denied a beneficial intervention, or continued in a trial if the research question is unlikely to be answered. Each of these three considerations must be monitored during the course of a trial to see if the trial should be stopped early.

- **Stopping for harm.** The most pressing reason to monitor clinical trials is to make sure that the intervention does not turn out unexpectedly to be harmful. If **harm** is judged to be clearly present and to outweigh benefits, the trial should be stopped.
- **Stopping for benefit.** If an intervention is more effective than was estimated when the trial was designed, statistically significant **benefit** can be observed early in the trial. When clear benefit has been proved, it may be unethical to continue the trial and delay offering the intervention to participants on placebo and to others who could benefit.
- **Stopping for futility.** If there is a very low probability of answering the research question, it may be unethical to continue participants in a trial that requires time and effort and that may cause some discomfort or risk. If a clinical trial is scheduled to continue for 5 years, for example, but after 4 years there is little difference in the rate of outcome events in the intervention and control groups, the "conditional power" (the likelihood of rejecting the null hypothesis in the remaining time, given the results thus far) becomes very small and consideration should be given to stopping the trial. Sometimes trials are stopped early if investigators are unable to recruit or retain enough participants to provide adequate power to answer the research question, or adherence to the intervention is very poor.

The research question might be answered by other trials before a given trial is finished. It is desirable to have more than one trial that provides evidence concerning a given research question, but if definitive evidence for either benefit or harm becomes available during a trial, it may be unethical to continue the trial.

Most clinical trials should include an **interim monitoring plan**. Trials funded by the National Institutes of Health (NIH) generally require interim monitoring, even if the intervention is considered safe (such as a behavioral intervention for weight loss). How interim monitoring will occur should be considered in the planning of any clinical trial. In small trials with interventions likely to be safe, the trial investigators might monitor safety or appoint a single independent data and safety monitor. In large trials and trials in which adverse effects of the intervention are unknown or potentially dangerous, interim monitoring is generally performed by a committee, usually known as the Data and Safety Monitoring Board (**DSMB**), consisting of experts in the disease or condition under study, biostatisticians, clinical trialists, ethicists, and sometimes a representative of the patient group being studied. These experts are not involved in the trial, and should have no personal or financial interest in its continuation. DSMB guidelines and procedures should be detailed in writing before the trial begins. Guidance for developing DSMB procedures is provided by the FDA and the NIH. Items to include in these guidelines are outlined in Table 11.3.

Stopping a trial should always be a careful decision that balances ethical responsibility to the participants and the advancement of scientific knowledge. Whenever a trial is stopped early, the chance to provide more conclusive results will be lost. The decision is often complex, and potential risks to participants must be weighed against possible benefits. Statistical tests of significance using one of the methods that compensates for multiple looks at the findings (Appendix 11B) provide important but not conclusive information for stopping a trial. Trends over time and effects on related outcomes should be evaluated for consistency, and the impact of stopping the study early on the credibility of the findings should be carefully considered (Example 11.2).

There are many statistical methods for monitoring the interim results of a trial. Analyzing the results of a trial repeatedly ("multiple peeks") is a form of multiple hypothesis testing and increases the probability of a type I error. For example, if $\alpha = 0.05$ is used for each interim

TABLE 11.3 MONITORING A CLINICAL TRIAL

- Elements to monitor
  - Recruitment
  - Randomization
  - Adherence to intervention and blinding
  - Follow-up completeness
  - Important variables
    - Outcomes
    - Adverse effects
    - Potential co-interventions
- Who will monitor
  - Trial investigator or a single monitor if small trial with minor hazards
  - Independent data and safety monitoring board otherwise
- Methods for interim monitoring
  - Specify statistical approach and frequency of monitoring in advance
  - Importance of judgment and context in addition to statistical stopping rules
- Changes in the protocol that can result from monitoring
  - Terminate the trial
  - Modify the trial
    - Stop one arm of the trial
    - Add new measurements necessary for safety monitoring
    - Discontinue high-risk participants
  - Extend the trial in time
  - Enlarge the trial sample

test and the results of a trial are analyzed four times during the trial and again at the end, the probability of making a type I error is increased from 5% to about 14% (18). To address this problem, statistical methods for interim monitoring generally decrease the α for each interim test so that the overall α is close to 0.05. There are multiple approaches to deciding how to "spend α" (Appendix 11B).

Analyzing the Results: Intention-to-Treat and Per-Protocol

Statistical analysis of the primary hypothesis of a clinical trial is generally straightforward. If the outcome is dichotomous, the simplest approach is to compare the proportions in the study groups using a **chi-squared test**. When the outcome is continuous, a ***t* test** may be used, or a nonparametric alternative if the outcome is not normally distributed. In many clinical trials, the duration of follow-up is different for each participant, necessitating the use of survival time methods. More sophisticated statistical models such as **Cox proportional hazards** analysis can accomplish this and at the same time adjust for chance maldistributions of baseline confounding variables (19).

One important issue that should be considered in the analysis of clinical trial results is the primacy of the intention-to-treat analytic approach to dealing with "**crossovers**," participants assigned to the active treatment group who do not get treatment or discontinue it, and those assigned to the control group who end up getting active treatment. An analysis done by **intention-to-treat** compares outcomes between the study groups with every participant analyzed according to his randomized group assignment, regardless of whether he adhered to the assigned intervention. Intention-to-treat analyses may underestimate the full effect of the treatment, but they guard against more important sources of biased results.

An alternative to the intention-to-treat approach is to perform "**per-protocol**" analyses that include only participants who adhered to the protocol. This is defined in various ways, but often includes only participants in both groups who were adherent to the assigned study medication, completed a certain proportion of visits or measurements, and had no other protocol violations. A subset of the per-protocol analysis is an "**as-treated**" analysis in which only participants who were adherent to the assigned intervention are included. These analyses *seem* reasonable because participants can only be affected by an intervention they actually receive. However, participants who adhere to the study treatment and protocol may be different from those who do not in ways that are related to the outcome. In the Postmenopausal Estrogen-Progestin Interventions (PEPI) trial, 875 postmenopausal women were randomly assigned to four different estrogen or estrogen plus progestin regimens and placebo (20). Among women assigned to the unopposed estrogen arm, 30% had discontinued treatment after 3 years because of endometrial hyperplasia, a precursor of endometrial cancer. If these women were eliminated in a per protocol analysis, the association of estrogen therapy and endometrial cancer would be missed.

The major disadvantage of the intention-to-treat approach is that participants who choose not to take the assigned intervention will, nevertheless, be included in the estimate of the effects of that intervention. Therefore, substantial discontinuation or crossover between treatments will cause intention-to-treat analyses to underestimate the magnitude of the effect of treatment. For this reason, results of trials are often evaluated with both intention-to-treat and per-protocol analyses. For example, in the Women's Health Initiative randomized trial of the effect of estrogen plus progestin treatment on breast cancer risk, the hazard ratio was 1.24 ($P = 0.003$) from the intention-to-treat analysis and 1.49 in the as-treated analysis ($P < 0.001$) (21). If the results of intention-to-treat and per protocol analyses differ, the intention-to-treat results generally predominate for estimates of efficacy because they preserve the value of randomization and, unlike per-protocol analyses, can only bias the estimated effect in the conservative direction (favoring the null hypothesis). However, for estimates of harm (e.g., the breast cancer findings), as-treated or per-protocol analyses provide the most conservative estimates, as interventions can only be expected to cause harm in exposed persons.

Results can only be analyzed by intention-to-treat if follow-up measures are completed regardless of whether participants adhere to treatment. Therefore, this should always be the goal.

Subgroup Analyses

Subgroup analyses are defined as comparisons between randomized groups in a subset of the trial cohort. The main reason for doing these analyses is to discover **effect modification** ("**interaction**") in subgroups, for example whether the effect of a treatment is different in men than in women. These analyses have a mixed reputation because they are easy to misuse and can lead to wrong conclusions. With proper care, however, they can provide useful ancillary information and expand the inferences that can be drawn from a clinical trial. To preserve the value of randomization, subgroups should be defined by measurements that were made before randomization. For example, a trial of denosumab to prevent fractures found that the drug decreased risk of non-vertebral fracture by 20% among women with low bone density. Preplanned subgroup analyses revealed that the treatment was effective (35% reduction in fracture risk; $P < 0.01$) among women with low bone density at baseline and that treatment was ineffective in women with higher bone density at baseline ($P = 0.02$ for effect modification) (22). It is important to note that the value of randomization is preserved: The fracture rate among women randomized to denosumab is compared with the rate among women randomized to placebo in each subgroup. Subgroup analyses based on post-randomization factors such as adherence to randomized treatment do not preserve the value of randomization and often produce misleading results.

Subgroup analyses can produce misleading results for several reasons. Being smaller than the entire trial population, there may not be sufficient power to find important differences; investigators should avoid claiming that a drug "was ineffective" in a subgroup when the finding

might reflect insufficient power to find an effect. Investigators often examine results in a large number of subgroups, increasing the likelihood of finding a different effect of the intervention in one subgroup by chance. For example, if 20 subgroups are examined, differences in one subgroup at $P < 0.05$ would be expected to occur by chance. To address this issue, planned subgroup analyses should be defined before the trial begins, and the number of subgroups analyzed should be reported with the results of the study (23). Claims about different responses in subgroups should be supported by evidence that there is a statistically significant interaction between the effect of treatment and the subgroup characteristic, and a separate study should confirm the effect modification before it is considered established.

SUMMARY

1. There are several variations on the randomized trial design that can substantially increase efficiency under the right circumstances:
   a. The **factorial design** allows two or more independent trials to be carried out for the price of one.
   b. **Cluster randomization** permits efficient studies of naturally occurring groups.
   c. **Non-inferiority or equivalence trials** compare a new intervention to an existing "standard of care."
   d. **Adaptive designs** increase efficiency by allowing design changes based on interim analyses, for example altering the **dose** of study drug, the **number** of participants, and the **duration** of follow-up.
2. There are also other useful clinical trial designs:
   a. **Time series designs** have a single group with outcomes compared within each participant during periods on and off an intervention.
   b. **Crossover designs** combine within and between group designs to enhance control over confounding (if **carryover effects** are not a problem) and **minimize sample size**.
3. Trials for regulatory approval of **new drugs** are classified as:
   a. **Phase I**, small trials to explore dosage and safety
   b. **Phase II**, medium-sized randomized or time series trials of drug effects at several doses
   c. **Phase III**, large randomized trials to demonstrate that benefits outweigh harms as the basis for FDA approval
   d. **Phase IV**, large post-marketing observational studies to confirm benefits and detect rare adverse effects
4. **Pilot studies** are important steps to help determine **acceptability** of interventions and **feasibility**, **size**, **cost**, and **duration** of planned trials.
5. In **conducting a trial**, if a substantial number of study participants **do not adhere** to the study intervention or are **lost to follow-up**, the results of the trial are likely to be underpowered, biased, or uninterpretable.
6. During a trial, **interim monitoring** by an independent **data and safety monitoring board (DSMB)** is needed to assure the **quality** of the study, and to decide if the trial should **stop early** due to evidence of **harm**, **benefit**, or **futility**.
7. **Intention-to-treat** analysis takes advantage of the control of confounding provided by randomization and should be the primary analysis approach for **assessing efficacy**. **Per protocol** analyses, a secondary approach that provides an estimate of the effect size in adherent participants (interpreted with caution), is the most conservative analysis of the harmful effects of treatment.
8. **Subgroup analyses** can detect whether the effect of treatment is modified by other variables; to minimize misinterpretations, the investigator should specify the subgroups in advance, test possible **effect modifications (interactions)** for statistical significance, and report the number of subgroups examined.

APPENDIX 11A
Specifying the Non-Inferiority Margin in a Non-Inferiority Trial

One of the most difficult issues in designing a **non-inferiority trial** is establishing the loss of efficacy of the new treatment that would be unacceptable (7), referred to as "Δ" and often called the **non-inferiority margin**. This decision is based on both statistical and clinical considerations of the potential efficacy and advantages of the new treatment, and requires expert judgment. Here's an example of how this works:

EXAMPLE 11.1 Designing a Study of a New Drug Compared to Warfarin in Patients with Atrial Fibrillation

Warfarin reduces risk for stroke in high-risk patients with atrial fibrillation, so a new drug should be compared to this standard of care. When warfarin is used to reduce the risk of stroke in this situation, it is difficult to dose correctly, requires frequent blood tests to monitor level of anticoagulation, and can cause major bleeding. If a new drug were available that did not have these drawbacks, it could be reasonable to prefer this drug to warfarin, even if its efficacy in reducing risk of stroke was slightly lower.

One approach to setting Δ is to perform a meta-analysis of previous trials of warfarin compared to placebo, and set Δ at some proportion of the distance between the null and lower bound for the treatment effect of warfarin. Alternatively, since studies included in meta-analyses often vary in quality, it may be better to base Δ on the results of the best quality randomized trial of warfarin that has similar entry criteria, warfarin dosage and outcome measures. It is important to set Δ such that there is a high likelihood, taking all benefits and harms into account, that the new therapy is better than placebo (6, 7).

Suppose that a meta-analysis of good-quality trials of warfarin compared to placebo shows that treatment with warfarin reduces the rate of stroke in high-risk patients with atrial fibrillation from 10% per year to about 5% per year (absolute treatment effect = 5%, 95% CI 4–6%). Given the advantages of our new drug, what loss of efficacy is clinically unacceptable? Perhaps an absolute efficacy that is 2% lower than warfarin would be acceptable? In this case, we would declare the new treatment non-inferior to warfarin if the lower limit of the confidence interval around the difference in stroke rates between warfarin and the new treatment is less than 2% (Figure 11.2). In a non-inferiority trial, it is also possible that the new treatment is found to be superior to the established treatment (topmost example in Figure 11.2).

APPENDIX 11B
Interim Monitoring of Trial Outcomes and Early Stopping

Interim monitoring of trial results to decide whether to stop a trial is a form of multiple hypothesis testing, and thereby increases the probability of a type I error. To address this problem, α for each test (α_i) is generally decreased so that the overall α is approximately = 0.05. There are multiple statistical methods for decreasing α_i.

One of the easiest to understand is the Bonferroni method, where $\alpha_i = \alpha/N$ if N is the total number of tests performed. For example, if the overall α is 0.05 and five tests will be performed, α_i for each test is 0.01. This method has two disadvantages, however: it requires using an equal threshold for stopping the trial at any interim analysis, and it results in a low α for the final analysis. Most investigators would rather use a more strict threshold for stopping a trial earlier rather than later and use an α close to 0.05 for the final analysis. In addition, this approach is too conservative because it assumes that each test is independent. Interim analyses are not independent, because each successive analysis is based on cumulative data, some of which were included in prior analyses. For these reasons, the Bonferroni method is not generally used.

A commonly used method suggested by O'Brien and Fleming (24) uses a very small α_i for the initial hypothesis test, then gradually increases it for each test such that α_i for the final test is close to the overall α. O'Brien and Fleming provide methods for calculating α_i if the investigator chooses the number of tests to be done and the overall α. At each test, $Z_i = Z^* (N_i)^{1/2}$, where $Z_i = Z$ value for the ith test; Z^* is determined so as to achieve the overall significance level; and N is the total number of tests planned. For example, for five tests and overall $\alpha = 0.05$, $Z^* = 2.04$; the initial $\alpha = 0.00001$ and the final $\alpha_5 = 0.046$. This method is unlikely to lead to stopping a trial very early unless there is a striking difference in outcome between randomized groups. In addition, this method avoids the awkward situation of getting to the end of a trial and accepting the null hypothesis when the P value is 0.04 or 0.03 but the α_i for the final test is diluted down to 0.01.

A major drawback to the O'Brien–Fleming method is that the number of tests and the proportion of data to be tested must be decided before the trial starts. In some trials, additional interim tests become necessary when important trends occur. DeMets and Lan (25) developed a method using a specified α-spending function that provides continuous stopping boundaries. The α_i at a particular time (or after a certain proportion of outcomes) is determined by the function and by the number of previous "looks." Using this method, the number of "looks" and the proportion of data to be analyzed at each "look" do not need to be specified before the trial. Of course, for each additional unplanned interim analysis conducted, the final α is a little smaller.

A different set of statistical methods based on curtailed sampling techniques suggests termination of a trial if future data are unlikely to change the conclusion. The multiple testing problem is irrelevant because the decision is based only on estimation of what the data will show at the end of the trial. A common approach is to compute the probability of rejecting the null hypothesis at the end of the trial, conditioned on the accumulated data. A range of conditional power is typically calculated, first assuming that *Ho* is true (i.e., that any future outcomes in the treated and control groups will be equally distributed) and also assuming that H_a is true (i.e., that outcomes will be distributed unequally in the treatment and control groups as specified by H_a). Other estimates can also be used to provide a full range of reasonable effect sizes. If the

conditional power to reject the null hypothesis across the range of assumptions is low, the null hypothesis is not likely to be rejected and the trial might be stopped.

Examples of two trials that were stopped early are presented in Example 11.2

EXAMPLE 11.2 Two Trials That Have Been Stopped Early

Cardiac Arrhythmia Suppression Trial (CAST) (26). The occurrence of premature ventricular contractions in survivors of myocardial infarction (MI) is a risk factor for sudden death. The CAST evaluated the effect of antiarrhythmic therapy (encainide, flecainide, or moricizine) in patients with asymptomatic or mildly symptomatic ventricular arrhythmia after MI on risk for sudden death. During an average of 10 months of follow-up, the participants treated with active drug had a higher total mortality (7.7% versus 3.0%) and a higher rate of death from arrhythmia (4.5% versus 1.5%) than those assigned to placebo. The trial was planned to continue for 5 years but this large and highly statistically significant difference led to the trial being stopped after 18 months.

Physicians' Health Study (27). The Physicians' Health Study was a randomized trial of the effect of aspirin (325 mg every other day) on cardiovascular mortality. The trial was stopped after 4.8 years of the planned 8-year follow-up. There was a statistically significant reduction in risk of non-fatal myocardial infarction in the treated group (relative risk = 0.56), but no difference in the number of cardiovascular disease deaths. The rate of cardiovascular disease deaths observed in the study was far lower than expected (88 after 4.8 years of follow-up versus 733 expected), and the trial was stopped because of the beneficial effect of aspirin on risk for nonfatal MI coupled with the very low conditional power to detect a favorable impact on cardiovascular mortality.

REFERENCES

1. Ridker PM, Cook NR, Lee I, et al. A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. *N Engl J Med* 2005;352:1293–1304.
2. The Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. *Control Clin Trials* 1998;19:61–109.
3. Walsh M, Hilton J, Masouredis C, et al. Smokeless tobacco cessation intervention for college athletes: results after 1 year. *Am J Public Health* 1999;89:228–234.
4. Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. *Am J Epidemiol* 1981;114:906–914.
5. Piaggio G, Elbourne DR, Altman DG, et al. Reporting of non-inferiority and equivalence randomized trials. An extension of the CONSORT Statement. *JAMA* 2006;295:1152–1160.
6. Piaggio G, Elbourne DR, Pocock SJ, et al. Reporting of non-inferiority and equivalence randomized trials. An extension of the CONSORT 2010 statement. *JAMA* 2012;308:2594–2604.
7. Kaul S, Diamond GA. Good enough: a primer on the analysis and interpretation of non-inferiority trials. *Ann Intern Med* 2006;145:62–69.
8 D'Agostino RB Sr., Massaro JM, Sullivan LM, et al. Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Statist Med* 2003;22:169–186.
9. Chang M, Chow S, Pong A. Adaptive design in clinical research: issues, opportunities, and recommendations. *J Biopharm Stat* 2006;16:299–309.
10. Chalmers T, Celano P, Sacks H, et al. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983;309:1358–1361.
11. Pocock S. Current issues in the design and interpretation of clinical trials. *Br Med J* 1985;296:39–42.
12. Nickles CJ, Mitchall GK, Delmar CB, et al. An n-of-1 trial service in clinical practice: testing the effectiveness of stimulants for attention-deficit/hyperactivity disorder. *Pediatrics* 2006;117:2040–2046.
13. Chestnut CH III, Silverman S, Andriano K, et al. A randomized trial of nasal spray salmon calcitonin in postmenopausal women with established osteoporosis: the prevent recurrence of osteoporotic fractures study. *Am J Med* 2000;109:267–276.

14. Cummings SR, Chapurlat R. What PROOF proves about calcitonin and clinical trials. *Am J Med* 2000;109:330–331.
15. Hulley S, Grady D, Bush T, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA* 1998;280:605–613.
16. Jamal SA, Hamilton CJ, Eastell RJ, Cummings SR. Effect of nitroglycerin ointment on bone density and strength in postmenopausal women. *JAMA* 2011;305:800–805.
17. Pfeffer M, Stevenson L. Beta-adrenergic blockers and survival in heart failure. *N Engl J Med* 1996;334:1396–1397.
18. Armitage P, McPherson C, Rowe B. Repeated significance tests on accumulating data. *J R Stat Soc* 1969;132A:235–244.
19. Friedman LM, Furberg C, DeMets DL. *Fundamentals of clinical trials*, 3rd ed. St. Louis, MO: Mosby Year Book, 1996.
20. Writing Group for the PEPI Trial. Effects of estrogen or estrogen/progestin regimens on heart disease risk factors in postmenopausal women. *JAMA* 1995;273:199–208.
21. Writing group for WHI investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. *JAMA* 2001;288:321-333.
22. McClung MR, Boonen S, Torring O, et al. Effect of denosumab treatment on the risk of fractures in subgroup of women with postmenopausal osteoporosis. *J Bone Mineral Res* 2012;27:211–218.
23. Wang R, Lagakos SW, Ware JH, et al. Statistics in medicine—Reporting of subgroup analyses in clinical trials. *NEJM* 2007;357:2189–2194.
24. O'Brien P, Fleming T. A multiple testing procedure for clinical trials. *Biometrics* 1979;35:549–556.
25. DeMets D, Lan G. The alpha spending function approach to interim data analyses. *Cancer Treat Res* 1995;75:1–27.
26. Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med* 1989;321:406–412.
27. Physicians' Health Study Investigations. Findings from the aspirin component of the ongoing Physicians' Health Study. *N Engl J Med* 1988;318:262–264.

CHAPTER 12

Designing Studies of Medical Tests

Thomas B. Newman, Warren S. Browner, Steven R. Cummings, and Stephen B. Hulley

Medical tests, such as those performed to screen for a risk factor, diagnose a disease, or estimate a patient's prognosis, are an important aspect of clinical research. The study designs discussed in this chapter can be used when **studying whether**, **and in whom**, **a particular test should be performed**.

Most designs for studies of medical tests resemble the observational designs in Chapters 7 and 8. There are, however, some important differences between most observational studies and those used to evaluate medical tests. Most important, the goal of most observational studies is to identify statistically significant associations (Chapter 5) that represent causal relationships (Chapter 9). In contrast, demonstrating that a test result has a statistically significant association with a particular condition is not nearly enough to determine whether that test would be useful clinically, and for studies of medical tests, causality is often irrelevant. Thus, odds ratios and *P* values are secondary considerations for studies of medical tests, which focus instead on *descriptive* parameters such as **sensitivity**, **specificity**, and **likelihood ratios** along with their associated **confidence intervals**.

DETERMINING WHETHER A TEST IS USEFUL

For a test to be useful it must pass muster on a series of increasingly difficult questions that address its **reproducibility**, **accuracy**, **feasibility**, and, most importantly, its **effects on clinical decisions** and **outcomes** (Table 12.1). Favorable answers to these questions are necessary but not sufficient criteria for a test to be worth doing. For example, if a test gives very different results depending on who does it or where it is done, it is unlikely to be useful. If the test seldom supplies new information, it is unlikely to affect clinical decisions. Even if it affects decisions, if these decisions do not improve the clinical outcome of patients who were tested at reasonable risk and cost, the test still may not be useful.

Of course, if using a test improves the outcomes of tested patients, favorable answers to the other questions can be inferred. However, studies of whether doing a test improves patient outcomes are the most difficult to do. Instead, the potential effects of a test on outcomes are usually inferred by comparing the accuracy, safety, or costs with those of existing tests. When developing a new diagnostic or prognostic test, it may be worthwhile to consider what aspects of current practice are most in need of improvement. For example, are current tests unreliable, inaccurate, expensive, dangerous, or difficult to perform?

General Issues for Studies of Medical Tests

- **Spectrum of disease severity and of test results.** Because the goal of most studies of medical tests is to draw inferences about populations by making measurements on samples, the way the sample is selected has a major effect on the validity of the inferences. **Spectrum**

TABLE 12.1 QUESTIONS TO DETERMINE THE USEFULNESS OF A MEDICAL TEST, POSSIBLE DESIGNS TO ANSWER THEM, AND STATISTICS FOR REPORTING RESULTS

QUESTION	POSSIBLE DESIGNS	STATISTICS FOR RESULTS*
How reproducible is the test?	Studies of intra- and inter-observer and intra- and inter-laboratory variability	Proportion agreement, kappa, coefficient of variation, mean, and distribution of differences (avoid correlation coefficient)
How accurate is the test?	Cross-sectional, case–control, or cohort-type designs in which test result is compared with a gold standard	Sensitivity, specificity, positive and negative predictive value, receiver operating characteristic curves, and likelihood ratios
How often do test results affect clinical decisions?	Diagnostic yield studies, studies of pre- and posttest clinical decision making	Proportion abnormal, proportion with discrepant results, proportion of tests leading to changes in clinical decisions; cost per abnormal result or per decision change
What are the costs, risks, and acceptability of the test?	Prospective or retrospective studies	Mean costs, proportions experiencing adverse effects, proportions willing to undergo the test
Does doing the test improve clinical outcome or have adverse effects?	Randomized trials, cohort or case–control studies in which the predictor variable is receiving the test and the outcomes include morbidity, mortality, or costs related either to the disease or to its treatment	Risk ratios, odds ratios, hazard ratios, numbers needed to treat, rates and ratios of desirable and undesirable outcomes

*Most statistics in this table should be presented with confidence intervals.

bias occurs when the spectrum of disease (or non-disease) in the sample differs from that of the patients to whom the investigator wishes to generalize. Early in the development of a diagnostic test, it may be reasonable to investigate whether a test can distinguish between subjects with clear-cut, late stage disease and healthy controls; if the answer is no, the investigator can go back to the lab to work on a modification or a different test. Later, however, when the research question addresses the clinical utility of the test, the spectrum of both disease and non-disease should be representative of the patients in whom the test will be used. For example, a test developed by comparing symptomatic pancreatic cancer patients to healthy controls could later be evaluated on a more difficult but clinically realistic sample, such as consecutive patients with unexplained abdominal pain or weight loss.

Spectrum bias can occur from an inappropriate spectrum of test results as well as an inappropriate spectrum of disease. For example, consider a study of inter-observer agreement among radiologists reading mammograms. If they are asked to classify the films as normal or abnormal, their agreement will be much higher if the "positive" films the investigator selects for them to examine are selected because they are clearly abnormal, and the "negative" films are selected because they are free of all suspicious abnormalities.

- **Importance of blinding.** Many studies of diagnostic tests involve judgments, such as whether to consider a radiograph abnormal, or whether a patient meets the criteria for diagnosing a particular disease. Whenever possible, investigators should blind those interpreting test results from other information about the patient being tested. In a study of the contribution of ultrasonography to the diagnosis of appendicitis, for example, those reading the sonograms should not know the results of the history and physical

examination.[1] Similarly, the pathologists making the final determination of who does and does not have appendicitis (the gold standard to which sonogram results will be compared) should not know the results of the ultrasound examination. Blinding prevents biases, preconceptions, and information from sources other than the test from affecting these judgments.

- **Sources of variation, generalizability, and the sampling scheme.** For some research questions, differences among patients are the main source of variation in the results of a test. For example, some infants with bacteremia (bacteria in the blood) will have an elevated white blood cell count, whereas others will not. The proportion of bacteremic infants with high white blood cell counts is not expected to vary much according to which laboratory does the blood count. On the other hand, many test results depend on the person doing the test or the setting in which the test is done. For example, the sensitivity, specificity, and inter-rater reliability for interpreting mammograms depend on the readers' skill and experience, as well as the quality of the equipment. When accuracy may vary from reader to reader or institution to institution, it is helpful to study different readers and institutions to assess the consistency of the results.
- **Gold standard for diagnosis.** Some diseases have a **gold standard** that is generally accepted to indicate the presence or absence of the target disease, such as the pathological examination of a tissue biopsy specimen for cancer. Other diseases have definitional gold standards, such as defining coronary artery disease as a 50% obstruction of at least one major coronary artery as seen with coronary angiography. Still others, such as rheumatologic diseases, require that a patient have a specified number of signs, symptoms, or laboratory abnormalities to meet the criteria for having the disease. Of course, if any signs, symptoms, or laboratory tests used to diagnose a disease are used as part of the gold standard, a study comparing them to that gold standard can make them look falsely good. This is called **incorporation bias** because the test being studied is *incorporated* into the gold standard; avoiding it is one of the previously mentioned reasons for blinding.

  It is also important to consider whether the gold standard is truly gold. If the gold standard is imperfect it can make a test either look worse than it really is (if in reality the test outperforms the gold standard) or better than it really is (if the index test makes the same mistakes as the gold standard).
- **What constitutes a positive test?** Particularly if a test has continuous results (like a serum erythropoietin level), it may be tempting for an investigator to look at all the results in those with the outcome (say, anemia of chronic disease) and those without the outcome (other types of anemia), and then select the best **cut point** to define a positive test. However, this is a type of **overfitting** (i.e., random variation in the particular sample studied that makes the test performance look better than it is in the population). Better approaches are to base the cut point on clinical or biological knowledge from other studies or to divide continuous tests into intervals, then calculate likelihood ratios for each interval (see the following text). To minimize overfitting, cut points for defining intervals should be specified in advance, or reasonable round numbers should be used. Overfitting is a particular issue for clinical prediction rules, which are discussed later in this chapter.

STUDIES OF TEST REPRODUCIBILITY

Sometimes the results of tests vary according to when or where they were done or who did them. **Intra-observer variability** describes the lack of reproducibility in results when the same observer or laboratory performs the test on the same specimen at different times. For example, if a radiologist is shown the same chest radiograph on two occasions, what percent of the time will he agree with himself on the interpretation, assuming he is unaware of his prior interpretation? **Inter-observer variability** describes the lack of reproducibility among two or more

[1] Alternatively, the accuracy of the history and physical examination alone could be compared with the accuracy of history and physical examination plus ultrasound.

observers: If another radiologist is shown the same film, how likely is he to agree with the first radiologist?

Often, the level of reproducibility (or lack thereof) is the main research question. In other cases, reproducibility is studied with the goal of quality improvement, either for clinical care or for a research study. When reproducibility is poor—because either intra- or inter-observer variability is large—a measurement is unlikely to be useful, and it may need to be either improved or abandoned.

Studies of reproducibility per se address precision, not accuracy or validity (Chapter 4), so all observers can agree with one another and still be wrong. When a gold standard is available, investigators of intra- and inter-observer reproducibility may compare subjects' observations with a gold standard to determine accuracy. When no gold standard is available, investigators must rely on the other methods of assessing validity described in Chapter 4.

Designs

The basic design to assess test reproducibility involves comparing test results from more than one observer or that were performed on more than one occasion. For tests that involve several steps, differences in any one of which might affect reproducibility, the investigator will need to decide on the breadth of the study's focus. For example, measuring inter-observer agreement of pathologists on a set of Pap smear slides in a single hospital may overestimate the overall reproducibility of Pap smears because the variability in how the sample was obtained and how the slide was prepared would not be captured.

The extent to which an investigator needs to isolate the steps that might lead to inter-observer disagreement depends partly on the goals of his study. Most studies should estimate the reproducibility of the entire testing process, because this is what determines whether the test is worth using. On the other hand, an investigator who is developing or improving a test may want to focus on the specific steps that are problematic in order to improve the process. In either case, the investigator should lay out the exact process for obtaining the test result in the operations manual (Chapters 4 and 17) and then describe it in the methods section when reporting the study results.

Analysis

- **Categorical variables.** The simplest measure of inter-observer agreement is the percent of observations on which the observers agree exactly. However, when the observations are not evenly distributed among the categories (e.g., when the proportion that are "abnormal" on a dichotomous test is not close to 50%), the percent agreement can be hard to interpret, because it does not account for agreement that could result simply from both observers having some knowledge about the prevalence of abnormality. For example, if 95% of subjects are normal, two observers who randomly choose which 5% of tests to call "abnormal" will agree that results are "normal" about 90% of the time. The percent agreement is also a suboptimal measure when a test has more than two possible results that are intrinsically ordered (e.g., normal, borderline, abnormal), because it counts partial disagreement (e.g., normal/borderline) the same as complete disagreement (normal/abnormal).

  A better measure of inter-observer agreement, called **kappa** (Appendix 12A), measures the extent of agreement beyond what would be expected from observers' knowledge of the prevalence of abnormality,[2] and can give credit for partial agreement. Kappa ranges from –1 (perfect disagreement) to 1 (perfect agreement). A kappa of 0 indicates no more agreement than would be expected from the observers' estimates of the prevalence of each level of abnormality. Kappa values above 0.8 are generally considered very good; levels of 0.6 to 0.8 are good.

[2] Kappa is often described as the extent of agreement beyond that expected by chance, but the estimate of the agreement expected by chance is based on the prevalence of abnormality assigned by each observer.

- **Continuous variables.** Measures of inter-observer variability for continuous variables depend on the design of the study. Some studies measure the agreement between just two machines or methods (e.g., temperatures obtained from two different thermometers). The best way to describe the data from such a study is to gather the data on the pairs of measurements (each pair consists of two measurements made at close to the same time in the same subject) and report the mean difference between those pairs with some measure of the spread of values, such as the standard deviation or how often the difference exceeds a clinically relevant threshold. For example, if a clinically important difference in body temperature is 0.3°C, a study comparing temperatures from tympanic and rectal thermometers could estimate the mean (± standard deviation) difference between the two techniques, and report how often the two measurements differed by more than 0.3°C.[3]

  Other studies examine the inter-assay, inter-observer, or inter-instrument variability of tests across a large group of different technicians, laboratories, or machines. These results are commonly summarized using the **coefficient of variation (CV)**, which is the standard deviation of all of the results obtained from a single specimen divided by the mean value. Often, the CVs of two or more different assays or instruments are compared; the one with the smallest CV is the most precise (though it may not be the most accurate).

STUDIES OF THE ACCURACY OF TESTS

Studies in this section address the question, "To what extent does the test give the right answer?" This assumes, of course, that a **gold standard** is available to reveal what the right answer is.

Designs

- **Sampling.** Studies of diagnostic test accuracy can have designs analogous to case–control or cross-sectional studies. In the **case–control** design of a diagnostic test, those with and without the disease are sampled separately, and the test results in the two groups are compared. As previously noted, case–control sampling may be appropriate early in the development of a diagnostic test, when the research question is whether the test warrants further study. Later, when the research question is the *clinical utility* of the test, the spectra of disease and non-disease should resemble those of the people to whom the test will be applied clinically; this is much more difficult to achieve with case–control sampling than with samples designed to be representative of the whole target population.

  Studies of tests that sample those with and without the disease separately are subject to bias in the measurement or reporting of the test result, because its measurement necessarily comes after the measurement of disease status. In addition, studies with this sampling scheme usually cannot be used to estimate **predictive values** (discussed in the following text).

  A **consecutive** sample of patients being evaluated for a particular diagnosis generally will yield more valid and interpretable results, including predictive values. For example, Tokuda et al. (3) found that the degree of chills (e.g., feeling cold versus whole body shaking under a thick blanket) was a strong predictor of bacteremia in a series of 526 consecutive febrile adult emergency department patients. Because the subjects were enrolled before it was known whether they were bacteremic, the spectrum of patients in this study should be reasonably representative of patients who present to emergency departments with fever.

[3] Although commonly used, the correlation coefficient is best avoided in studies of the reliability of laboratory tests because it is highly influenced by outlying values and does not allow readers to determine how frequently differences between the two measurements are clinically important. Confidence intervals for the mean difference should also be avoided because their dependence on sample size makes them potentially misleading. A narrow confidence interval for the mean difference between the two measurements does not imply that they generally closely agree—only that the mean difference between them is being measured precisely. See Bland and Altman (1) or Newman and Kohn (2) for additional discussion of these points.

A sampling scheme that we call **tandem testing** is sometimes used to compare two (presumably imperfect) tests with one another. Both tests are done on a representative sample of subjects and the gold standard is selectively applied to those with positive results on either or both tests. The gold standard should also be applied to a random sample of patients with concordant negative results, to make sure that they really don't have the disease. This design, which allows the investigator to determine which test is more accurate without the expense of doing the gold standard test in all the subjects with negative results, has been used in studies comparing different cervical cytology methods (4).

Prognostic test studies require **cohort designs**. In a prospective design, the test is done at baseline, and the subjects are then followed to see who develops the outcome of interest. A retrospective cohort study can be used when a new test becomes available, such as viral load in HIV-positive patients, if a previously defined cohort with banked blood samples is available. Then the viral load can be measured in the stored blood to see whether it predicts prognosis. The nested case–control design (Chapter 8) is particularly attractive if the outcome of interest is rare and the test is expensive.

- **Predictor variable: the test result.** Although it is simplest to think of the results of a diagnostic test as being either positive or negative, many tests have categorical, ordinal, or continuous results. In order to take advantage of all available information in the test, investigators should generally report the results of ordinal or continuous tests rather than dichotomizing as "normal or abnormal." Most tests are more indicative of disease if they are very abnormal than if they are slightly abnormal, and have a borderline range in which they do not provide much information.
- **Outcome variable: the disease (or its outcome).** The outcome variable in a **diagnostic test study** is the presence or absence of the disease, which is best determined with a gold standard. Wherever possible, the assessment of outcome should not be influenced by the results of the diagnostic test being studied. This is best accomplished by blinding those doing the gold standard test so that they do not know the results of the index test.

  Sometimes, particularly with screening tests, uniform application of the gold standard is not ethical or feasible. For example, Smith-Bindman et al. (5) studied the accuracy of mammography according to characteristics of the interpreting radiologist. Women with positive mammograms were referred for further tests, eventually with pathologic evaluation as the gold standard. However, it is not reasonable to do breast biopsies in women whose mammograms are negative. Therefore, to determine whether these women had false-negative mammograms, the authors linked their mammography results with local tumor registries and considered whether or not breast cancer was diagnosed in the year following mammography to be the gold standard. This solution assumes that all breast cancers that exist at the time of mammography will be diagnosed within 1 year, and that all cancers diagnosed within 1 year existed at the time of the mammogram. Measuring the gold standard differently depending on the result of the test creates a potential for bias, discussed in more detail at the end of the chapter, but sometimes that is the only feasible option.

  The outcome variable in a **prognostic test study** involves what happens to patients with a disease, such as how long they live, what complications they develop, or what additional treatments they require. Again, blinding is important, especially if clinicians caring for the patients may make decisions based upon the prognostic factors being studied. For example, Rocker et al. (6) found that attending physicians' estimates of prognosis, but not those of bedside nurses, were independently associated with intensive care unit mortality. This could be because the attending physicians were more skilled at estimating severity of illness, but it could also be because physician prognostic estimates had a greater effect than those of the nurses on decisions to withdraw life support. To distinguish between these possibilities, it would be helpful to obtain estimates of prognosis from attending physicians other than those involved in making or framing decisions about withdrawal of support.

Analysis

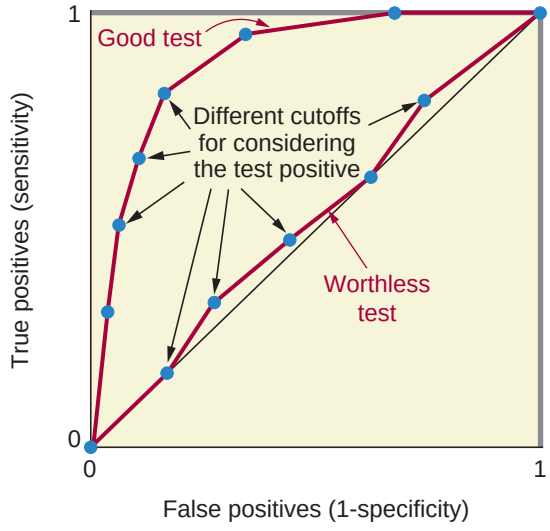
- **Sensitivity, specificity, and positive and negative predictive values.** When results of a dichotomous test are compared with a dichotomous gold standard, the results can be summarized in a 2 × 2 table (Table 12.2). The **sensitivity** of a test is defined as the proportion of subjects with the disease in whom the test gives the right answer (i.e., is positive); **specificity** is the proportion of subjects without the disease in whom the test gives the right answer (i.e., is negative). If the sample of patients who were studied is representative of the group of patients in whom the test would be used, two additional parameters can be calculated. The **positive predictive value** is the proportion of subjects with positive tests who have the disease; the **negative predictive value** is the proportion of subjects with negative tests who don't have the disease.
- **Receiver operating characteristic curves.** Many diagnostic tests yield ordinal or continuous results. With such tests, several values of sensitivity and specificity are possible, depending on the cutoff chosen to define a positive test. This trade-off between sensitivity and specificity can be displayed using a graphic technique originally developed in electronics: **receiver operating characteristic (ROC) curves**. The investigator selects several cutoff points and determines the sensitivity and specificity at each point. He then graphs the sensitivity (or true-positive rate) on the *Y*-axis as a function of 1 – specificity (the false-positive rate) on the *X*-axis. An ideal test is one that reaches the upper left corner of the graph (100% true-positives and no false-positives). A worthless test follows the diagonal from the lower left to the upper right corners: at any cutoff the true-positive rate is the same as the false-positive rate (Figure 12.1). The area under the ROC curve, which thus ranges from 0.5 for a useless test to 1.0 for a perfect test, is a useful summary of the overall accuracy of a test and can be used to compare the accuracy of two or more tests.
- **Likelihood ratios.** Although the information in a diagnostic test with continuous or ordinal results can be summarized using sensitivity and specificity or ROC curves, there is a better way. Likelihood ratios allow the investigator to take advantage of all information in a test. For each test result, the likelihood ratio is the ratio of the likelihood of that result in someone with the disease to the likelihood of that result in someone without the disease.

$$\text{Likelihood ratio} = \frac{\text{P (Result | Disease)}}{\text{P (Result | No Disease)}}$$

TABLE 12.2 SUMMARIZING RESULTS OF A STUDY OF DICHOTOMOUS TESTS IN A 2 × 2 TABLE

		GOLD STANDARD			
		DISEASE	NO DISEASE	TOTAL	
TEST	***Positive***	a True-positive	b False-positive	a + b	Positive predictive value = a/(a + b)
	Negative	c False-negative	d True-negative	c + d	Negative predictive value = d/(c + d)
	Total	a + c	b + d		
		Sensitivity = a/(a + c)	Specificity = d/(b + d)		

Positive and negative predictive values can be calculated from a 2 × 2 table like this only when the prevalence of disease is (a + c)/(a + b + c + d). This will not be the case if subjects with and without disease are sampled separately (e.g., 100 of each in a study with case–control sampling).



■ **FIGURE 12.1** Receiver operating characteristic curves for good and worthless tests.

The P is read as "probability of" and the "|" is read as "given." Thus, P(Result|Disease) is the probability of the result given disease, and P(Result|No Disease) is the probability of that result given no disease. The likelihood ratio is a ratio of these two probabilities.[4]

The higher the likelihood ratio, the better the test result for *ruling in* a disease; a likelihood ratio greater than 100 is very high (and unusual among tests). On the other hand, the lower a likelihood ratio (the closer it is to 0), the better the test result is for *ruling out* the disease. A likelihood ratio of 1 means that the test result provides no information at all about the likelihood of disease; those close to 1 (say from 0.8 to 1.25) provide little helpful information.

An example of likelihood ratios is shown in Table 12.3, which presents results from a study of complete blood counts in newborns at risk for serious infections (7). A white blood cell count less than 5,000 cells/μL was much more common among infants with serious infections than among other infants. The calculation of likelihood ratios simply quantifies this: 19% of the infants with infections had white blood cell counts less than 5,000 cells/μL, compared with only 0.52% of those without infections. Therefore, the likelihood ratio is 19%/0.52% = 36.

---

[4] For dichotomous tests the likelihood ratio for a *positive* test is

$$\frac{\text{Sensitivity}}{1-\text{Specificity}}$$

and the likelihood ratio for a *negative* test is

$$\frac{1-\text{Sensitivity}}{\text{Specificity}}.$$

Detailed discussions of how to use likelihood ratios and prior information (the prior probability of disease) to estimate a patient's probability of disease after knowing the test result (the posterior probability) are available in Newman and Kohn (2). The formula is

$$\text{Prior odds} \times \text{Likelihood ratio} = \text{Posterior odds}$$

where prior and posterior odds are related to their respective probabilities by

$$\text{odds} = \frac{P}{1-P}.$$

TABLE 12.3 EXAMPLE OF CALCULATION OF LIKELIHOOD RATIOS FROM A STUDY OF COMPLETE BLOOD COUNTS TO PREDICT SERIOUS INFECTIONS IN YOUNG NEWBORNS (7)

WHITE BLOOD CELL COUNT (PER μL)	SERIOUS INFECTION		LIKELIHOOD RATIO
	YES	NO	
<5,000	46 19%	347 0.52%	36
5,000–9,999	53 22%	5,103 7.6%	2.9
10,000–14,999	53 22%	16,941 25%	0.86
15,000–19,999	45 18%	21,168 31%	0.58
≥20,000	48 20%	23,818 35%	0.56
Total	245 100%	67,377 100%	

- **Absolute risks, risk ratios, risk differences, and hazard ratios.** The analysis of studies of **prognostic tests** is similar to that of other cohort studies. If everyone in a prognostic test study is followed for a set period of time (say 3 years) with few losses to follow-up, then the results can be summarized with absolute risks, risk ratios, and risk differences. Especially when follow-up is complete and of short duration, results of studies of prognostic tests are sometimes summarized like those of diagnostic tests, using sensitivity, specificity, predictive value, likelihood ratios, and ROC curves. On the other hand, when the study subjects are followed for varying lengths of time, a survival-analysis technique that accounts for the length of follow-up time and estimates hazard ratios is preferable (8).
- **Net reclassification improvement.** For new tests or biomarkers intended to predict future disease events, it is important to quantify what the new tests add to existing prediction models. While one way to do this is to look at the amount they increase the area under the ROC curve, changes in the area under the ROC curve are often small, even for well-established predictors, and are difficult to translate into projected changes in clinical decisions and patient outcomes (9, 10). A more direct approach, which is most useful when treatment thresholds are well-established, is to examine how often a model or clinical prediction rule including the new test changes the classification of patients from one risk category (and treatment decision) to another, compared with the old model. If the new test improves prediction, more subjects who develop the outcome ("cases") should move *up* to a higher risk category than move *down* to a lower risk category; the opposite should be true for those who do not develop the outcome ("controls"): their risk should move *down* in more subjects than it moves *up*. **Net reclassification improvement (NRI)** quantifies these differences as follows (11):

$$\text{NRI} = \text{P(up|case)} - \text{P(down|case)} + \text{P(down|control)} - \text{P(up|control)}$$

where P(up|case) is the proportion of cases in whom the model with the new marker led to the subject moving to a higher risk category and the other terms are correspondingly defined. For example, Shepherd et al. (12) found that adding the calculated mammographic fibroglandular volume (i.e., the estimated amount of breast tissue at risk of malignancy) to a model that included traditional clinical risk factors improved the prediction of subsequent breast cancer or ductal carcinoma in situ with an NRI of 21% (P = 0.0001).

STUDIES TO CREATE CLINICAL PREDICTION RULES

Studies to create **clinical prediction rules** differ from studies of existing tests (or rules) because the goal is to improve clinical decisions by using mathematical methods to develop a new (composite) test, rather than to evaluate one that already exists.

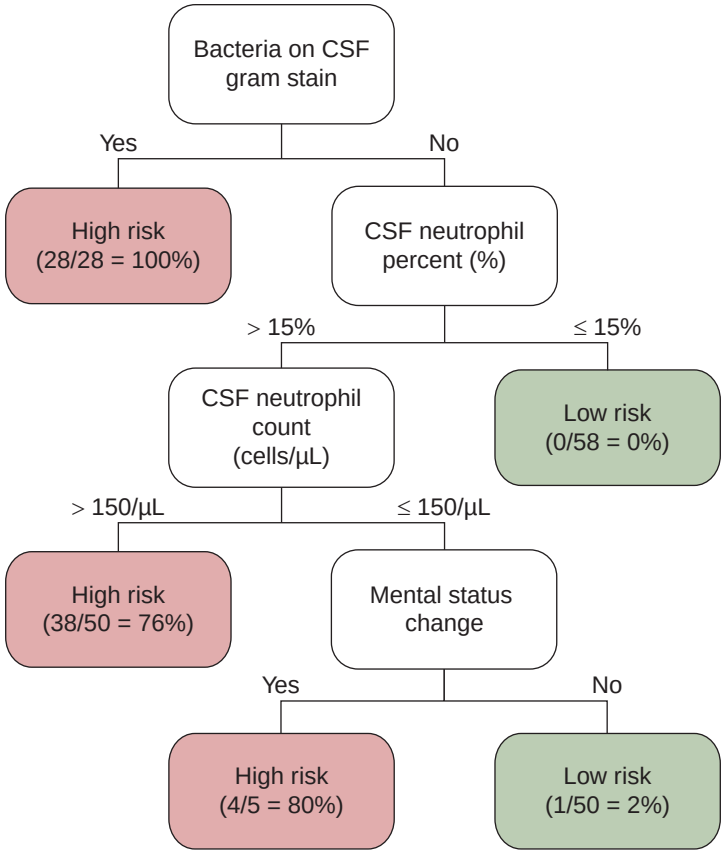
Subjects for these studies should be similar to those in whom the rule will be applied. Clinical prediction rules are likely to be most helpful when intended to guide a specific clinical decision, such as the decision to start treatment with statins (for which the Framingham Risk Score is used). Therefore, subjects should be those in whom the specific clinical decision needs to be made, especially those in whom it is currently difficult or uncertain (13). Many studies to develop clinical decision rules include subjects only from a single center, but those developed using data from multiple centers are more likely to be generalizable.

Mathematical methods for creating prediction rules generally involve a multivariate technique for selecting candidate predictor variables and combining their values to generate a prediction. The candidate variables should include all known and plausible **predictor variables** that can be easily, reliably, and inexpensively measured. A multivariate model, such as **logistic regression** or the **Cox (proportional hazards) model**, can quantify the independent contribution of candidate predictor variables for predicting the outcome. Those most strongly and consistently associated with outcome can be included in the rule, and points can be assigned to different values of the predictor variables depending on the coefficients in the model. For example, Wells et al. (14) used logistic regression analysis on 40 potential clinical predictors of pulmonary embolism to create a prediction score based on just 7 variables (Table 12.4). This now popular score is used to assign a pretest probability of pulmonary embolism, to guide further testing decisions and the interpretation of their results (15).

An alternative technique, which does not require modeling and is helpful for generating rules of high sensitivity, is **recursive partitioning**, or **Classification and Regression Tree (CART)** analysis. This technique creates a tree that asks a series of yes/no questions, taking the user down different branches depending on the answers. At the end of each branch will be an estimated probability of the outcome. The tree can be designed to have high sensitivity by instructing the software to make the penalty for false negatives higher than that for false positives. An example of such a tree, used to predict bacterial meningitis among adults with meningitis (16), is shown in Figure 12.2.

TABLE 12.4 EXAMPLE OF A CLINICAL PREDICTION RULE (FOR PULMONARY EMBOLISM) DERIVED FROM A LOGISTIC REGRESSION ANALYSIS (14)

CLINICAL CHARACTERISTIC	POINTS
Previous pulmonary embolism or deep vein thrombosis	+ 1.5
Heart rate >100 beats per minute	+ 1.5
Recent surgery or immobilization (within the last 30 days)	+ 1.5
Clinical signs of deep vein thrombosis	+ 3
Alternative diagnosis less likely than pulmonary embolism	+ 3
Hemoptysis (coughing blood)	+ 1
Cancer (treated within the last 6 mo)	+ 1
ESTIMATED CLINICAL PROBABILITY OF PULMONARY EMBOLISM (15)	**TOTAL SCORE**
Low (Probability ~1%–2%)	0–1
Intermediate (Probability ~16%)	2–6
High (Probability ~40%)	≥7



■ **FIGURE 12.2** Example of a Classification and Regression Tree to distinguish bacterial from viral meningitis in adults (16). White boxes serve to divide subjects into those at high risk of bacterial meningitis (red boxes) and those at low risk (green boxes); the numbers show the proportions with bacterial meningitis[5] in the red and green "terminal branches" of the tree.

Regardless of the method chosen to develop the rule, it is important that it be **validated** in a group of patients different from those in whom it was derived. One reason for this is to avoid **overfitting** (i.e., taking advantage of the tendency in a single sample for random error to increase the predictive strength of some factors). Overfitting can be addressed by dividing the cohort into **derivation** (typically 50% to 67% of the sample) and **validation** data sets, and testing the rule derived from the derivation cohort using data from the validation cohort. However, this validates the rule only in a population very similar to that from which it was derived (i.e., it addresses only internal validity). To address external validity, it is important to determine how well the rule performs in different populations ("prospective validation") (17).

■ STUDIES OF THE EFFECT OF TEST RESULTS ON CLINICAL DECISIONS

A test may be accurate, but if the disease is very rare, the test may be so seldom positive that it is hardly ever worth doing. Other tests may not affect clinical decisions because they do not provide new information beyond what was already known (e.g., from the medical history and physical examination). The study designs in this section address the **yield** of diagnostic tests and their effects on clinical decisions.

[5] The numbers in the figure include both derivation and validation data sets.

Types of Studies

- **Diagnostic yield studies.** Diagnostic yield studies address such questions as:
  - When a test is ordered for a particular indication, how often is it abnormal?
  - Can abnormal results be predicted from other information available at the time of testing?
  - In which group(s) of patients does the testing have the most or least value?
  - What happens to patients with abnormal results? Do benefits outweigh harms?

  Diagnostic yield studies estimate the proportion of positive tests among patients with a particular indication for the test. Unfortunately, showing that a test is often positive is not sufficient to indicate the test should be done. However, a diagnostic yield study showing a test is almost always negative may be sufficient to question its use for that indication.

  For example, Siegel et al. (18) studied the yield of stool cultures in hospitalized patients with diarrhea. Although not all patients with diarrhea receive stool cultures, it seems reasonable to assume that those who do are, if anything, more likely to have a positive culture than those who do not. Overall, only 40 (2%) of 1,964 stool cultures were positive. Moreover, none of the positive results were in the 997 patients who had been in the hospital for more than 3 days. Because a negative stool culture is unlikely to affect management in these patients with a low likelihood of bacterial diarrhea, the authors concluded that stool cultures are of little value among patients with diarrhea who have been in the hospital for more than 3 days.
- **Before/after studies of clinical decision-making.** These designs directly address the effect of a test result on clinical decisions. The design generally involves a comparison between what clinicians do (or say they would do) before and after obtaining results of a diagnostic test. For example, Carrico et al. (19) prospectively studied the value of abdominal ultrasound scans in 94 children with acute lower abdominal pain. They asked the clinicians requesting the sonograms to record their diagnostic impression and what their treatment would be if a sonogram were not available. After doing the sonograms and providing the clinicians with the results, they asked again. They found that sonographic information changed the initial treatment plan in 46% of patients.

  Of course (as discussed later), altering a clinical decision does not guarantee that a patient will benefit, and some altered decisions could actually be harmful. Studies that demonstrate effects on decisions are most useful when the natural history of the disease and the efficacy of treatment are clear. In the preceding example, there would very likely be a benefit from changing the decision from "discharge from hospital" to "laparoscopy" in children with appendicitis, or from "laparoscopy" to "observe" in children with nonspecific abdominal pain.

STUDIES OF FEASIBILITY, COSTS, AND RISKS OF TESTS

Another important area for clinical research relates to the practicalities of diagnostic testing. What proportion of patients will return a postcard with tuberculosis skin test results? What are the medical effects of false-positive screening tests in newborns, and the psychological effects on the parents? What proportion of colonoscopies are complicated by colonic perforation?

Design Issues

Studies of the **feasibility**, **costs**, and **risks** of tests are generally descriptive. The sampling scheme is important because tests often vary among the people or institutions doing them, and among the patients receiving them.

A straightforward choice is to study everyone who receives the test, as in a study of the return rate of postcards after tuberculosis skin testing. Alternatively, for some questions, the subjects in the study may be only those with results that were positive or falsely positive. For example, Bodegard et al. (20) studied families of infants who had tested falsely positive on a

newborn screening test for hypothyroidism and found that fears about the baby's health persisted for at least 6 months in almost 20% of the families.

Adverse effects can occur not just from false-positive results, but also from the testing itself. For example, Rutter et al. (21) employed an electronic medical record to do a retrospective cohort study of serious adverse events (perforation, hemorrhage, and acute diverticulitis) in the 30 days following colonoscopy among patients in the Group Health Cooperative of Puget Sound.

Analysis

Results of these studies can usually be summarized with simple descriptive statistics like means and standard deviations, medians, ranges, and frequency distributions. Dichotomous variables, such as the occurrence of adverse effects, can be summarized with proportions and their 95% confidence intervals (CIs). For example, in the aforementioned study Rutter et al. (21) reported perforations in 21/43,456 colonoscopies; this is 0.48 per 1,000 with a 95% confidence interval from 0.30 to 0.74 per 1,000.

There are generally no sharp lines that divide tests into those that are or are not feasible, or those that have or do not have an unacceptably high risk of adverse effects. For this reason it is helpful in the design stage of the study to specify criteria for deciding that the test is acceptable. What rate of follow-up would be insufficient? What rate of complications would be too high?

STUDIES OF THE EFFECT OF TESTING ON OUTCOMES

The best way to determine the value of a medical test is to see whether patients who are tested have a better clinical outcome (e.g., live longer or with better quality of life) than those who are not. Randomized trials are the ideal design for making this determination, but trials of diagnostic tests are often difficult to do. The value of tests is therefore usually estimated from observational studies. The key difference between the designs described in this section and the experimental and observational designs discussed elsewhere in this book is that the predictor variable for this section is *performing the test*, rather than a treatment, risk factor, or the result of a test.

Designs

Testing itself is unlikely to have any direct benefit on the patient's health. It is only when a test result leads to effective preventive or therapeutic interventions that the patient may benefit (22). Therefore, one important caveat about outcome studies of testing is that the predictor variable actually being studied is not just a test (e.g., a fecal occult blood test), but also all of the medical care that follows (e.g., procedures for following up abnormal results, colonoscopy, etc.).

It is best if the outcome variable of these studies is a measure of morbidity or mortality, not simply a diagnosis or stage of disease. For example, showing that men who are screened for prostate cancer have a greater proportion of cancers diagnosed at an early stage does not by itself establish the value of screening (23, 24). Many of those cancers would not have caused any problem if they had not been detected.

The outcome should be broad enough to include plausible adverse effects of testing and treatment, and may include psychological as well as medical effects of testing. Therefore, a study of the value of prostate-specific antigen screening for prostate cancer should include treatment-related impotence or incontinence in addition to cancer-related morbidity and mortality. When many more people are tested than are expected to benefit (as is usually the case), less severe adverse outcomes among those without the disease may be important, because they will occur much more frequently. While negative test results may be reassuring and comforting to some patients (25), in others the psychological effects of labeling or false-positive results, loss of insurance, and troublesome (but nonfatal) side effects of preventive medications or surgery may outweigh infrequent benefits (24).

- **Observational studies.** Observational studies are generally quicker, easier, and less costly than clinical trials. However, they have important disadvantages as well, especially because patients who are tested tend to differ from those who were not tested in important ways that may be related to the risk of a disease or its prognosis. For example, those getting the test could be at relatively *low* risk of an adverse health outcome, because people who volunteer for medical tests and treatments tend to be healthier than average, an example of **volunteer bias**. On the other hand, those tested may be at relatively *high* risk, because patients are more likely to be tested when there are indications that lead them or their clinicians to be concerned about a disease, an example of **confounding by indication** for the test (Chapter 9).

  An additional common problem with observational studies of testing is the lack of standardization and documentation of any interventions or changes in management that follow positive results. If a test does not improve outcome in a particular setting, it could be because follow-up of abnormal results was poor, because patients were not compliant with the planned intervention, or because the particular intervention used in the study was not ideal.
- **Clinical trials.** The most rigorous design for assessing the benefit of a diagnostic test is a clinical trial, in which subjects are randomly assigned to receive or not to receive the test. Presumably the result of the test is then used to guide clinical management. A variety of outcomes can be measured and compared in the two groups. **Randomized trials** minimize or eliminate confounding and selection bias and allow measurement of all relevant outcomes such as mortality, morbidity, cost, and satisfaction. Standardizing the testing and intervention process enables others to reproduce the results.

  Unfortunately, randomized trials of diagnostic tests are often **not practical**, especially for diagnostic tests already in use in the care of sick patients. Randomized trials are generally more feasible and important for tests that might be used in large numbers of apparently healthy people, such as new screening tests.

  Randomized trials, however, may bring up **ethical issues** about withholding potentially valuable tests. Rather than randomly assigning subjects to undergo a test or not, one approach to minimizing this ethical concern is to randomly assign some subjects to receive

EXAMPLE 12.1 An Elegant Observational Study of a Screening Test

Selby et al.(26) did a nested case–control study in the Kaiser Permanente Medical Care Program to determine whether screening sigmoidoscopy reduces the risk of death from colon cancer. They compared the rates of previous sigmoidoscopy among patients who had died of colon cancer with controls who had not. They found an adjusted odds ratio of 0.41 (95% CI, 0.25 to 0.69), suggesting that sigmoidoscopy resulted in an almost 60% decrease in the death rate from cancer of the rectum and distal colon.

A potential problem is that patients who undergo sigmoidoscopy may differ in important ways from those who do not, and that those differences might be associated with a difference in the expected death rate from colon cancer. To address this possible confounding, Selby et al. examined the apparent efficacy of sigmoidoscopy at preventing death from cancers of the proximal colon, above the reach of the sigmoidoscope. If patients who underwent sigmoidoscopy were less likely to die of colon cancer for other reasons, then sigmoidoscopy would appear to be protective against these cancers as well. However, sigmoidoscopy had no effect on mortality from cancer of the proximal colon (adjusted odds ratio = 0.96; 95% CI, 0.61 to 1.50), suggesting that confounding was not the reason for the apparent reduction in distal colon cancer mortality. Specifying alternate endpoints (in advance!) that are expected *not* to be associated with the predictor of interest (cancer of the proximal colon in this case), and then showing that they are not, can greatly strengthen causal inference (27).

an intervention that increases the use of the test, such as frequent postcard reminders and assistance in scheduling. The primary analysis must still follow the "**intention-to-treat**" rule—that is, the entire group that was randomized to receive the intervention must be compared with the entire comparison group. However, this rule will tend to create a conservative bias; the observed efficacy of the intervention will underestimate the actual efficacy of the test, because some subjects in the control group will get the test and some subjects in the intervention group will not. This problem can be addressed in secondary analyses that include testing rates in both groups and assume all the difference in outcomes between the two groups is due to different rates of testing. The actual benefits of testing in the subjects as a result of the intervention can then be estimated algebraically (8, 28).

Analysis

Analysis of studies of the effect of testing on outcome are those appropriate to the specific design used—odds ratios for case–control studies, and risk ratios or hazard ratios for cohort studies or clinical trials. A convenient way to express the results is to project the results of the testing procedure to a large cohort (e.g., 100,000), and list the number of initial tests, follow-up tests, people treated, side effects of treatment, costs, and deaths in tested and untested groups.

■ PITFALLS IN THE DESIGN OR ANALYSIS OF DIAGNOSTIC TEST STUDIES

As with other types of clinical research, compromises in the design of studies of diagnostic tests may threaten the validity of the results, and errors in analysis may hinder their interpretation. Some of the most common and serious of these, along with steps to avoid them, are outlined in the following text.

Inadequate Sample Size

If the outcome of a diagnostic test study is common, obtaining an adequate sample size is likely to be feasible. When the disease or outcome is rare, a very large number of people may be needed. Many laboratory tests, for example, are not expensive, and a yield of 1% or less might justify doing them, especially if they can diagnose a serious treatable illness. For example, Sheline and Kehr (29) retrospectively reviewed routine admission laboratory tests, including the Venereal Disease Research Laboratory (VDRL) test for syphilis among 252 psychiatric patients and found that the laboratory tests identified one patient with previously unsuspected syphilis. If this patient's psychiatric symptoms were indeed due to syphilis, it would be hard to argue that it was not worth the $3,186 spent on VDRLs to make this diagnosis. But if the true rate of unsuspected syphilis were close to the 0.4% seen in this study, a study of this sample size could easily have found no cases.

Inappropriate Exclusion

When calculating proportions, it is inappropriate to exclude subjects from the numerator without excluding similar subjects from the denominator. For example, in a study of routine laboratory tests in emergency department patients with new seizures (30), 11 of 136 patients (8%) had a correctable laboratory abnormality (e.g., hypoglycemia) as a cause for their seizure. In 9 of the 11 patients, however, the abnormality was suspected on the basis of the history or physical examination. The authors therefore reported that only 2 of 136 patients (1.5%) had abnormalities not suspected on the basis of the history or physical examination. But if all patients with suspected abnormalities are excluded from the numerator, then similar patients should have been excluded from the denominator as well. The correct denominator for this proportion

is therefore not all 136 patients tested, but only those who were not suspected of having any laboratory abnormalities on the basis of their medical history or physical examination.

Dropping Borderline or Uninterpretable Results

Sometimes a test may fail to give any answer at all, such as if the assay failed, the test specimen deteriorated, or the test result fell into a gray zone of being neither positive nor negative. It is not usually legitimate to ignore these problems, but how to handle them depends on the specific research question and study design. In studies dealing with the expense or inconvenience of tests, failed attempts to do the test are clearly important results.

Patients with "nondiagnostic" imaging studies or a borderline result on a test need to be counted as having had that specific result on the test. In effect, this may change a dichotomous test (positive, negative) to an ordinal one—positive, indeterminate and negative. ROC curves can then be drawn and likelihood ratios can be calculated for "indeterminate" as well as positive and negative results.

Verification Bias: Selective Application of a Single Gold Standard

A common sampling strategy for studies of medical tests is to study (either prospectively or retrospectively) patients who are tested for disease who also receive the gold standard for diagnosis. However, this causes a problem if the test being studied is also used to decide who gets the gold standard. For example, consider a study of predictors of fracture in children presenting to the emergency department with ankle injuries, in which only children who had ankle x-rays were included. If those with a particular finding (for example, ankle swelling) were more likely to get an x-ray, this could affect the sensitivity and specificity of ankle swelling as a test for fracture. This bias, called **verification bias**, is illustrated numerically in Appendix 12B. Verification bias can be avoided by using strict criteria for application of the gold standard that do not include the test or finding being studied. If this is not practical, it is possible to estimate and correct for verification bias if the gold standard can be applied to a random sample of those who test negative.

Differential Verification Bias: Different Gold Standards for Those Testing Positive and Negative

Another strategy is to use a different gold standard for those in whom the usual gold standard is not indicated. For example, subjects with ankle injuries in whom no x-ray was performed could be included by contacting them by telephone a few weeks after the injury and classifying them as not having had a fracture if they recovered uneventfully. However, this can cause **differential verification bias**, also called **double gold standard bias** (31). This bias can occur any time the gold standard differs among those with positive and negative test results. In the previously mentioned study of mammography (5) the gold standard for those with positive mammograms was a biopsy, whereas for those with negative mammograms, it was follow-up to see if a cancer became evident in the next year. Having different gold standards for the disease is a problem if the gold standards don't always have the same results, as would occur if breast cancer that would be detected by biopsy in the case of a positive mammogram would not become evident in the 1-year follow-up of those with a negative mammogram.

Another example is a study of ultrasonography to diagnose intussusception in young children (32). All children with a positive ultrasound scan for intussusception received the gold standard contrast enema. In contrast, the majority of children with a negative ultrasound were observed in the emergency department and intussusception was ruled out clinically. For cases of intussusception that resolve spontaneously, the two gold standards would give different results: the contrast enema would be positive, whereas clinical follow-up would be negative. A numerical illustration of differential verification bias in this study is provided in Appendix 12C.

Differential verification bias can be avoided by applying the same gold standard to all subjects. When this is not feasible (as was the case in the mammography study), investigators should make every effort to use other studies (e.g., autopsy studies examining the prevalence of asymptomatic cancers among patients who died from other causes in a study of a cancer screening test) to assess the degree to which this bias might threaten the validity of the study.

SUMMARY

1. The usefulness of **medical tests** can be assessed using designs that address a series of increasingly stringent questions (Table 12.1). For the most part, standard **observational designs** provide **descriptive statistics** of test characteristics with confidence intervals.
2. The **subjects** for a study of a diagnostic test should be chosen from patients who have a **spectrum** of disease and non-disease appropriate for the research question, in most cases reflecting the anticipated use of the test in clinical practice.
3. If possible, the investigator should **blind** those interpreting the test results and determining the gold standard from other information about the patients being tested.
4. Measuring the **reproducibility** of a test, including the **intra-** and **inter-observer variability**, is often a good first step in evaluating a test.
5. Studies of the **accuracy of tests** require a **gold standard** for determining if a patient has, or does not have, the disease or outcome being studied.
6. The results of studies of the accuracy of diagnostic tests can be summarized using **sensitivity**, **specificity**, **predictive value**, **ROC curves**, and **likelihood ratios**. Studies of the value of prognostic tests can be summarized with **risk ratios**, **hazard ratios**, or **reclassification improvement**.
7. Studies to develop **new clinical prediction rules** are subject to problems of **overfitting** and lack of **generalizability**, requiring that new rules be **validated** in additional population samples.
8. The most rigorous design for studying the utility of a diagnostic test is a **clinical trial**, with subjects randomized to receive the test or not, and with **mortality**, **morbidity**, **cost**, and **quality of life** among the outcomes.
9. If trials are not ethical or feasible, **observational studies** of **benefits**, **harms**, and **costs**, with appropriate attention to possible **biases** and **confounding**, can be helpful.

APPENDIX 12A
Calculation of Kappa to Measure Inter-Observer Agreement

Consider two observers listening for an S4 gallop on cardiac examination (Table 12A.1). They record it as either present or absent. The simplest measure of inter-observer agreement is the proportion of observations on which the two observers agree. This proportion can be obtained by summing the numbers along the diagonal from the upper left to the lower right and dividing it by the total number of observations. In this example, out of 100 patients there were 10 patients in whom both observers heard a gallop, and 75 in whom neither did, for (10 + 75)/100 = 85% agreement.

TABLE 12.A.1 INTER-OBSERVER AGREEMENT ON PRESENCE OF AN S4 GALLOP

	GALLOP HEARD BY OBSERVER 1	NO GALLOP HEARD BY OBSERVER 1	TOTAL, OBSERVER 2
Gallop heard by observer 2	10	5	15
No gallop heard by observer 2	10	75	85
Total, observer 1	20	80	100

When the observations are not evenly distributed among the categories (e.g., when the proportion "abnormal" on a dichotomous test is substantially different from 50%), or when there are more than two categories, another measure of inter-observer agreement, called *kappa* (κ), is sometimes used. Kappa measures the extent of agreement beyond what would be expected by chance alone, given the observed "*marginal values*" (i.e., the row and column totals). Kappa ranges from –1 (perfect disagreement) to 1 (perfect agreement). A kappa of 0 indicates that the amount of agreement was exactly that expected from the row and column totals. κ is estimated as:

$$\kappa = \frac{\text{Observed agreement (\%)} - \text{Expected agreement (\%)}}{100\% - \text{Expected agreement (\%)}}$$

The "expected" proportion in each cell is simply the proportion in that cell's row (i.e., the row total divided by the sample size) times the proportion in that cell's column (i.e., the column total divided by the sample size). The expected agreement is obtained by adding the expected proportions in the cells along the diagonal of the table, in which the observers agreed.

For example, in Table 12A.1 the observers appear to have done quite well: They have agreed 85% of the time. But how well did they do compared with agreement expected from their marginal totals? By chance alone (given the observed marginal values) they will agree about 71% of the time: (20% × 15%) + (80% × 85%) = 71%. Because the observed agreement was 85%, kappa is (85% –71%)/(100% –71%) = 0.48—respectable, if somewhat less impressive than 85% agreement.

When there are more than two categories of test results, it is important to distinguish between ordinal variables, which are intrinsically ordered, and nominal variables, which are not. For ordinal variables, kappa as calculated above fails to capture all the information in the data, because it does not give partial credit for coming close. To give credit for partial agreement, a *weighted* kappa should be used. (See Newman and Kohn [29] for a more detailed discussion.)

APPENDIX 12B
Numerical Example of Verification Bias

Consider two studies examining ankle swelling as a predictor of fractures in children with ankle injuries. The first study is a **consecutive sample** of 200 children. In this study, all children with ankle injuries are x-rayed, regardless of swelling. The sensitivity and specificity of ankle swelling are 80% and 75%, as shown in Table 12B.1:

TABLE 12B.1 ANKLE SWELLING AS A PREDICTOR OF FRACTURE USING A CONSECUTIVE SAMPLE

	FRACTURE	NO FRACTURE
Swelling	32	40
No swelling	8	120
Total	40	160
	Sensitivity = 32/40 = 80%	Specificity = 120/160 = 75%

The second study is a **selected** sample, in which only half the children without ankle swelling are x-rayed. Therefore, the numbers in the "No swelling" row will be reduced by half. This raises the apparent sensitivity from 32/40 (80%) to 32/36 (89%) and lowers the apparent specificity from 120/160 (75%) to 60/100 (60%), as shown in Table 12B.2.

TABLE 12B.2 VERIFICATION BIAS: ANKLE SWELLING AS A PREDICTOR OF FRACTURE USING A SELECTED SAMPLE

	FRACTURE	NO FRACTURE
Swelling	32	40
No swelling	4	60
Total	36	100
	Sensitivity = 32/36 = 89%	Specificity = 60/100 = 60%

APPENDIX 12C
Numerical Example of Differential Verification Bias

Results of the study by Eshed et al. of ultrasonography to diagnose intussusception (32) are shown in Table 12C.1.

TABLE 12C.1 RESULTS OF A STUDY OF ULTRASOUND DIAGNOSIS OF INTUSSUSCEPTION

	INTUSSUSCEPTION	NO INTUSSUSCEPTION
Ultrasound +	37	7
Ultrasound –	3	104
Total	40	111
	Sensitivity = 37/40 = 93%	Specificity = 104/111 = 94%

The 104 subjects with a negative ultrasound listed as having "No Intussusception" actually included 86 who were followed clinically and did not receive a contrast enema. If about 10% of these subjects (i.e., nine children) actually had an intussusception that resolved spontaneously, but that would still have been identified if they had a contrast enema, and all subjects had received a contrast enema, those nine children would have changed from true-negatives to false-negatives, as shown in Table 12C.2.

TABLE 12C.2 EFFECT ON SENSITIVITY AND SPECIFICITY IF NINE CHILDREN WITH SPONTANEOUSLY RESOLVING INTUSSUSCEPTION HAD RECEIVED THE CONTRAST ENEMA GOLD STANDARD INSTEAD OF CLINICAL FOLLOW-UP

	INTUSSUSCEPTION	NO INTUSSUSCEPTION
Ultrasound +	37	7
Ultrasound –	3 + 9 = 12	104 – 9 = 95
Total	49	102
	Sensitivity = 37/49 = 76%	Specificity = 95/102 = 93%

A similar, albeit less pronounced, effect occurs if some children with positive scans had intussusceptions that would have resolved spontaneously if given the chance (31).

REFERENCES

1. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1(8476):307–310.
2. Newman TB, Kohn M. *Evidence-based diagnosis*. New York: Cambridge University Press, 2009:10–38.
3. Tokuda Y, Miyasato H, Stein GH, et al. The degree of chills for risk of bacteremia in acute febrile illness. *Am J Med* 2005;118(12):1417.
4. Sawaya GF, Washington AE. Cervical cancer screening: which techniques should be used and why? *Clin Obstet Gynecol* 1999;42(4):922–938.

5. Smith-Bindman R, Chu P, Miglioretti DL, et al. Physician predictors of mammographic accuracy. *J Natl Cancer Inst* 2005;97(5):358–367.
6. Rocker G, Cook D, Sjokvist P, et al. Clinician predictions of intensive care unit mortality. *Crit Care Med* 2004;32(5):1149–1154.
7. Newman TB, Puopolo KM, Wi S, et al. Interpreting complete blood counts soon after birth in newborns at risk for sepsis. *Pediatrics* 2010;126(5):903–909.
8. Vittinghoff E, Glidden D, Shiboski S, et al. *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*, 2nd ed. New York: Springer, 2012.
9. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Ann Intern Med* 2009;150(11):795–802.
10. Cook NR. Assessing the incremental role of novel and emerging risk factors. *Curr Cardiovasc Risk Rep* 2010;4(2):112–119.
11. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27(2):157–172; discussion 207–212.
12. Shepherd JA, Kerlikowske K, Ma L, et al. Volume of mammographic density and risk of breast cancer. *Cancer Epidemiol Biomarkers Prev* 2011;20(7):1473–1482.
13. Grady D, Berkowitz SA. Why is a good clinical prediction rule so hard to find? *Arch Intern Med* 2011;171(19):1701–1702.
14. Wells PS, Anderson DR, Rodger M, et al. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer.*Thromb Haemost* 2000;83(3):416–420.
15. Wells PS, Anderson DR, Rodger M, et al. Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and d-dimer. *Ann Intern Med* 2001;135(2):98–107.
16. Tokuda Y, Koizumi M, Stein GH, et al. Identifying low-risk patients for bacterial meningitis in adult patients with acute meningitis. *Intern Med* 2009;48(7):537–543.
17. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA* 1997;277(6):488–494.
18. Siegel DL, Edelstein PH, Nachamkin I. Inappropriate testing for diarrheal diseases in the hospital. *JAMA* 1990;263(7):979–982.
19. Carrico CW, Fenton LZ, Taylor GA, et al. Impact of sonography on the diagnosis and treatment of acute lower abdominal pain in children and young adults. *American Journal of Roentgenology* 1999;172(2):513–516.
20. Bodegard G, Fyro K, Larsson A. Psychological reactions in 102 families with a newborn who has a falsely positive screening test for congenital hypothyroidism. *Acta Paediatr Scand Suppl* 1983;304:1–21.
21. Rutter CM, Johnson E, Miglioretti DL, et al. Adverse events after screening and follow-up colonoscopy. *Cancer Causes Control* 2012;23(2):289–296.
22. Etzioni DA, Yano EM, Rubenstein LV, et al. Measuring the quality of colorectal cancer screening: the importance of follow-up. *Dis Colon Rectum* 2006;49(7):1002–1010.
23. Welch HG. *Should I be tested for cancer? Maybe not, and here's why*. Berkeley, CA: University of California Press, 2004.
24. Welch HG, Schwartz LM, Woloshin S. *Overdiagnosed: making people sick in pursuit of health*. Boston, MA: Beacon Press, 2011.
25. Detsky AS. A piece of my mind. Underestimating the value of reassurance. *JAMA* 2012;307(10):1035–1036.
26. Selby JV, Friedman GD, Quesenberry CJ, et al. A case-control study of screening sigmoidoscopy and mortality from colorectal cancer [see comments]. *N Engl J Med* 1992;326(10):653–657.
27. Prasad V, Jena AB. Prespecified falsification end points: can they validate true observational associations? *JAMA* 2013;309(3):241–242.
28. Sheiner LB, Rubin DB. Intention-to-treat analysis and the goals of clinical trials. *Clin Pharmacol Ther* 1995;57(1):6–15.
29. Sheline Y, Kehr C. Cost and utility of routine admission laboratory testing for psychiatric inpatients. *Gen Hosp Psychiatry* 1990;12(5):329–334.
30. Turnbull TL, Vanden Hoek TL, Howes DS, et al. Utility of laboratory studies in the emergency department patient with a new-onset seizure. *Ann Emerg Med* 1990;19(4):373–377.
31. Newman TB, Kohn MA. *Evidence-based diagnosis*. New York: Cambridge University Press, 2009:101–102.
32. Eshed I, Gorenstein A, Serour F, et al. Intussusception in children: can we rely on screening sonography performed by junior residents? *Pediatr Radiol* 2004;34(2):134–137.

CHAPTER 13

Research Using Existing Data

Deborah G. Grady, Steven R. Cummings, and Stephen B. Hulley

Many research questions can be answered quickly and efficiently using data or specimens that have already been collected. There are three general approaches to using these existing resources. **Secondary data analysis** is the use of existing data to investigate research questions other than the main ones for which the data were originally gathered. **Ancillary studies** add one or more measurements to a study, often in a subset of the participants, to answer a separate research question. **Systematic reviews** combine the results of multiple previous studies of a given research question, often including calculation of a summary estimate of effect that has greater precision than the individual study estimates. Making creative use of existing data and specimens is a fast and effective way for new investigators with limited resources to begin to answer important research questions, gain valuable experience in a research area, and sometimes have a publishable finding in a short time frame.

ADVANTAGES AND DISADVANTAGES

The main **advantages** of studies using existing data are speed and economy. A research question that might otherwise require much time and money to investigate can sometimes be answered **rapidly** and **inexpensively**. For example, in the database of the Study of Osteoporotic Fractures, a prospective cohort study originally designed to study risk factors for fracture, Yaffe and colleagues used repeated measurements that had been made of physical activity and of cognitive function to discover that women who walked more had a 36% lower risk of cognitive decline than women who walked less (1).

Studies using existing data or specimens also have **disadvantages**. The selection of the population to study, which data to collect, the quality of data gathered, and how variables were measured and recorded are all predetermined. The existing data may have been collected from a population that is not ideal (e.g., men only rather than men and women), the measurement approach may not be what the investigator would prefer (history of hypertension, a dichotomous historical variable, in place of actual blood pressure), and the quality of the data may be poor (frequent missing or incorrect values). Important confounders and outcomes may not have been measured or recorded. All these factors contribute to the main disadvantage of using existing data: The investigator has little or **no control** over what data have been collected, and how.

SECONDARY DATA ANALYSIS

Secondary data sets may come from medical records, health care billing files, death certificates, public databases, and many other sources, but other **research studies**, either conducted at the investigator's institution or elsewhere, are one of the richest sources of secondary data. Many studies collect more data than the investigators analyze and these data can be used to document interesting results that have gone unnoticed. Access to such data is generally controlled by the study's **principal investigator** or a **steering committee**; the new researcher should therefore seek out information about studies by other investigators that may have made measurements relevant to the research question. One of the most important ways a good mentor can be helpful

to a new investigator is by providing knowledge of and access to relevant data sets. Most **NIH-funded** studies are required to make their data publicly available.

Other fruitful sources of secondary data are large regional and **national data sets** that are publicly available and do not have a principal investigator. Computerized databases of this sort are as varied as the reasons people have for collecting information. We will give several examples that deserve special mention, and readers can locate others in their own areas of interest.

- **Tumor registries** are government-supported agencies that collect complete statistics on cancer incidence, treatment, and outcome in defined geographic areas. These registries currently include about one quarter of the U.S. population, and the area of coverage is expected to increase during the coming years. One purpose of these registries is to provide data to outside investigators. Combined data for all the registries are available from the Surveillance, Epidemiology, and End Results (**SEER**) Program. For example, investigators used the SEER registry of breast cancer diagnoses to find that the annual incidence of estrogen-receptor positive breast cancer declined 13% in postmenopausal women between 2001 and 2003; this trend paralleled the reduction in use of hormone therapy by postmenopausal women, suggesting that stopping hormone therapy reduced the risk of breast cancer (2).
- **Death certificate registries** can be used to follow the mortality of any cohort. The **National Death Index** includes all deaths in the United States since 1978. This can be used to ascertain the vital status of subjects of an earlier study or of those who are part of another data set that includes important predictor variables. A classic example is the follow-up of men with coronary disease who were randomly assigned to high-dose nicotinic acid or placebo to lower serum cholesterol in the Coronary Drug Project. No study had ever shown an effect of lipid treatment on mortality and there was no difference in death rates at the end of the 5 years of randomized treatment, but a mortality follow-up 9 years later using the National Death Index revealed a significant benefit (3). Whether an individual is alive or dead is public information, so follow-up was available even for men who had dropped out of the study.

  The National Death Index can be used when either the Social Security number or the name and birth date are known. Ascertainment of the fact of death is 99% complete with this system, and additional information from death certificates (notably cause of death) can then be obtained from state records. On the state and local level, many jurisdictions now have computerized vital statistics systems, in which individual data (such as information from birth or death certificates) are entered as they are received.
- **NHANES,** the National Health and Nutrition Examination Survey is a series of surveys that assess the health and nutritional status of both adults and children in the United States. The surveys employ population-based cluster random selection to identify a nationally representative sample, and include self-reported data (e.g., demographic, socioeconomic, dietary, and health-related behaviors), physical examinations, laboratory tests, and other measurements. NHANES data can provide population-based estimates of disease prevalence, risk factors, and other variables. For example, bone mineral density (BMD) of the hip was measured during two examinations: 1988–1994 and 2005–2006. The results provide normal values for women and men of various races in the United States that are used to define 'osteoporosis' as 2.5 standard deviations below the average BMD value for young adults in NHANES (4). Investigators also used the repeated measurements to discover that BMD has been improving and the prevalence of osteoporosis has been declining (5).

Secondary data can be especially useful for studies to evaluate patterns of utilization and clinical outcomes of medical treatment. This approach can complement the information available from randomized trials and examine questions that trials cannot answer. These types of existing data include **electronic** administrative and clinical **databases** such as those developed by Medicare, the Department of Veterans Affairs, Kaiser Permanente Medical Groups, the Duke Cardiovascular Disease Databank, and **registries** such as the San Francisco Mammography Registry and the National Registry of Myocardial Infarction. Information from these sources

(many of which can be found on the Web) can be very useful for studying **rare adverse events** and for assessing real-world utilization and effectiveness of an intervention that has been shown to work in a clinical trial setting. For example, the National Registry of Myocardial Infarction was used to examine risk factors for intracranial hemorrhage after treatment with recombinant tissue-type plasminogen activator (tPA) for acute myocardial infarction (MI). The registry included 71,073 patients who received tPA; among these, 673 had intracranial hemorrhage confirmed by computed tomography or magnetic resonance imaging. A multivariate analysis showed that a tPA dose exceeding 1.5 mg/kg was significantly associated with developing an intracranial hemorrhage when compared with lower doses (6). Given that the overall risk of developing an intracranial hemorrhage was less than 1%, a clinical trial collecting primary data to examine this outcome would have been prohibitively large and expensive.

Another valuable contribution from this type of secondary data analysis is a better understanding of the difference between efficacy and effectiveness. The randomized clinical trial is the gold standard for determining the **efficacy** of a therapy in a select population under highly controlled circumstances in limited clinical settings. In the "real world," however, the patients who are treated, the choice of drugs and dosage by the treating physician, and adherence to medications by the patient are much more variable. These factors may make the application of therapy in the general population less effective than what is observed in trials. The **effectiveness** of treatments in actual practice can sometimes be studied using secondary data. For example, primary angioplasty has been demonstrated to be superior to thrombolytic therapy in clinical trials among patients with acute MI (7). But this may only be true when success rates for angioplasty are as good as those achieved in the clinical trial setting. Secondary analyses of community data sets have not found a benefit of primary angioplasty over thrombolytic therapy (8, 9). However, it is important to remember that observational studies of treatments have several limitations—most importantly potential confounding by differences in characteristics of those treated and those not treated. Bias and confounding are particularly difficult to assess using secondary databases that are not designed to study the effectiveness of treatments, and a randomized trial comparing treatments conducted in community settings is a better approach, when feasible.

Secondary data analysis is often the best approach for describing how therapies are used in clinical practice. Although clinical trials can demonstrate efficacy of a new therapy, this benefit can only occur if the therapy is adopted by practicing physicians. Understanding **utilization rates**, addressing **regional variation** and use in specific populations (such as the elderly, ethnic minorities, the economically disadvantaged, and women), can have useful public health implications. For example, using publicly available data from a 5% random sample of Medicare beneficiaries, investigators demonstrated substantial regional variation in the prevalence of diagnosed glaucoma after adjustment for potential confounders, suggesting over- or underdiagnosis in certain regions of the country (10).

Two or more existing data sets may also be linked to answer a research question. Investigators who were interested in how military service affects health used the 1970 to 1972 draft lottery involving 5.2 million 20-year-old men who were assigned eligibility for military service randomly by date of birth (the first data set) and subsequent mortality based on **death certificate registries** (the second source of data). The predictor variable (date of birth) was a randomly assigned proxy for military service during the Vietnam era. Men who had been randomly assigned to be eligible for the draft had significantly greater mortality from suicide and motor vehicle accidents in the ensuing 10 years (11). The study was done very inexpensively, yet it was a more unbiased approach to examining the effect of military service on specific causes of subsequent death than other studies of this topic with much larger budgets.

When individual data are not available, **aggregate data sets** can sometimes be useful. Aggregate data include information only for groups of persons (e.g., death rates from cervical cancer in each of the 50 states), not for individuals. With such data, associations can only be measured among these groups by comparing group information on a risk factor (such as tobacco sales by

region) with the rate of an outcome (lung cancer by region). Studies of associations based on aggregate data are called **ecologic studies**.

The advantage of aggregate data is its availability. Its major drawback is that associations are especially susceptible to confounding: Groups tend to differ from each other in many ways, not just with regard to the predictor variable of interest. As a result, associations observed in the aggregate do not necessarily hold for the individual. For example, sales of cigarettes may be greater in states with high suicide rates, but individuals who commit suicide may not be the ones doing most of the smoking. This situation is referred to as the **ecologic fallacy**. Aggregate data are most appropriately used to test the plausibility of a new hypothesis or to generate new hypotheses. Interesting results can then be pursued in another study that uses individual data.

Getting Started

After choosing a research topic and becoming familiar with the literature in that area (including a thorough literature search and advice from a senior mentor), the next step is to investigate whether the research question can be addressed with an existing data set. The help of a **senior colleague** can be invaluable in finding an appropriate data set. An experienced researcher has defined areas of interest in which he stays current and is aware of important data sets and the investigators who control these data, both at his own institution and elsewhere. This person can help identify and gain access to the appropriate data. Often, the research question needs to be altered slightly (by modifying the definition of the predictor or outcome variables, for example) to fit the available data.

The best solution may be close at hand, a database at the **home institution**. For example, a University of California, San Francisco (UCSF) fellow who was interested in the role of lipoproteins in coronary disease noticed that one of the few interventions known to lower the level of lipoprotein(a) was estrogen. Knowing that the Heart and Estrogen/Progestin Replacement Study (HERS), a major clinical trial of hormone treatment to prevent coronary disease, was managed at UCSF, the fellow approached the investigators with his interest. Because no one else had specifically planned to examine the relationship between this lipoprotein, hormone treatment, and coronary heart disease events, the fellow designed an analysis and publication plan. After receiving permission from the HERS study leadership, he worked with coordinating center statisticians, epidemiologists, and programmers to carry out an analysis that he subsequently published in a leading journal (12).

Sometimes a research question can be addressed that has little to do with the original study. For example, another fellow from UCSF was interested in the value of repeated screening Pap tests in women over 65 years old. He realized that the mean age of participants in the HERS trial was 67 years, that participants were required to have a normal Pap test to enter the trial, and that participants then underwent screening Pap tests annually during follow-up. By following up on Pap test outcomes, he was able to document that 110 Pap tests were abnormal among 2,763 women screened over a 2-year period, and that only one woman was ultimately found to have abnormal follow-up histology. Therefore, all but one of the abnormal Pap tests were falsely positive (13). This study strongly influenced the next U.S. Preventive Services Task Force recommendation that Pap tests should not be performed in low-risk women over age 65 with previous normal tests.

Sometimes it is necessary to venture **further afield**. Working from a list of predictor and outcome variables whose relation might help to answer the research question, an investigator can seek to locate databases that include these variables. Some studies have websites that provide free access to the study data without requiring permission. When the data are not available online, phone calls or e-mail messages to the authors of previous studies or to government officials might result in access to files containing useful data. It is essential to conquer any anxiety about contacting strangers to ask for help. Most people are surprisingly cooperative, either by providing data themselves or by suggesting other places to try.

Once the data for answering the research question have been located, the next challenge is to obtain **permission** to use them. It is a good practice to use official titles and your institutional domain name on correspondence or e-mail, and to copy your mentor as someone who will be recognized as an expert in the field. Young investigators should determine if their mentors are acquainted with the investigators who control the database, as an introduction may be more effective than a cold contact. It is generally most effective to work with an investigator, or a member of the study staff, who is interested in the research topic and involved in the study that has the data of interest. This investigator can facilitate access to the data, assure understanding of the study methods and how the variables were measured, and often becomes a valued colleague and collaborator. Data sets from multicenter studies and clinical trials generally have clear procedures for obtaining access to the data that include the requirement for a written proposal that must be approved by an analysis or publications committee.

The investigator should be very specific about what information is sought and confirm the request in writing. Many studies have guidelines for requesting data that specify what data are being requested, how the analyses will be done, and the timelines for completing the work. It is a good idea to keep the size of the request to a minimum and to offer to pay the cost of preparing the data. If the data set is controlled by a group of researchers, the investigator can suggest a collaborative relationship. In addition to providing an incentive to share the data, this can engage a co-investigator who is familiar with the database. It is wise to clearly define such a relationship early on, including who will be first author of the planned publications.

■ ANCILLARY STUDIES

Research using secondary data takes advantage of the fact that most of the data needed to answer a research question are already available. In an **ancillary study**, the investigator **adds** one or several **measurements** to an existing study to answer a different research question. For example, in the HERS trial of the effect of hormone therapy on risk for coronary events in 2,763 elderly women, an investigator added measurement of the frequency and severity of urinary incontinence. Adding a brief questionnaire at the next planned exam created a large trial of the effect of hormone therapy on urinary incontinence, with little additional time or expense (14).

Ancillary studies have many of the **advantages** of secondary data analysis with fewer constraints. They are both inexpensive and efficient, and the investigator can design a few key ancillary measurements specifically to answer the research question. Ancillary studies can be added to any type of study, including cross-sectional and case–control studies, but large prospective cohort studies and randomized trials are particularly well suited.

Ancillary studies have the problem that the measurements may be most informative when added before the study begins, and it may be difficult for an outsider to identify studies in the planning phase. Even when a variable was not measured at baseline, however, a single measurement during or at the end of a trial can produce useful information. By adding cognitive function measures at the end of the HERS trial, the investigators were able to compare the cognitive function of elderly women treated with hormone therapy for 4 years with the cognitive function of those treated with placebo (15).

A good opportunity for ancillary studies is provided by the banks of **stored sera**, **DNA**, **images**, and so on, that are found in most large clinical trials and cohort studies. The opportunity to propose new measurements using these stored specimens can be an extremely cost-effective approach to answering a novel research question, especially if it is possible to make these measurements on a subset of specimens using a nested case–control or case–cohort design (Chapter 8). In HERS, for example, a nested case–control study that carried out genetic analyses on stored specimens showed that the excess number of thromboembolic events in the hormone-treated group was not due to an interaction with factor V Leiden (16).

Getting Started

Opportunities for ancillary studies should be actively pursued, especially by new investigators who have limited time and resources. A good place to start is to identify studies with research questions that include either the predictor or the outcome variable of interest. For example, an investigator interested in the effect of weight loss on pain associated with osteoarthritis of the knee might start by identifying studies that include good measurement of painful osteoarthritis (by validated questionnaires) or databases with records of joint replacements that also have preceding measurements of weight. Additionally, the investigator may look for trials of interventions (such as diet, exercise, behavior change, or drugs) for weight loss. Such studies can be identified by searching lists of studies funded by the federal government (http://clinicaltrials.gov or http://report.nih.gov), by contacting pharmaceutical companies that manufacture drugs for weight loss, and by talking with experts in weight loss who are familiar with ongoing studies. To create an ancillary study, the investigator would simply add a measure of arthritis symptoms at a follow-up exam of subjects enrolled in these studies.

After identifying a study that provides a good opportunity for ancillary measures, the next step is to obtain the cooperation of the study investigators. Most researchers will consider adding brief ancillary measures to an established study if they address an important question and do not substantially interfere with the conduct of the main study. Investigators will be reluctant to add measures that require a lot of the participant's time (e.g., cognitive function testing) or are invasive and unpleasant (colonoscopy) or costly (positron emission tomography scanning).

Generally, formal permission from the principal investigator or the appropriate study committee is required to add an ancillary study. Most large, multicenter studies have established procedures requiring a written application. The proposed ancillary study is often reviewed by a committee that can approve, reject, or revise the ancillary study. Many ancillary measures require funding, and the ancillary study investigator must find a way to pay these costs. Of course, the marginal cost of an ancillary study is far less than the cost of conducting the same study independently. Ancillary studies are also very well suited for some types of NIH funding that provide only modest support for measurements and analyses but substantial support for career development (Chapter 19). Some large studies have their own mechanisms for funding ancillary studies, especially if the research question is important and considered relevant by the funding agency.

The **disadvantages** of ancillary studies are few. If the study will be collecting data from participants, new measures can be added, but variables already being measured generally cannot be changed. In some cases there may be practical problems in obtaining permission from the investigators or sponsor to perform the ancillary study, in training those who will make the measurements, or in obtaining separate informed consent from participants. These issues, including a clear understanding of authorship of scientific papers that result from the ancillary study and the rules governing their preparation and submission, need to be clarified before starting the study.

■ SYSTEMATIC REVIEWS

Systematic reviews identify a set of completed studies that address a particular research question, and evaluate the results of these studies to arrive at conclusions about a body of research. In contrast to other approaches to reviewing the literature, a systematic review uses a well-defined approach to identify all relevant studies, display the characteristics and results of eligible studies, and, when appropriate, calculate a summary estimate of the overall results. The **statistical aspects** of a systematic review (calculating summary effect estimates and variance, statistical tests of heterogeneity, and statistical estimates of publication bias) are called **meta-analysis**.

A systematic review can be a great opportunity for a new investigator. Although it takes a surprising amount of time and effort, a systematic review generally does not require substantial financial or other resources. Completing a good systematic review requires that the investigator

become intimately familiar with the literature on the research question. For new investigators, this detailed knowledge of published studies is invaluable. Publication of a good systematic review can also establish a new investigator as an "expert" on the research question. Moreover, the findings, with power enhanced by the larger sample size available from the combined studies and peculiarities of individual study findings revealed by comparison with the others, often represent an important scientific contribution. Systematic review findings can be particularly useful for developing **practice guidelines**.

The elements of a good systematic review are listed in Table 13.1. A good source of information on methods for conducting high-quality systematic reviews can be found in the *Cochrane Handbook for Systematic Reviews* (http://handbook.cochrane.org). Just as for other studies, the methods for completing each of these steps should be described in a written protocol before the systematic review begins.

The Research Question

A good systematic review has a well-formulated, clear research question that meets the **FINER** criteria (Chapter 2). Feasibility depends largely on the existence of a set of studies of the question. The research question should describe the disease or condition of interest, the population and setting, the intervention and comparison treatment (for trials), and the outcomes of interest. For example,

> "Among persons admitted to an intensive care unit with acute coronary syndrome, does treatment with aspirin plus intravenous heparin reduce the risk of myocardial infarction and death during the hospitalization more than treatment with aspirin alone?"

This research question led to a meta-analysis that found that adding aspirin to heparin improved outcomes, which was published in a top medical journal (17) and had an important impact on practice patterns.

Identifying Completed Studies

Systematic reviews are based on a comprehensive and unbiased **search** for completed studies. The search should follow a well-defined strategy established before the results of the individual studies are known. The process of identifying studies for potential inclusion in the review and the sources for finding such articles should be explicitly documented before the study. Searches should not be limited to MEDLINE, which may not list non-English-language references. Depending on the research question, electronic databases such as AIDSLINE, CANCERLIT, and EMBASE should be included, as well as manual review of the bibliography of relevant published studies, previous reviews, evaluation of the Cochran Collaboration database, and consultation with experts. The search strategy should be clearly described so that other investigators can replicate the search.

TABLE 13.1 ELEMENTS OF A GOOD SYSTEMATIC REVIEW

1. Clear research question
2. Comprehensive and unbiased identification of completed studies
3. Clear definition of inclusion and exclusion criteria
4. Uniform and unbiased abstraction of the characteristics and findings of each study
5. Clear and uniform presentation of data from individual studies
6. Calculation of a weighted summary estimate of effect and confidence interval based on the findings of all eligible studies when appropriate
7. Assessment of the heterogeneity of the findings of the individual studies
8. Assessment of potential publication bias
9. Subgroup and sensitivity analyses

Criteria for Including and Excluding Studies

The protocol for a systematic review should provide a good rationale for including and excluding studies, and these **criteria** should be established ***a priori*** (Table 13.2). Once these criteria are established, each potentially eligible study should be reviewed for eligibility independently by two or more investigators, with disagreements resolved by another reviewer or by consensus. When determining eligibility, it may be best to blind reviewers to the date, journal, authors, and results of trials.

Published systematic reviews should **list studies** that were considered for inclusion and the specific reason for excluding a study. For example, if 30 potentially eligible studies are identified, these 30 studies should be fully referenced and a reason should be given for each exclusion.

Collecting Data from Eligible Studies

Data should be abstracted from each study in a uniform and unbiased fashion. Generally, this is done **independently** by two or more abstractors using predesigned forms (Table 13.3). The data abstraction forms should include any data that will subsequently appear in the text, tables, or figures describing the studies included in the systematic review, or in tables or figures presenting the outcomes. When the two abstractors disagree, a third abstractor can settle the difference, or a consensus process may be used. The process for abstracting data from studies for the systematic review should be clearly described in the manuscript.

The published reports of some studies that might be eligible for inclusion in a systematic review may not include important information, such as design features, risk estimates, and standard deviations. Often it is difficult to tell if design features such as blinding were not implemented or were just not described in the publication. The reviewer can sometimes calculate

TABLE 13.2 CRITERIA FOR INCLUDING OR EXCLUDING STUDIES FROM META-ANALYSES

CRITERIA	EXAMPLE–OMEGA-3 FATTY ACIDS AND CARDIOVASCULAR EVENTS*
1. Period during which the studies were published	Studies published before August 2012
2. Study design	Randomized, controlled trials implemented in primary or secondary cardiovascular disease prevention settings
3. Study population	Studies of adults randomized to omega-3 fatty acids or control
4. Intervention or risk factor	Omega-3 fatty acid administration, either by diet or supplements, any dose, administered for at least one year
5. Acceptable control groups	A non-omega-3 fatty acid diet or supplement
6. Other study design requirements (e.g., blinding for trials or control for specific potential confounders for observational studies)	None
7. Acceptable outcomes	All-cause mortality, cardiac death, sudden death, myocardial infarction, and stroke
8. Maximal acceptable loss to follow-up	Not stated
9. Minimal acceptable length of follow-up	Not stated

*This example of how these criteria are used is drawn from a published meta-analysis showing no effect of omega-3 fatty acids in preventing cardiovascular disease events. (24)

TABLE 13.3 ELEMENTS TO INCLUDE ON DATA ABSTRACTION FORMS FOR META-ANALYSES

1. Eligibility criteria (how does the study meet pre-established eligibility criteria?)
2. Design features (study design, control group, blinding, control for confounding, etc.)
3. Characteristics and number of participants in each study group (demographics, illness severity, etc.)
4. Intervention (for trials) or risk factors (for observational studies).
   - For interventions—dose, duration of treatment, etc.
   - For observational studies—type and level of risk factor, etc.
5. Main outcome, secondary outcomes, and outcomes in pre-established subgroups
6. Elements to allow assessment of quality of included studies (randomization, blinding, adherence, loss to follow-up, control for confounding, etc.)

relative risks and confidence intervals from crude data presented from randomized *trials*, but it is generally unacceptable to calculate risk estimates and confidence intervals based on crude data from *observational* studies because there is not sufficient information to adjust for potential confounders. Every effort should be made to contact the authors to retrieve important information that is not included in the published description of a study. If this necessary information cannot be calculated or obtained, the study findings are generally excluded.

Presenting the Findings Clearly

Systematic reviews generally include three types of information. First, important **characteristics** of each study included in the systematic review are presented in tables. These often include characteristics of the study population, sample size, number or rate of outcomes, length of follow-up, and methods used in the study. Second, the review displays the **analytic findings** of the individual studies (relative risk, odds ratio, risk difference, and confidence intervals or *P* values) in a table or figure. Finally, in the absence of significant heterogeneity (see below), the meta-analysis presents **summary estimates and confidence intervals** based on the findings of all the included studies as well as sensitivity and subgroup analyses.

The summary effect estimates represent a main outcome of the meta-analysis, but should be presented in the context of all the information abstracted from the individual studies. The characteristics and findings of individual studies included in the systematic review should be displayed clearly in tables and figures so that the reader can form opinions that do not depend solely on the statistical summary estimates.

Meta-Analysis: Statistics for Systematic Reviews

- **Summary effect estimate and confidence interval**. Once all completed studies have been identified, those that meet the inclusion and exclusion criteria have been chosen, and data have been abstracted from each study, a **summary estimate** (summary relative risk, summary odds ratio, summary risk difference, etc.) and **confidence interval** are generally calculated. The summary effect is essentially an average effect weighted by the inverse of the variance of the outcome of each study. Methods for calculating the summary effect and confidence interval are discussed in Appendix 13. Those not interested in the details of calculating mean weighted estimates from multiple studies should at least be aware that different approaches can give different results. For example, recent meta-analyses of the effectiveness of condoms for preventing heterosexual transmission of HIV have given summary estimates ranging from 80% to 94% decrease in transmission rates, although they are based on the results of almost identical sets of studies (18, 19).
- **Heterogeneity**. Combining the results of several studies is not appropriate if the studies differ in clinically important ways, such as the population, intervention, outcome, control

condition, blinding, and so on. It is also inappropriate to combine the findings if the results of the individual studies differ widely. Even if the methods used in the studies appear to be similar, the fact that the results vary markedly suggests that something important was different in the individual studies. This variability in the findings of the individual studies is called heterogeneity (and the study findings are said to be **heterogeneous**); if there is little variability, the study findings are said to be **homogeneous**.

How can the investigator decide whether methods and findings are similar enough to combine into summary estimates? First, he can review the individual studies to determine if there are substantial differences in study design, study populations, intervention, or outcome. Then he can examine the results of the individual studies. If some trials report a substantial beneficial effect of an intervention and others report considerable harm, heterogeneity is clearly present. Sometimes, it is difficult to decide if heterogeneity is present. For example, if one trial reports a 50% risk reduction for a specific intervention but another reports only a 30% risk reduction, is heterogeneity present? Statistical approaches (tests of homogeneity) have been developed to help answer this question (Appendix 13), but ultimately, the **assessment of heterogeneity** requires judgment. Every reported systematic review should include some discussion of heterogeneity and its effect on the summary estimates.

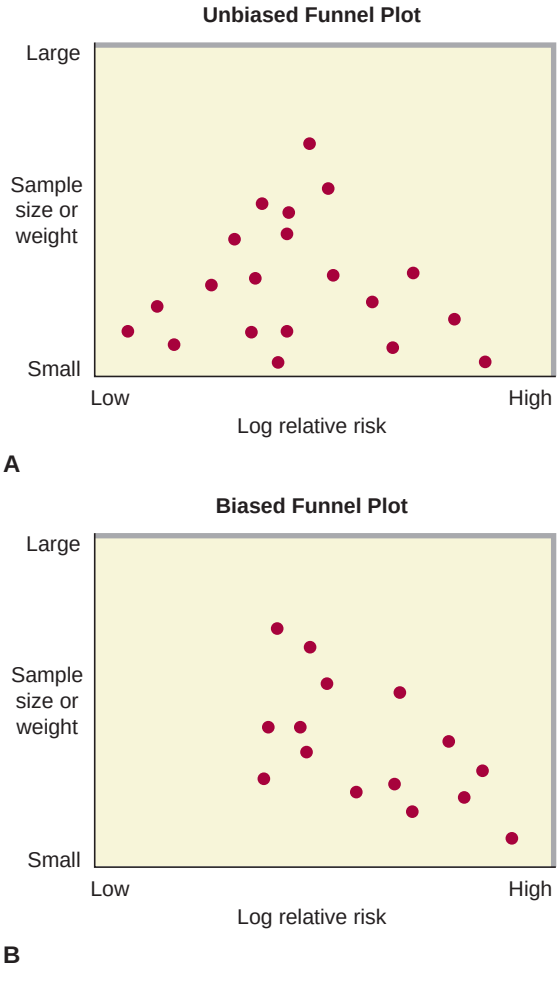
Assessment of Publication Bias

Publication bias occurs when published studies are not representative of all studies that have been done, usually because positive results tend to be submitted and published more often than negative results. There are two main ways to deal with publication bias. **Unpublished studies** can be identified and the results included in the summary estimate. Unpublished results may be identified by querying investigators and reviewing abstracts, meeting presentations, and doctoral theses. The results of unpublished studies can be included with those of the published studies in the overall summary estimate, or sensitivity analyses can determine if adding these unpublished results substantially changes the summary estimate determined from published results. However, including unpublished results in a systematic review is problematic for several reasons. It is often difficult to identify unpublished studies and even more difficult to abstract the required data. Frequently, inadequate information is available to determine if the study meets inclusion criteria for the systematic review or to evaluate the quality of the methods (which, lacking the rigor of peer review, may be inferior). For these reasons, unpublished data are not often included in meta-analyses.

Alternatively, the extent of potential publication bias can be estimated and this information used to temper the conclusions of the systematic review. Publication bias exists when unpublished studies have different findings from published studies. Unpublished studies are more likely to be small (large studies usually get published, regardless of the findings) and to have found no association between the risk factor or intervention and the outcome (markedly positive studies usually get published, even if small). If there is no publication bias, there should be no association between a study's size (or the variance of the outcome) and the findings. The degree of this association is often measured using **Kendall's Tau**, a coefficient of correlation. A strong or statistically significant correlation between study outcome and sample size suggests publication bias. In the absence of publication bias, a plot of study sample size versus outcome (e.g., log relative risk) should have a bell or **funnel shape** with the apex near the summary effect estimate.

The funnel plot in Figure 13.1A suggests that there is little publication bias because small studies with both negative and positive findings were published. The plot in Figure 13.1B, on the other hand, suggests publication bias because the distribution appears truncated in the corner that should contain small, negative studies.

When substantial publication bias is likely, summary estimates should not be calculated or should be interpreted cautiously. Every reported systematic review should include some discussion of potential publication bias and its effect on the summary estimates.



■ **FIGURE 13.1** **A:** Funnel plot that does not suggest publication bias because there are studies with a range of large and small sample sizes, and low relative risks are reported by some smaller studies. **B:** Funnel plot suggestive of publication bias because few of the smaller studies report low relative risks.

Subgroup and Sensitivity Analyses

Subgroup analyses may be possible using data from all or some subset of the studies included in the systematic review. For example, in a systematic review of the effect of postmenopausal estrogen therapy on endometrial cancer risk, some of the studies presented the results by duration of estrogen use. Subgroup analyses of the results of studies that provided such information demonstrated that longer duration of use was associated with higher risk for cancer (20).

Sensitivity analyses indicate how "sensitive" the findings of the meta-analysis are to certain decisions about the design of the systematic review or inclusion of certain studies. For example, if the authors decided to include studies with a slightly different design or methods in the systematic review, the findings are strengthened if the summary results are similar whether or not the questionable studies are included. Systematic reviews should generally include sensitivity analyses if any of the design decisions appear questionable or arbitrary.

Meta-analyses can increase the **power** to answer a research question, but have the disadvantage that they do not include individual-level data to allow adjustment for potential confounding or to perform individual subgroup analyses. In some situations, it may be possible to obtain the individual-level data from the relevant individual studies and perform **pooled analyses**. In these cases, the pooled data from individual studies can be used to adjust for confounding or

assess subgroup effects just as would be done in a large single study. For example, the Early Breast Cancer Trialists Collaborative Group pooled individual-level data from 123 randomized trials to evaluate the efficacy of different chemotherapy regimens for early breast cancer (21). However, it is generally difficult to obtain individual-level data from relevant studies, and uncommon that these studies have measured variables in ways that are similar enough to be combined into one data set.

Garbage In, Garbage Out

The biggest drawback to a systematic review is that it can produce a reliable-appearing summary estimate based on the results of individual studies that are of poor quality. There are several approaches used to assess the quality of different study designs in meta-analyses, but the process of assessing quality is complex and problematic. We favor relying on relatively strict criteria for good study design when setting the inclusion criteria. If the individual studies that are summarized in a systematic review are of poor quality, no amount of careful analysis can prevent the summary estimate from being unreliable. A special instance of this problem is encountered in systematic reviews of observational data. If the results of these studies are not adjusted for potential confounding variables, the results of the meta-analysis will also be unadjusted and potentially confounded.

SUMMARY

This chapter describes three approaches to making creative use of existing data and specimens, a fast and effective way for new investigators with limited resources to acquire valuable experience and an early publication.

Secondary Data Analysis

1. This approach to using existing data sets has the **advantage** of greatly reducing the time and cost of doing research and the **disadvantage** of providing little or no control over the study population, design, or measurements.
2. Sources of data for secondary analysis include **existing research projects**, **electronic medical records**, **administrative databases** and public databases such as **tumor registries**, **death certificate registries**, and national surveys such as **NHANES**.
3. Large community-based data sets are useful for studying **effectiveness** (the real-world effects of an intervention in various communities); for assessing **utilization rates** and **regional variation**, and for discovering **rare adverse events**.
4. Studies of associations based on **aggregate data** are called **ecological studies**; these can provide useful information but are subject to special biases termed **ecological fallacies**.

Ancillary Study

1. An ancillary study is a secondary data analysis in which the investigator makes one or more **new measurements** to answer a new research question with relatively **little cost and effort**.
2. Good opportunities for ancillary studies may be found in **cohort studies** or **clinical trials** that include either the predictor or outcome variable for the new research question.
3. **Stored serum**, **DNA**, **images**, and so on, provide the opportunity for nested case–control designs.
4. Most large studies have written **policies** that allow investigators (including outside scientists) to propose and carry out secondary data analyses and ancillary studies.

Systematic Review

1. A good systematic review, like any other study, requires a **written protocol** before the study begins that includes the **research question**, methods for **identifying all eligible studies**, methods for **abstracting data** from the studies, and **statistical methods**.
2. The statistical aspects of combining studies on a topic, termed **meta-analysis**, include the **summary effect estimate and confidence interval**, tests for evaluating **heterogeneity** and potential **publication bias**, and **subgroup** and **sensitivity analyses**.
3. The **characteristics** and **findings** of individual studies should be displayed clearly in tables and figures so that the reader can form opinions that do not depend solely on the statistical summary estimates.
4. A major challenge is assessing **quality** of the studies in a systematic review, which can strongly influence the findings of the review.

APPENDIX 13
Statistical Methods for Meta-Analysis

SUMMARY EFFECTS AND CONFIDENCE INTERVALS

The primary goal of meta-analysis is to calculate a **summary effect estimate** and confidence interval. An intuitive way to do this is to multiply each study outcome, such as the relative risk (an effect estimate), by the sample size (a weight that reflects the precision of the relative risk), add these products, and divide by the sum of the weights. In actual practice, the inverse of the variance of the effect estimate from each individual study ($1/\text{variance}_i$) is used as the weight for each study. The inverse of the variance is a better estimate of the precision of the effect estimate than the sample size because it takes into account the number of outcomes and their distribution. The weighted mean effect estimate is calculated by multiplying each study weight ($1/\text{variance}_i$) by the log of the relative risk (or any other risk estimate, such as the log odds ratio, risk difference, etc.), adding these products, and dividing by the sum of the weights. Small studies generally result in a large variance (and a wide confidence interval around the risk estimate) and large studies result in a small variance (and a narrow confidence interval around the risk estimate). Therefore, in a meta-analysis, large studies get a lot of weight (1/small variance) and small studies get little weight (1/big variance).

To determine if the summary effect estimate is statistically significant, the variability of the estimate of the summary effect is calculated. There are various formulas for calculating the variance of summary risk estimates (22, 23). Most use something that approximates the inverse of the sum of the weights of the individual studies ($1/\Sigma\ \text{weight}_i$). The variance of the summary estimate is used to calculate the 95% confidence interval around the summary estimate ($\pm\ 1.96 \times \text{variance}^{1/2}$).

RANDOM- VERSUS FIXED-EFFECTS MODELS

There are multiple statistical approaches available for calculating a summary estimate (22, 23). The choice of statistical method is usually dependent on the type of outcome (relative risk, odds ratio, risk difference, etc.). In addition to the statistical method, the investigator must also choose to use either a fixed-effects or random-effects model. The **fixed-effects** model simply calculates the variance of a weighted summary estimate based on the inverse of the sum of the weights of each individual study. The **random-effects** model adds variance to the summary effect in proportion to the variability of the results of the individual studies. Summary effect estimates are generally similar using either the fixed- or random-effects model, but the variance of the summary effect is greater in the random-effects model to the degree that the results of the individual studies differ, and the confidence interval around the summary effect is correspondingly larger, so that summary results are less likely to be statistically significant. Many journals require authors to use a random-effects model because it is considered "conservative" (i.e., less likely to find a statistically significant effect if one does not exist). Meta-analyses should state clearly whether they used a fixed- or random-effects model.

Simply using a random-effect model does not obviate the problem of heterogeneity. If the studies identified by a systematic review are clearly heterogeneous, a summary estimate should not be calculated.

STATISTICAL TESTS OF HOMOGENEITY

Tests of homogeneity assume that the findings of the individual trials are the same (the null hypothesis) and use a statistical test (test of homogeneity) to determine if the data (the individual study findings) refute this hypothesis. A chi-squared test is commonly used (22). If the data do support the null hypothesis (P value ≥0.10), the investigator accepts that the studies are homogeneous. If the data do not support the hypothesis (P value <0.10), he rejects the null hypothesis and assumes that the study findings are heterogeneous. In other words, there are meaningful differences in the populations studied, the nature of the predictor or outcome variables, or the study results.

All meta-analyses should report tests of homogeneity with a P value. These tests are not very powerful and it is hard to reject the null hypothesis and prove heterogeneity when the sample size—the number of individual studies—is small. For this reason, a P value of 0.10 rather than 0.05 is typically used as a cutoff. If substantial heterogeneity is present, it is inappropriate to combine the results of trials into a single summary estimate.

REFERENCES

1. Yaffe K, Barnes D, Nevitt M, et al. A prospective study of physical activity and cognitive decline in elderly women: women who walk. *Arch Intern Med* 2001;161:1703–1708.
2. Kerlikowske K, Miglioretti D, Buist D, et al. Declines in invasive breast cancer and use of postmenopausal hormone therapy in a screening mammography population. *J Natl Cancer Inst.* 2007;99:1335–1339.
3. Canner PL. Mortality in CDP patients during a nine-year post-treatment period. *J Am Coll Cardiol* 1986;8:1243–1255.
4. Looker AC, Johnston CC Jr., Wahner HW, et al. Prevalence of low femoral bone density in older U.S. women from NHANES III. *J Bone Miner Res* 1995;10:796–802.
5. Looker AC, Melton LJ, Harris TB, et al. Prevalence and trends in low femur bone density among older US adults: NHANES 2005–2006 compared with NHANES III. *J Bone Miner Res* 2010;25:64–71.
6. Gurwitz JH, Gore JM, Goldberg RJ, et al. Risk for intracranial hemorrhage after tissue plasminogen activator treatment for acute myocardial infarction. Participants in the National Registry of Myocardial Infarction 2. *Ann Intern Med* 1998;129:597–604.
7. Weaver WD, Simes RJ, Betriu A, et al. Comparison of primary coronary angioplasty and intravenous thrombolytic therapy for acute myocardial infarction: a quantitative review. *JAMA* 1997;278:2093–2098; published erratum appears in *JAMA* 1998;279:876.
8. Every NR, Parsons LS, Hlatky M, et al. A comparison of thrombolytic therapy with primary coronary angioplasty for acute myocardial infarction. Myocardial infarction triage and intervention investigators. *N Engl J Med* 1996;335:1253–1260.
9. Tiefenbrunn AJ, Chandra NC, French WJ, et al. Clinical experience with primary percutaneous transluminal coronary angioplasty compared with alteplase (recombinant tissue-type plasminogen activator) in patients with acute myocardial infarction: a report from the Second National Registry of Myocardial Infarction (NRMI–2). *J Am Coll Cardiol* 1998;31:1240–1245.
10. Cassard SD, Quigley HA, Gower EW, et al. Regional variations and trends in the prevalence of diagnosed glaucoma in the Medicare population. *Ophthalmology* 2012;119:1342–1351.
11. Hearst N, Newman TB, Hulley SB. Delayed effects of the military draft on mortality: a randomized natural experiment. *N Engl J Med* 1986;314:620–624.
12. Shlipak M, Simon J, Vittinghoff E, et al. Estrogen and progestin, lipoprotein (a), and the risk of recurrent coronary heart disease events after menopause. *JAMA* 2000;283:1845–1852.
13. Sawaya GF, Grady D, Kerlikowske K, et al. The positive predictive value of cervical smears in previously screened postmenopausal women: the Heart and Estrogen/Progestin Replacement Study (HERS). *Ann Intern Med* 2000;133:942–950.
14. Grady D, Brown J, Vittinghoff E, et al. Postmenopausal hormones and incontinence: the Heart and Estrogen/Progestin Replacement Study. *Obstet Gynecol* 2001;97:116–120.
15. Grady D, Yaffe K, Kristof M, et al. Effect of postmenopausal hormone therapy on cognitive function: the Heart and Estrogen/Progestin Replacement Study. *Am J Med* 2002;113:543–548.
16. Herrington DM, Vittinghoff E, Howard TD, et al. Factor V Leiden, hormone replacement therapy, and risk of venous thromboembolic events in women with coronary disease. *Arterioscler Thromb Vasc Biol* 2002;22:1012–1017.
17. Oler A, Whooley M, Oler J, et al. Heparin plus aspirin reduces the risk of myocardial infarction or death in patients with unstable angina. *JAMA* 1996;276:811–815.

18. Pinkerton SD, Abramson PR. Effectiveness of condoms in preventing HIV transmission. *Soc Sci Med* 1997;44:1303–1312.
19. Weller S, Davis K. Condom effectiveness in reducing heterosexual HIV transmission. *Cochrane Database Syst Rev* 2002;(1):CD003255.
20. Grady D, Gebretsadik T, Kerlikowske K, et al. Hormone replacement therapy and endometrial cancer risk: a meta-analysis. *Obstet Gynecol* 1995;85:304–313.
21. Peto R, Davies C, Godwin J, et al. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. Early Breast Cancer Trialists' Collaborative Group. *Lancet* 2012;379:432–441.
22. Petitti DB. *Meta-analysis, decision analysis and cost effectiveness analysis: methods for quantitative synthesis in medicine,* 2nd ed. New York: Oxford University Press, 2000.
23. Cooper H, Hedges LV. *The handbook of research synthesis*. New York: Russell Sage Foundation, 1994.
24. Rizos EC, Ntzani EE, Bika E, et al. Association between omega-3 fatty acid supplementation and risk of major cardiovascular disease events. *JAMA* 2012;308:1024–1033.

SECTION III

Implementation

CHAPTER 14

Addressing Ethical Issues

Bernard Lo and Deborah G. Grady

Research with human participants raises ethical concerns because people accept inconvenience and risks to advance scientific knowledge and to benefit others. The public, who participate in and help fund clinical research, needs to trust that research follows high ethical standards.

In this chapter we begin with the **history** of research oversight and then review **ethical principles** and **federal regulations** guiding research with human participants, especially requirements for **institutional review board (IRB)** approval and **informed consent**. We finally turn to issues of **scientific misconduct**, **authorship**, **conflicts of interest**, and ethical issues in specific types of research.

HISTORY OF REGULATIONS ON CLINICAL RESEARCH

Current regulations and guidelines for clinical research have responded to abuses,including Nazi physician "research" during World War II, research in the U.S. on prisoners, residents of long-term care facilities and other vulnerable populations, and the Tuskegee Study (Case 14.1).

CASE 14.1 The Tuskegee Study (1)

In 1932 U.S. government agencies started the Tuskegee study to document the natural history and long-term effects of untreated syphilis. Subjects were impoverished, poorly educated African American men in rural Alabama. They received meals, some basic medical care, and burial insurance. Researchers falsely told the subjects that they were receiving treatment for syphilis, for example misrepresenting lumbar punctures done for research purposes as "special free treatments." When antibiotics for syphilis became available during World War II and later recommended as a public health measure, researchers took steps to keep subjects from receiving treatment. In response to the Tuskegee study, in 1974 the federal government issued regulations on human subjects research, which required informed consent from subjects and review by IRBs for federally sponsored human subjects research. In 1997, President Clinton formally apologized for the Tuskegee study.

ETHICAL PRINCIPLES

Ethical lapses from the Tuskegee study and several others inspired current regulations for the protection of research participants. Three ethical principles, which had been violated in these studies, were articulated to guide research with human participants (2). First, recognizing that all persons have the right to make their own decisions about research participation, the principle of **respect for persons** requires investigators to obtain informed and voluntary consent

from research participants, to allow them to discontinue participation in research at any time, and to protect participants with impaired decision-making capacity.

Second, the principle of **beneficence** requires that the scientific knowledge to be gained from the study must outweigh the inconvenience and risk experienced by research participants, and that risks be minimized. Risks include both physical harm from research interventions and also psychosocial harm, such as breaches of confidentiality, stigma, and discrimination. The risks of participating in the study can be reduced, for example, by screening potential participants to exclude those likely to suffer harm, ensuring confidentiality and monitoring participants for adverse effects.

Third, the principle of **justice** requires that the benefits and burdens of research be distributed fairly. Disadvantaged and vulnerable populations, such as people with low income, limited education, poor access to health care, or impaired decision-making capacity, should not be selectively targeted as participants if other populations would also be suitable to address the research questions. Studying vulnerable groups primarily because of easy access, cooperation, and follow-up takes unfair advantage of them.

Justice also requires equitable access to the benefits of research. Traditionally, clinical research has been regarded as risky, and potential participants have been thought of as guinea pigs that needed protection from dangerous interventions that would confer little or no personal benefit. Increasingly, however, clinical research is regarded as providing access to new therapies for such conditions as HIV infection and cancer. Patients who seek promising new drugs for fatal conditions want increased access to clinical research, not greater protection, and such access should be available regardless of income, insurance, or education. Children, women, and members of ethnic minorities historically have been under-represented in clinical research, resulting in a weak evidence base and potentially suboptimal clinical care. The principle of justice requires that these groups be included in research studies. NIH-funded clinical researchers must have adequate representation of children, women, and members of ethnic minorities in studies, or justify why these groups might be under-represented.

FEDERAL REGULATIONS FOR RESEARCH ON HUMAN SUBJECTS

Federal regulations apply to all federally funded research and to research that will be submitted to the U.S. Food and Drug Administration (FDA) in support of a new drug or device application. In addition, universities require that all research on human participants conducted by affiliated faculty and staff comply with core regulations regarding informed consent and IRB review, including research funded privately or conducted off-site. Although the regulations refer to human "subjects," the term "**participants**" is preferred by some because it emphasizes that people are active participants in research, rather than subjects to be experimented upon.

Several **definitions** in these regulations are important to understand:

- **Research** is "systematic investigation designed to develop or contribute to generalizable knowledge" (3). Unproven clinical care that is directed toward benefiting the individual patient and not toward publication is not considered research. Some quality improvement projects might be treated as research, although most meet criteria for exemption, which we discuss later.
- **Human subjects** are living individuals about whom an investigator obtains either "data through intervention or interaction with the individual" or "identifiable private information."
- **Private information** comprises (1) information that a person can reasonably expect is not being observed or recorded and (2) information that has been provided for specific purposes and that "the individual can reasonably expect will not be made public (e.g., a medical record)." Information is identifiable if "the identity of the subject is or may be readily ascertained by the investigator."
- Coded research data are **not identifiable** if the key that links data to participants is destroyed before the research begins or if the investigators have no access to the key.

The Federal Regulations on the Protection of Human Subjects are available on the website of the Office for Human Research Protections (3). Researchers who have questions about these federal regulations should consult their IRB. These federal regulations provide two main protections for human participants—IRB approval and informed consent.

Institutional Review Board (IRB) Approval

Federal regulations require that research with human participants be approved by an **IRB**. The IRB mission is to ensure that the research is ethically acceptable and that the welfare and rights of research participants are protected. Although most IRB members are researchers, IRBs must also include community members and persons knowledgeable about legal and ethical issues concerning research.

When approving a research study, the IRB must determine that (3):

- **Risks** to participants are **minimized**
- **Risks** are **reasonable** in relation to anticipated benefits and the importance of the knowledge that is expected to result
- **Selection** of participants is **equitable**
- **Informed consent** will be obtained from participants or their legally authorized representatives
- **Confidentiality** is adequately maintained

The IRB system is decentralized. Each local IRB implements federal regulations using its own forms, procedures, and guidelines, and there is no appeal to a higher body. As a result, a multicenter study might be approved by one IRB but not by other IRBs. Usually these differences can be resolved through discussions or protocol modifications.

IRBs and federal regulations have been criticized for several reasons (4, 5). They might place undue emphasis on consent forms, fail to scrutinize the research design, and not adequately consider the scientific merit of the research. Although IRBs need to review any protocol revisions and monitor adverse events, typically they do not check whether research was actually carried out in accordance with the approved protocols. Many IRBs lack the resources and expertise to adequately fulfill their mission of protecting research participants. For these reasons, federal regulations and IRB approval should be regarded only as a minimal ethical standard for research. Ultimately, the **judgment and character of the investigator** are the most essential element for assuring that research is ethically acceptable.

Exceptions to Full IRB Review

- Most research using surveys and interviews, as well as secondary analyses of de-identified existing records and specimens may be **exempted** from IRB review (Table 14.1). The ethical justification for such exemptions is that the research involves low risk, almost all people would consent to such research, and obtaining consent from each participant would make

TABLE 14.1 RESEARCH THAT IS EXEMPT FROM FEDERAL RESEARCH REGULATIONS

1. Surveys, interviews, or observations of public behavior unless:
   - Participants can be identified *and*
   - Disclosure of participants' responses could place them at risk for legal liability or damage their reputation, financial standing, or employability. For example, questionnaires on such issues as drug addiction, depression, HIV risk behaviors, or illegal immigration are not exempt.
2. Studies of existing records, data, or specimens, provided that data are:
   - Publicly available (e.g., data sets released by state and federal agencies) OR
   - Recorded by the investigator in such a manner that participants cannot be identified, for example, because the investigator cannot obtain the key to the code.
3. Research on normal educational practices.

TABLE 14.2 RESEARCH THAT MAY UNDERGO EXPEDITED IRB REVIEW

1. Certain minimal risk research procedures, including:
   - Collection of specimens through venipuncture, saliva or sputum collection, or skin or mucosal swabs.
   - Collection of specimens through noninvasive procedures routinely employed in clinical practice, such as electrocardiograms and magnetic resonance imaging. X-rays, which expose participants to radiation, require full IRB review.
   - Research involving data, records, or specimens that have been collected or will be collected for clinical purposes.
   - Research using surveys or interviews that is not exempt from IRB review.
2. Minor changes in previously approved research protocols.
3. Renewal of IRB approval for studies that are complete except for data analysis or long-term follow-up.

such studies prohibitively expensive or difficult. Many IRBs, however, require researchers to submit some information about the project, to verify that it qualifies for exemption.

- An IRB may allow certain minimal risk research to undergo **expedited review** by a single reviewer rather than the full committee (Table 14.2). The Office for Human Research Protections website lists the types of research eligible for expedited review (6). The concept of **minimal risk to participants** plays a key role in federal regulations, as indicated in Table 14.2. Minimal risk is defined as that "ordinarily encountered in daily life or during the performance of routine physical or psychological tests." Both the magnitude and probability of risk must be considered. The IRB must judge whether a specific project may be considered minimal risk.

Informed and Voluntary Consent

Investigators must obtain informed and voluntary consent from research participants.

Disclosure of Information to Participants

The federal regulations require investigators to discuss several topics with potential participants, including:

- **The nature of the research project.** The prospective participant should be told explicitly that research is being conducted, what the purpose of the research is, and who is being recruited as a participant. The specific study hypothesis need not be stated.
- **The procedures of the study.** Participants need to know what they will be asked to do in the research project. On a practical level, they should be told how much time will be required and how often. Procedures that are not standard clinical care should be identified as such. If the study involves blinding or randomization, these concepts should be explained in terms the participant can understand. In interview or questionnaire research, participants should be informed of the topics to be addressed.
- **The risks and potential benefits of the study and the alternatives to participating in the study.** Medical, psychosocial, and economic risks and benefits should be described in lay terms. Also, potential participants need to be told the alternatives to participation; for example, whether the intervention in a clinical trial is available outside the study. Concerns have been voiced that often the information provided to participants understates the risks and overstates the benefits (7). For example, research on new drugs is sometimes described as offering benefits to participants. However, most promising new interventions, despite encouraging preliminary results, show no significant advantages over standard therapy. Participants commonly have a "therapeutic misconception" that the research intervention is designed to provide them a personal benefit (8). Investigators should make clear that it is

not known whether the study drug or intervention is more effective than standard therapy and that promising drugs can cause serious harms.

Consent Forms

Written consent forms are generally required to document that the process of informed consent—discussions between an investigator and the participant—has occurred. The consent form needs to contain the required information discussed in the previous section. Alternatively, a short form may be used, which states that the required elements of informed consent have been presented orally. If the short form is used, there must be a witness to the oral presentation, who must sign the short consent form in addition to the participant.

IRBs usually have template consent forms that they prefer investigators to use. IRBs may require more information to be disclosed than federal regulations require.

Participants' Understanding of Disclosed Information

Research participants commonly have serious misunderstandings about the goals of research and the procedures and risks of the specific protocol (9). In discussions and consent forms, researchers should avoid technical jargon and complicated sentences. IRBs have been criticized for excessive focus on consent forms rather than on whether participants have understood crucial information (9). Strategies to increase **comprehension** by participants include having a study team member or a neutral educator spend more time talking one-on-one with study participants, simplifying consent forms, using a question-and-answer format, providing information over several visits, and using audiotapes or videotapes (10). In research that involves substantial risk or is controversial, investigators should consider assessing participants' comprehension and documenting that the participant can correctly answer questions about key aspects of the research (11, 12).

The Voluntary Nature of Consent

Ethically valid consent must be voluntary as well as informed. Researchers must minimize the possibility of coercion or undue influence. Examples of **undue influence** are excessive payments to participants and enrolling students as research participants. Undue influence is ethically problematic if it leads participants to significantly discount the risks of a research project or seriously undermines their ability to decline to participate. Participants must understand that declining to participate in the study will not compromise their medical care and that they may withdraw from the project at any time.

Exceptions to Informed Consent

Some scientifically important studies would be difficult or impossible to carry out if informed consent were required from each participant.

Research with Leftover De-Identified Specimens and Data

CASE 14.2 Research with Neonatal Blood Specimens

Shortly after birth, infants have a heel stick to collect blood onto filter paper to screen for genetic diseases. In most states, parental permission is not required for this mandated screening; hence the specimens represent the entire population of newborns. Specimens left over after clinical screening have been valuable for research on genetic causes of birth defects and preterm birth, environmental exposures during pregnancy, and gene–environment interactions.

Informed consent and IRB review are not required to use de-identified specimens in research (Table 14.1), but many IRBs still require investigators to notify them of such research. When original research is submitted for publication, many journals require authors to declare that an IRB approved the protocol or determined that review was not needed.

Waiver of Informed Consent

Some valuable research projects require identified existing information and specimens. Such studies do not qualify for exemption from IRB review, but may qualify for a waiver of informed consent.

CASE 14.2 Research with Neonatal Blood Specimens (*Continued*)

A research team would like to use identified neonatal blood specimens to study the association between maternal environmental exposures to selected chemicals and low birth weight, prematurity, and perinatal deaths. Researchers can link identified specimens to birth certificates, death certificates, and hospital records. Because of the large number of children who need to be studied to achieve adequate power to detect associations, it would not be feasible to obtain permission from parents or guardians.

Under federal regulations, IRBs may grant waivers of informed consent if all of the conditions in Table 14.3 apply. Most IRBs would waive consent for the proposed study of maternal environmental exposures.

Rationale for Exceptions to Informed Consent

Some scientifically important research presents such low risks that consent would be burdensome, while doing little to protect research participants. Every patient has benefited from knowledge obtained from research that used existing records and specimens. Fairness in the sense of reciprocity suggests that people who receive such benefits should be willing to participate in similar very low risk research to benefit others.

Objections to Exceptions to Informed Consent

Even though the federal regulations permit de-identified neonatal blood specimens to be used for research without parental permission, there is significant public opposition.

CASE 14.2 Research with Neonatal Blood Specimens (*Continued*)

Parents in several states have objected to the storage of specimens for unspecified research without their permission or the opportunity to withdraw from research, bringing lawsuits in two states. The plaintiffs did not contest the collection of blood for neonatal screening but objected that even de-identification of the specimens failed to address their concerns about loss of privacy and autonomy.

Because such objections to research might undermine the clinical uptake of neonatal screening, states are increasingly giving parents an opportunity to opt out of research uses of neonatal specimens collected in state screening programs. Such attention to parental wishes may be beyond what the federal research regulations require. Thus, what is legally permitted in research might not always be ethically acceptable, particularly for sensitive research.

TABLE 14.3 RESEARCH THAT MAY RECEIVE A WAIVER OF INFORMED CONSENT

1. The research involves no more than minimal risk to the participants; *and*:
2. The waiver or alteration will not adversely affect the rights and welfare of the participants; *and*
3. The research could not practicably be carried out without the waiver; *and*
4. Whenever appropriate, the subjects will be provided with additional pertinent information after participation. This provision allows some research involving deception, for example, when disclosing the propose of the research would undermine study validity.

Participants Who Lack Decision-Making Capacity

When participants are not capable of giving informed consent, permission to participate in the study should be obtained from the participant's legally authorized representative (the parent or guardian in the case of young children). Also, the protocol should be subjected to additional scrutiny, to ensure that the research question could not be studied in a population that is capable of giving consent.

Minimizing Risks

Researchers need to anticipate risks that might occur in research projects and reduce them, for instance by identifying and excluding persons who are very susceptible to adverse events, appropriate monitoring for adverse events, and substituting less invasive measurements. An important aspect of minimizing risk is maintaining participants' confidentiality.

Confidentiality

Breaches of confidentiality might cause stigma or discrimination, particularly if the research addresses sensitive topics such as sexual attitudes or practices, use of alcohol or drugs, illegal conduct, and psychiatric illness. Strategies for protecting confidentiality include coding research data, protecting or destroying the key that identifies participants, and limiting personnel who have access to identifiers. However, investigators should not make unqualified promises of confidentiality. Confidentiality may be overridden if research records are audited or subpoenaed, or if conditions are identified that legally must be reported, such as child abuse, certain infectious diseases, and serious threats of violence. In projects where such reporting can be foreseen, the protocol should specify how field staff should respond, and participants should be informed of these plans.

Investigators can forestall subpoenas in legal disputes by obtaining **confidentiality certificates** from the Public Health Service (13), which allow them to withhold identifiable research data if faced with a subpoena or court order to disclose them. However, these certificates have not been widely tested in court rulings, do not apply to audits by funding agencies or the FDA, and do not preclude the researcher from voluntarily disclosing information regarding child or elder abuse, domestic violence, or reportable communicable diseases. The research need not be federally funded to receive a certificate of confidentiality.

The HIPAA Health Privacy Rule

The federal Health Privacy Rule (commonly known as **HIPAA**, after the Health Insurance Portability and Accountability Act) protects individually identifiable health information, which is termed **protected health information**. Under the Privacy Rule, individuals must sign an authorization for the use of protected health information in a research project (14). This HIPAA authorization form is in addition to the informed consent form required by the IRB. Researchers must obtain authorization for each use of protected information for research; general consent for future research is not permitted. Authorization is not required if data are not identifiable and in

certain other situations. Researchers should contact their IRB with questions about the Privacy Rule and how it differs from the Federal Regulations on the Protection of Human Subjects.

RESEARCH PARTICIPANTS WHO REQUIRE ADDITIONAL PROTECTIONS

Some participants might be "at greater risk for being used in ethically inappropriate ways in research" because of difficulty giving voluntary and informed consent or increased susceptibility to adverse events (15).

Types of Vulnerability

Identifying different types of vulnerability allows researchers to adopt safeguards tailored to the specific type of vulnerability.

Cognitive or Communicative Impairments

Persons with impairment for cognition or communication might have difficulty understanding information about a study and weighing the risks and benefits of the study.

Power Differences

Persons who reside in institutions, such as prisoners or nursing home residents, might feel pressure to participate in research and to defer to persons who control their daily routine. Residents might not appreciate that they may decline to participate in research without retaliation by authorities or jeopardy to other aspects of their everyday lives.

If the investigator in the research project is also a participant's **treating physician**, the participant might hesitate to decline to participate in research, fearing that the physician would then be less interested in his or her care. Similarly, students and trainees might feel pressure to enroll in research conducted by their instructors or superiors.

Social and Economic Disadvantages

Persons with low socioeconomic status or poor access to health care might join a research study to obtain payment or medical care, even if they would regard the risks as unacceptable if they had a higher income. Participants with poor education or low health literacy might fail to comprehend information about the study or be more susceptible to influence by other people.

Protections for Vulnerable Participants

Federal regulations on research with vulnerable participants can be found on the federal Office for Human Research Protections website (3).

Research on Children

Investigators must obtain the permission of the parents and the assent of the child if developmentally appropriate. Research with children involving more than minimal risk is permissible only:

- If it offers the prospect of direct benefit to the child OR
- If the increase over minimal risk is minor and the research is likely to yield "generalizable knowledge of vital importance about the child's disorder or condition."

Research on Prisoners

Prisoners might not feel free to refuse to participate in research and might be unduly influenced by cash payments, breaks from prison routine, or parole considerations. Federal regulations

limit the types of research that are permitted in prisoners and require stricter IRB review and approval by the Department of Health and Human Services.

Research on Pregnant Women, Fetuses, and Embryos

Research that offers no prospect of direct benefit to the fetus is permitted only if "the purpose of the research is the development of important biomedical knowledge that cannot be obtained by any other means". Research that offers the prospect of direct benefit only to the fetus requires the informed consent of the father as well as the pregnant woman, even though research that offers children the prospect of direct benefit requires the permission of only one parent. These restrictions have been criticized for deterring research that would strengthen the evidence base for clinical care of pregnant women and their fetuses.

■ RESPONSIBILITIES OF INVESTIGATORS

Allegations of serious research misbehavior continue to occur today.

CASE 14.3 Cardiac Adverse Effects of Rofecoxib

In 2000, the results of the VIGOR randomized controlled trial were published. This study compared a new COX-2 selective nonsteroidal anti-inflammatory drug, rofecoxib, to an older, nonselective drug, naproxen (16). The manufacturer of rofecoxib sponsored the study. Rofecoxib caused significantly fewer gastrointestinal complications than naproxen (2.1 versus 4.5 per 100 patient-years), while providing similar efficacy for arthritis pain. The rofecoxib arm also had more heart attacks (0.4% versus 0.1%). Following this publication, rofecoxib was widely prescribed, with sales over $2.5 billion annually. Before the article was published, three additional heart attacks in the rofecoxib arm were reported to the FDA, but not to the university-based authors of the paper or to the journal. Two authors who were employees of the manufacturer knew of these additional cases. The journal that published the VIGOR study results later issued an expression of concern that the "article did not accurately represent the safety data available when the article was being reviewed for publication" (17). In addition to withholding unfavorable data, the publication set an earlier cutoff date for cardiovascular adverse events than for gastrointestinal adverse events without disclosing this to the journal or academic authors of the study, biasing the results in favor of rofecoxib.

Subsequently, another randomized trial showed that rofecoxib caused significantly more heart attacks and strokes than naproxen (18), and the manufacturer voluntarily withdrew the drug from the market.

In other highly influential publications, researchers intentionally made up or altered data, for example, in alleging a link between measles-mumps-rubella vaccine and childhood autism and in claiming to derive a human stem cell line using somatic cell nuclear transplantation (19, 20). Such misconduct undermines public and physician trust in research and threatens public funding for research.

Scientific Misconduct

The federal Office for Research Integrity defines **research misconduct** as fabrication, falsification, and plagiarism (21).

- **Fabrication** is making up results and recording or reporting them.
- **Falsification** is manipulating research materials, equipment, or procedures or changing or omitting data or results, so that the research record misrepresents the actual findings.
- **Plagiarism** is appropriating another person's ideas, results, or words without giving appropriate credit.

In this federal definition, misconduct must be intentional in the sense that perpetrators are aware that their conduct is wrong. In Case 14.3, intentional falsification of findings could not be proved. Research misconduct excludes honest error and legitimate scientific differences of opinion, which are a normal part of the research process. The federal definition does not address other wrong actions, such as double publication, refusal to share research materials, and sexual harassment; research institutions should deal with them under other policies.

When research misconduct is alleged, both the federal funding agency and the investigator's institution have the responsibility to carry out a fair and timely inquiry or investigation (22). During an investigation, both whistleblowers and accused scientists have rights that must be respected. Whistleblowers need to be protected from retaliation, and accused scientists need to be told the charges and given an opportunity to respond. Punishment for proven research misconduct may include suspension of a grant, debarment from future grants, and other administrative, academic, criminal, or civil sanctions.

CASE 14.3 Cardiac Adverse Effects of Rofecoxib (*Continued*)

Many patients who had taken rofecoxib and suffered a heart attack sued the manufacturer. During the legal process, internal sponsor e-mails were subpoenaed, which indicated that many articles on rofecoxib were commonly drafted by company employees or consultants, and academic investigators were often invited to be first author only after the manuscript had been drafted. The employees who drafted the articles frequently were not listed as authors or acknowledged.

Authorship

To merit authorship, researchers must make substantial contributions to:

- Study conception and design, or data analysis and interpretation, *and*
- Drafting or revising the article; *and*
- Giving final approval of the manuscript. (23)

Guest authorship and **ghost authorship** are unethical. Guest or honorary authors are listed as authors despite having made only trivial contributions to the paper; for example, by providing name recognition, access to participants, reagents, laboratory assistance, or funding. In Case 14.3, it is not appropriate for people to become authors after the study is completed, the data analyzed, and the first draft written. Ghost authors make substantial contributions to a paper but are not listed as authors. They are generally employees of pharmaceutical companies or medical writing companies. Omission of ghost writers misleads readers into underestimating the company's role in the manuscript. According to one study, 25% of original research articles in high-impact general journals have guest authors and 12% have ghost authors (24).

Disagreements commonly arise regarding who should be an author or the order of authors. These issues are best discussed explicitly and decided at the beginning of a project. Changes in authorship should be negotiated if decisions are made to shift responsibilities for the work. Suggestions have been made for carrying out such negotiations diplomatically (25). Because there is no agreement on criteria for position of authors, some journals describe the contributions of each author to the project in the published article.

Conflicts of Interest

A researcher's primary interests should be providing valid answers to important scientific questions and protecting the safety of participants. Researchers might have other interests, such as their reputation or income, that conflict with the primary goals of research and might impair their objectivity or undermine public trust in research (26).

Types of Conflicts of Interests

- **Financial conflicts of interest.** Studies of new drugs, devices, and tests are commonly funded by industry. The ethical concern is that certain financial ties might lead to bias in the design and conduct of the study, the overinterpretation of positive results, or failure to publish negative results (27, 28). If investigators hold patents on the study intervention or stock options in the company making the drug or device under study, they might reap large financial rewards if the treatment is shown to be effective, in addition to their compensation for conducting the study. Finally, receipt of large consulting fees, honoraria, or in-kind gifts might bias an investigator's judgment in favor of the company's product.
- **Dual roles for clinician-investigators.** If an investigator is the personal physician of an eligible research participant, the role of clinician and investigator might conflict. Patients might fear that their future care will suffer if they decline to participate in the research, and they might not distinguish between research and treatment. Furthermore, what is best for a particular patient might differ from what is best for the research project.

Responding to Conflicting Interests

All conflicts of interest should be disclosed, and some have such great potential for biasing research results that they should be managed or avoided.

- **Reduce the likelihood of bias.** In well-designed clinical trials, several standard precautions help keep competing interests in check. Investigators can be **blinded** to the intervention a participant is receiving to prevent bias in assessing outcomes. An independent **data and safety monitoring board** (see Chapter 11), whose members have no conflict of interest, can review interim data and terminate the study if the data provide convincing evidence of benefit or harm. The **peer review** process for grants, abstracts, and manuscripts also helps reduce bias.
- **Separate conflicting roles.** Physicians should separate the role of investigator in a research project from the role of clinician providing the participant's medical care. In general, physicians should not enroll their own patients in a research study where they are a co-investigator. If such patients are enrolled, a member of the research team who is not the treating physician should handle consent discussions.
- **Control of analysis and publications.** In research funded by a pharmaceutical company, academic-based investigators need to ensure that the contract gives them **control over the primary data and statistical analysis**, and the **freedom to publish findings**, whether or not the investigational drug is found to be effective (27, 28). The investigator has an ethical obligation to take responsibility for all aspects of the research. The sponsor may review the manuscripts, make suggestions, and ensure that patent applications have been filed before the article is submitted to a journal. However, the sponsor should not have power to veto or censor publication or to insist on specific language in the manuscript.
- **Disclose conflicting interests.** Research institutions require conflicts of interest to be disclosed to a designated office. The NIH and other funding agencies, local IRBs, scientific meetings, and medical journals require disclosure of conflicts of interest when grants, abstracts, or papers are submitted. Although disclosure alone is often an inadequate response to serious conflicts of interest, it might deter investigators from ethically problematic practices and allows reviewers and readers of journal articles to assess the potential for undue influence.
- **Manage conflicts of interest.** If a particular study presents significant conflicts of interest, the research institution, funding agency, or IRB may require additional safeguards, such as

closer monitoring of the informed consent process or modification of the conflicted investigator's role.

- **Prohibit certain situations.** To minimize conflicts of interest, funders or academic institutions might prohibit the patent holder on an intervention or an officer of the company manufacturing the intervention to serve as principal investigator in a clinical trial.

■ ETHICAL ISSUES SPECIFIC TO CERTAIN TYPES OF RESEARCH

Randomized Clinical Trials

Although randomized clinical trials are the most rigorous design for evaluating interventions (see Chapter 10), they present special ethical concerns for two reasons: The intervention is determined by chance and, in contrast to observational studies, researchers carry out an intervention on participants. One ethical justification for assigning treatment by randomization is that the study interventions are in **equipoise**, a concept that seems intuitively clear but is hotly debated and impossible to define precisely (29). There should be genuine uncertainty or controversy over which arm of the trial is superior, so that participants will not be significantly harmed if they allow their care to be determined by randomization rather than by their personal physician. Equipoise does not require an exact balance among study arms.

Participants in a clinical trial receive an intervention whose adverse effects might be unknown. Thus, trials require careful monitoring to make sure that participants are not being inappropriately harmed. It is the investigator's responsibility to establish careful methods for evaluating adverse effects (see Chapters 10 and 11). For most trials, this includes establishing an independent Data and Safety Monitoring Board that intermittently reviews study data and has the power to stop the trial if there is unexpected harm associated with the intervention (see Chapter 11).

Interventions for control groups also raise ethical concerns. If there is a standard of effective care for a condition, the control group should receive it (see Chapter 11). However, placebo controls may still be justified in short-term trials that do not offer serious risks to participants, such as studies of mild hypertension and mild, self-limited pain. Participants need to be informed of effective interventions that are available outside the research study.

It is unethical to continue a clinical trial if there is compelling evidence that one arm is safer or more effective. Furthermore, it would be wrong to continue a trial that will not answer the research question because of low enrollment, few outcome events, or high dropout rates. The periodic analysis of interim data in a clinical trial by an independent Data and Safety Monitoring Board can determine whether a trial should be terminated prematurely for these reasons (30). Such interim analyses should not be carried out by the researchers themselves, because unblinding investigators to interim findings can lead to bias if the study continues, and the investigators often have conflicting interest in continuing or stopping a study. Procedures for examining interim data and statistical stopping rules should be specified before enrolling participants (see Chapter 11).

Clinical trials in developing countries present additional ethical dilemmas (Chapter 18).

Research on Previously Collected Specimens and Data

Research with previously collected data and stored specimens offers the potential for significant discoveries. For example, DNA testing on a large number of stored biological specimens that are linked to clinical data might identify genes that increase the likelihood of developing a disease, having a poor prognosis, or responding to a particular treatment. Large biobanks of blood and tissue samples allow future studies to be carried out without the collection of additional samples. Research on previously collected specimens and data offers no physical risks to participants. However, there might be ethical concerns. Consent for unspecified future studies is problematic because no one can anticipate what kind of research might be carried out later.

Furthermore, participants might object to future use of data and samples in certain ways. If breaches of confidentiality occur, they might lead to stigma and discrimination. Groups participating in research might be harmed even if individual participants are not.

When biological specimens are collected, consent forms should allow participants to agree to or refuse certain broad categories of future research using the specimens. For example, participants might allow their specimens to be used:

- For future research that is approved by an IRB and scientific review panel; or
- Only for research on specific conditions; or
- Only in the current research study, not in future studies.

Participants should also know whether identifiable data and specimens will be shared with other researchers. Furthermore, participants should understand that research discoveries from the specimens might be patented and developed into commercial products.

OTHER ISSUES

Payment to Research Participants

Participants in clinical research deserve payment for their time and effort and reimbursement for out-of-pocket expenses such as transportation and child care. Practically speaking, compensation might also be needed to enroll and retain participants. A common practice is to offer higher payment for studies that are very inconvenient or risky. However, incentives also raise ethical concerns about undue inducement. If participants are paid more to participate in riskier research, persons of lower socioeconomic status might undertake risks against their better judgment. To avoid undue influence, it has been suggested that participants be compensated only for actual expenses and time, at an hourly rate for unskilled labor (31).

SUMMARY

1. Investigators must assure that their projects observe the **ethical principles** of **respect for persons**, **beneficence**, and **justice**.
2. Investigators must assure that research meets the requirements of applicable **federal regulations**, the key features being **informed consent** from participants and **IRB review**. During the informed consent process, investigators must explain to potential participants the **nature of the project** and the **risks**, **potential benefits**, and **alternatives**. Investigators must assure the **confidentiality** of participant information, observing the **HIPAA** Health Privacy Rule.
3. **Vulnerable populations**, such as **children**, **prisoners**, **pregnant women**, and people with **cognitive deficits** or **social disadvantage**, require additional protections.
4. Investigators must have **ethical integrity**. They must not commit **scientific misconduct**, which regulations define as **fabrication**, **falsification**, or **plagiarism**. Investigators need to disclose and appropriately manage **conflicts of interest** and should follow criteria for **appropriate authorship** by listing themselves as an author on a manuscript only if they made substantial intellectual contributions, and by ensuring that all persons who substantially contribute to a manuscript are listed as an author.
5. In certain types of research, additional ethical issues must be addressed. In randomized clinical trials, the intervention arms must be in **equipoise**, control groups must receive **appropriate interventions**, and the trial must not be continued once it has been demonstrated that one arm is more effective or harmful. When research is carried out on previously collected specimens and data, special attention needs to be given to **confidentiality**.

REFERENCES

1. Jones JH. The Tuskegee syphilis experiment. In: Emanuel EJ, Grady C, Crouch RA, et al., editors. *Oxford textbook of research ethics*. New York: Oxford University Press, 2008, 86–96.
2. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of biomedical and behavioral research*. 1979. Available at: www.hhs.gov/ohrp/humansubjects/guidance/belmont.html, accessed 8/27/12.
3. Department of Health and Human Services. *Protection of human subjects 45 CFR part 46*. 2005. Available at: www.dhhs.gov/ohrp/humansubjects/guidance/45cfr46.html, accessed 9/27/12.
4. Emanuel EJ, Menikoff J. Reforming the regulations governing research with human subjects. *N Engl J Med* 2011; 365:1145–50.
5. Lo B, Barnes M. Protecting research participants while reducing regulatory burdens. *JAMA* 2011;306:2260–2261.
6. Department of Health and Human Services. *Protocol review*. 2005. Available at: www.dhhs.gov/ohrp/policy/protocol/index.html, accessed 9/27/12.
7. King NMP, Churchill LR. Assessing and comparing potential benefits and risks of harm. In: Emanuel EJ, Grady C, Crouch RA, et al., editors. *The Oxford textbook of clinical research ethics*. New York: Oxford University Press, 2008, 514–526.
8. Henderson GE, Churchill LR, Davis AM, et al. Clinical trials and medical care: defining the therapeutic misconception. *PLoS Med* 2007;4:e324.
9. Federman DD, Hanna KE, Rodriguez LL. *Responsible research: a systems approach to protecting research participants*. 2002. Available at: www.nap.edu/catalog.php?record_id=10508, accessed 9/29/12.
10. Flory J, Emanuel E. Interventions to improve research participants' understanding in informed consent for research: a systematic review. *JAMA* 2004;292:1593–1601.
11. Lomax GP, Hall ZW, Lo B. Responsible oversight of human stem cell research: the California Institute for Regenerative Medicine's medical and ethical standards. *PLoS Med* 2007;4:e114.
12. Woodsong C, Karim QA. A model designed to enhance informed consent: experiences from the HIV prevention trials network. *Am J Public Health* 2005;95:412–419.
13. Wolf LE, Dame LA, Patel MJ, et al. Certificates of confidentiality: legal counsels' experiences with perspectives on legal demands for reseasch data. *J Empir Res Hum Res Ethics* 2012;7:1–9.
14. Nass SJ, Leavitt LA, Gostin LO. *Beyond the HIPAA Privacy Rule: enhancing privacy, improving health through research*. 2009. Available at: http://iom.edu/Reports/2009/Beyond-the-HIPAA-Privacy-Rule-Enhancing-Privacy-Improving-Health-Through-Research.aspx, accessed 9/29/12.
15. National Bioethics Advisory Commission. *Ethical and policy issues in research involving human participants*. Rockville, MD: National Bioethics Advisory Commission, 2001.
16. Bombardier C, Laine L, Reicin A, et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *N Engl J Med* 2000;343:1520–1528.
17. Curfman GD, Morrissey S, Drazen JM. Expression of concern. *N Engl J Med* 2005;353:2813–2814.
18. Bresalier RS, Sandler RS, Quan H, et al. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. *N Engl J Med* 2005;352:1092–1102.
19. Godlee F, Smith J, Marcovitch H. Wakefield's article linking MMR vaccine and autism was fraudulent. *BMJ* 2011;342:c7452.
20. Kennedy D. Responding to fraud. *Science* 2006;314:1353.
21. Office of Research Integrity. *Case summaries*. Available at: http://ori.hhs.gov/case_summary, accessed 9/29/12.
22. Mello MM, Brennan TA. Due process in investigations of research misconduct. *N Engl J Med* 2003;349:1280–1286.
23. International Committee of Medical Journal Editors. *Uniform requirements for manuscripts submitted to biomedical journals*. Available at: www.icmje.org/faq_urm.html, accessed 9/29/12.
24. Wislar JS, Flanagin A, Fontanarosa PB, Deangelis CD. Honorary and ghost authorship in high impact biomedical journals: a cross sectional survey. *BMJ* 2011;343:d6128.
25. Browner WS. Authorship. In: *Publishing and presenting clinical research*, 2nd ed. Philadelphia: Lippincott Williams & Willkins, 2006, 137–144.
26. Lo B, Field M. *Conflict of interest in medical research, education, and practice*. 2009. Available at: www.iom.edu/Reports/2009/Conflict-of-Interest-in-Medical-Research-Education-and-Practice.aspx, accessed 11/16/11.
27. DeAngelis CD, Fontanarosa PB. Ensuring integrity in industry-sponsored research: primum non nocere, revisited. *JAMA* 2010;303:1196–1198.
28. DeAngelis CD, Fontanarosa PB. Impugning the integrity of medical science: the adverse effects of industry influence. *JAMA* 2008;299:1833–1835.
29. Joffe S, Miller FG. Equipoise: asking the right questions for clinical trial design. *Nat Rev Clin Oncol* 2012;9:230–235.
30. Ellenberg SS, Fleming TR, DeMets DL. *Data monitoring committees in clinical trials*. Chichester, England: Wiley, 2003.
31. Grady C. Payment of clinical research subjects. *J Clin Invest* 2005;115:1681–1687.

CHAPTER 15

Designing Questionnaires, Interviews, and Online Surveys

Steven R. Cummings, Michael A. Kohn, and Stephen B. Hulley

Much of the information used in clinical research is gathered using **questionnaires**, administered on paper or electronically, or through **interviews**. For many studies, the validity of the results depends on the quality of these **instruments**. In this chapter we describe the components of questionnaires and interviews and outline procedures for developing them.

Clinical researchers have a rapidly increasing number of options for developing **online surveys**, including REDCap, a Web-based data management platform developed by a Vanderbilt University consortium, and commercial products such as SurveyMonkey, Zoomerang, Qualtrics, and QuesGen. These products provide online, easy-to-use survey development tools and utilities for automatic e-mailing to study participants or posting on the study website. The ongoing transition from paper-based to Web-based surveys has not changed the principles of designing good instruments: writing clear instructions and well-phrased questions that elicit informative responses (1).

DESIGNING GOOD INSTRUMENTS

Open-Ended and Closed-Ended Questions

There are two basic types of questions, open-ended and closed-ended, which serve somewhat different purposes. **Open-ended questions** are particularly useful when it is important to hear what respondents have to say in their own words. For example:

> What habits do you believe increase a person's chance of having a stroke?

Open-ended questions leave the respondent **free** to answer with fewer limits imposed by the researcher. They allow participants to report more information than is possible with a discrete list of answers, but the responses may be less complete. A major **disadvantage** is that open-ended questions usually require qualitative methods or special systems (such as coding dictionaries for symptoms and adverse events) to code and analyze the responses; this takes more time than entering responses to closed-ended questions, and may require subjective judgments. Open-ended questions are often used in exploratory phases of question design because they help the researcher understand a concept as respondents express it. Phrases and words used by respondents can form the basis for **closed-ended questions** that ask respondents to choose from two or more preselected answers:

Which of the following do you believe increase the chance of having a stroke? (Check all that apply.)

- ☐ Smoking
- ☐ Being overweight
- ☐ Stress
- ☐ Drinking alcohol

Because closed-ended questions provide a list of possible alternatives from which the respondent may choose, they are quicker and **easier to answer** and the answers are **easier to tabulate** and analyze. In addition, the list of possible answers often helps clarify the meaning of the question, and closed-ended questions are well suited for use in multi-item scales that produce a single score.

On the other hand, closed-ended questions have several **disadvantages**. They lead respondents in certain directions and do not allow them to express their own, potentially more accurate, answers. The set of answers may not be **exhaustive** (not include all possible options, e.g., "sexual activity" or "dietary salt"). One solution is to include an option such as "Other (please specify)" or "None of the above." When a single response is desired, the respondent should be so instructed and the set of possible responses should also be **mutually exclusive** (i.e., the categories should not overlap) to ensure clarity and parsimony.[1]

When the question allows more than one answer, instructing the respondent to mark "All that apply" is not ideal. This does not force the respondent to consider each possible response, and a missing item may represent either an answer that does not apply or an overlooked item. It is better to ask respondents to mark each possible response as either "yes" or "no":

Which of the following do you believe increases the chance of having a stroke?

	Yes	No	Don't know
Smoking	◯	◯	◯
Being overweight	◯	◯	◯
Stress	◯	◯	◯
Drinking alcohol	◯	◯	◯

The **visual analog scale (VAS)** is another option for recording answers to closed-ended questions using lines or other drawings. The participant is asked to mark a line at a spot along the continuum from one extreme to the other that best represents his answer. It is important that the words that anchor each end describe the most extreme values for the item of interest. Here is a VAS for pain severity:

Please mark the place on this line that best describes the severity of your pain in general over the past week.

None ——————————————— Unbearable

For convenience of measurement, the lines are often 10 cm long and the score is the distance, in centimeters, from the lowest extreme. For an example of an online VAS, see the website: http://www.epibiostat.ucsf.edu/dcr/.

[1] For online forms, the convention is to display mutually exclusive options as radio buttons (circles), and to use check boxes (squares) for responses to "all that apply" questions.

VASs are attractive because they rate characteristics on a continuous scale; they may be more sensitive to small changes than ratings based on categorical lists of adjectives. Many of the online survey tools, including REDCap, Qualtrics, and QuesGen, accommodate VASs.

Formatting

On questionnaires, it is customary to describe the purpose of the study and how the data will be used in a brief statement at the outset. Similar information is usually presented at the beginning of an interview as part of obtaining consent. To ensure accurate and standardized responses, all instruments must have instructions specifying how they should be filled out. This is true not only in self-administered questionnaires, but also for the forms that interviewers use to record responses.

Sometimes it is helpful to provide an example of how to complete a question, using a simple question that is easily answered:

> Instructions on How to Fill Out a Questionnaire That Assesses Dietary Intake
>
> These questions are about your usual eating habits during the past 12 months. Please mark your usual serving size and write down how often you eat each food in the boxes next to the type of food.
>
> For example, if you drink a medium (6 oz) glass of apple juice about three times a week, you would answer:
>
> Apple Juice ◯ Small (3 oz) ◉ Medium (6 oz) ◯ Large (9 oz) [3] time(s) per ◯ Day ◉ Week ◯ Month ◯ Year

To improve the flow of the instrument, questions concerning major subject areas should be grouped together and introduced by headings or short descriptive statements. To warm up the respondent to the process of answering questions, it is helpful to begin with emotionally neutral questions such as name and contact information. Highly sensitive questions about income or sexual function are often placed at the end of the instrument. For each question or set of questions with a format that differs from that of other questions on the instrument, instructions must clearly indicate how to respond.

If the instructions include different time frames, it is sometimes useful to repeat the time frame at the top of each new set of questions. For example, questions such as:

> How often have you visited a doctor during the past year?
>
> During the past year, how many times have you been a patient in an emergency department?
>
> How many times were you admitted to the hospital during the past year?

can be shortened and tidied as follows:

> During the past year, how many times have you
>
> - Visited a doctor?
> - Been a patient in an emergency department?
> - Been admitted to a hospital?

For **paper forms**, the **visual design** should make it as easy as possible for respondents—whether study subjects or research staff—to complete all questions in the correct sequence. If the format is too complex, respondents or interviewers may skip questions, provide the wrong information, or sometimes even refuse to complete the instruments. A **neat** format with **plenty of space** is more attractive and easier to use than one that is crowded or cluttered. Although investigators often assume that a questionnaire will appear shorter by having fewer pages, the task can be more difficult when more questions are crowded onto a page. Response scales should be spaced widely enough so that it is easy to circle or check the correct number without the mark accidentally including the answer "above" or "below." When an open-ended question is included, the space for responding should be big enough to allow respondents with large handwriting to write comfortably in the space. People with visual problems, including many elderly subjects, will appreciate large type and high contrast (black on white).

Possible answers to closed-ended questions should be lined up vertically and preceded by boxes or brackets to check, or by numbers to circle, rather than using open blanks:

How many different medicines do you take every day? (Check one)
◯ None
◯ 1–2
◯ 3–4
◯ 5–6
◯ 7 or more

Note that these response options are exhaustive and mutually exclusive.

Sometimes the investigator may wish to follow-up certain answers with more detailed questions. This is best accomplished by a **branching question**. Respondents' answers to the initial question (often referred to as a **screener**) determine whether they are directed to answer additional questions or skip ahead to later questions. For example:

Have you ever been told that you have high blood pressure?
◯ Yes ⟶ How old were you when you were first told you had high blood pressure?
_ _ years old
◯ No
↓
Go to question 11

Branching questions save time and allow respondents to avoid irrelevant or redundant questions. Directing the respondent to the next appropriate question is done by using arrows to point from response to follow-up questions and including directions such as "*Go to question 11*" (see Appendix 15).

Online surveys are generally clearer and easier for the respondents because they incorporate skip logic. A male study subject will not see a question about pregnancies and will only reach the question about pack years if he answered "yes" to the question about cigarette smoking. (See www.epibiostat.ucsf.edu/dcr/.) However, the skip logic must be carefully validated during the study's pretesting phase. Complex skip logic can result in dead ends and "orphan"

questions that are never reached. Good design, including consideration of respondents with visual problems,[2] is just as important for online forms as for paper forms

Wording

Every word in a question can influence the validity and reproducibility of the responses. The objective is to construct questions that are simple, free of ambiguity, and encourage accurate and honest responses without embarrassing or offending the respondent.

- **Clarity.** Make questions as clear and specific as possible. Concrete words are preferred over abstract words. For example, asking, "How much exercise do you usually get?" is less clear than asking, "During a typical week, how many hours do you spend in vigorous walking?"
- **Simplicity.** Use simple, common words and grammar that convey the idea, and avoid technical terms and jargon. For example, it is clearer to ask about "drugs you can buy without a prescription from a doctor" than to ask about "over-the-counter medications."
- **Neutrality.** Avoid "loaded" words and stereotypes that suggest a desirable answer. Asking "During the last month, how often did you drink too much alcohol?" may discourage respondents from admitting that they drink a lot of alcohol. "During the last month, how often did you drink more than five drinks in one day?" is a more factual, less judgmental, and less ambiguous question.

Sometimes it is useful to set a tone that permits the respondent to admit to behaviors and attitudes that may be considered undesirable. For example, when asking about a patient's compliance with prescribed medications, an interviewer or a questionnaire may use an introduction: "People sometimes forget to take medications their doctor prescribes. Does that ever happen to you?" Such wording can be tricky, however—it is important to give respondents permission to admit certain behaviors without encouraging them to exaggerate.

Collecting information about potentially **sensitive** areas like sexual behavior or income is especially difficult. Some people feel more comfortable answering these types of questions in self-administered questionnaires than in interviews, but a skillful interviewer can sometimes elicit open and honest answers. It may be useful to put potentially embarrassing responses on a card so that the respondent can answer by simply pointing to a response.

Setting the Time Frame

To measure the frequency of the behavior it is essential to have the respondent describe it in terms of some **unit of time**. If the behavior is usually the same day after day, such as taking one tablet of a diuretic every morning, the question can be very simple: "How many tablets do you take a day?"

Many behaviors change from day to day, season to season, or year to year. To measure these, the investigator must first decide what aspect of the behavior is most important to the study: **the average** or **the extremes**. A study of the effect of alcohol on the risk of cardiovascular disease may need a measurement of average consumption over time, but a study of the role of alcohol in the occurrence of falls may need to know how frequently the respondent drank enough alcohol to become intoxicated.

Questions about average behavior can be asked in two ways: asking about "usual" or "typical" behavior or counting actual behaviors during a period of time. For example, an investigator may determine average intake of beer by asking respondents to estimate their usual intake:

[2] The commercial providers of online survey tools pay considerable attention to issues of readability, partially because Section 508 of the Rehabilitation Act of 1973 requires federal agencies to make their electronic forms accessible to people with disabilities. Most commercial providers are certified as "508 compliant."

> About how many beers do you have during a typical week (one beer is equal to one 12-oz can or bottle, or one large glass)?
>
> [___] beers per week

This format is simple and brief. It assumes, however, that respondents can accurately average their behavior into a single estimate. Because drinking patterns often change markedly over even brief intervals, the respondent may have a difficult time deciding what is a typical week. Faced with questions that ask about usual or typical behavior, people often report the things they do most commonly and ignore the extremes. Asking about drinking on typical days, for example, will underestimate alcohol consumption if the respondent drinks large amounts on weekends.

An alternative approach is to quantify exposure during a **certain period of time**:

> During the last 7 days, how many beers did you have (one beer is equal to one 12-oz can or bottle, or one large glass)?
>
> [___] beers in the last 7 days

The goal is to ask about the shortest recent segment of time that accurately represents the characteristic over the whole period of interest for the research question. The best length of time depends on the characteristic. For example, patterns of sleep can vary considerably from day to day, but questions about sleep habits during the past week may adequately represent patterns of sleep during an entire year. On the other hand, the frequency of unprotected sex may vary greatly from week to week, so questions about unprotected sex should cover longer intervals.

Using **diaries** may be a more accurate approach to keep track of events, behaviors, or symptoms that happen episodically (such as falls) or that vary from day to day (such as vaginal bleeding). This may be valuable when the timing or duration of an event is important or the occurrence is easily forgotten. Participants can enter these data into **electronic devices**, and the approach allows the investigator to calculate an average daily score of the event or behavior being assessed. However, this approach can be time-consuming for participants and can lead to more missing data than the more common approach of asking retrospective questions. The use of diaries assumes that the time period assessed was typical, and that the self-awareness involved in using diaries has not altered the behavior being recorded in important ways.

Avoid Pitfalls

- **Double-barreled questions.** Each question should contain only one concept. Consider this question designed to assess caffeine intake: "How many cups of coffee or tea do you drink during a day?" Coffee contains much more caffeine than tea and differs in other ways, so a response that combines the two beverages is not as precise as it could be. When a question attempts to assess two things at one time, it is better to break it into two separate questions. "(1) How many cups of coffee do you drink during a typical day?" and "(2) How many cups of tea do you drink during a typical day?"
- **Hidden assumptions.** Sometimes questions make assumptions that may not apply to all people who participate in the study. For example, a standard depression item asks how often, in the past week: "I felt that I could not shake off the blues even with help from my family." This assumes that respondents have families and ask for emotional support; for those who

do not have a family or who do not seek help from their family, it is difficult to answer the question.

- **The question and answer options don't match.** It is important that the question match the options for the answer, a task that seems simple but is often done incorrectly. For example, the question "Have you had pain in the last week?" should not be matched with response options of "never," "seldom," "often," "very often." (The question should be changed to "How often have you had pain in the last week?" or the answer should be changed to "yes" or "no.") Another common problem occurs when questions about intensity are given agree/disagree options. For example, a respondent may be given the statement "I am sometimes depressed" and then asked to respond with "agree" or "disagree." Disagreeing with this statement could mean that the person is often depressed, or never depressed. It is usually clearer to use a simple question about how often the person feels depressed matched with options about frequency (never, sometimes, often).

Scales and Scores to Measure Abstract Variables

It is difficult to quantitatively assess an **abstract concept** such as quality of life from a single question. Therefore, abstract characteristics are commonly measured by generating scores from a series of questions organized into a scale (2, 3).

Using multiple items to assess a concept may have other advantages over single questions or several questions asked in different ways that cannot be combined. Compared with the alternative approaches, **multi-item scales** can increase the range of possible responses (e.g., a multi-item quality-of-life scale might generate scores that range from 1 to 100 whereas a single question rating quality of life might produce four or five responses from "poor" to "excellent"). A disadvantage of multi-item scales is that they produce results (quality of life = 46.2) that can be difficult to understand intuitively.

Likert scales are commonly used to quantify attitudes, behaviors, and domains of health-related quality of life. These scales provide respondents with a list of statements or questions and ask them to select a response that best represents the rank or degree of their answer. Each response is assigned a number of points. For example, consider a questionnaire to measure the strength of a person's opinion that a diet high in fruits and vegetables improves health:

For each item, circle the one number that best represents your opinion:

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
a. Eating more fruits and vegetables reduces the risk of heart disease.	1	2	3	4	5
b. Vegetarians are healthier than people who include meat in their diet.	1	2	3	4	5
c. Increasing the intake of fruits and vegetables slows the rate of aging.	1	2	3	4	5

An investigator can compute an overall **score** for a respondent's answers by simply summing the score for each item, or averaging the points for all non-missing items. For example, a person who answered that he or she strongly agreed that eating more fruits and vegetables reduces the risk of heart disease (one point), and that vegetarians are healthier than people who include meat in their diet (one point), but disagreed that increasing the intake of fruits and vegetables slows aging (four points), would have a total score of 6. Simply adding up or averaging item scores assumes that all the items have the same weight and that each item is measuring the same general characteristic.

The **internal consistency** of a scale can be tested statistically using measures such as **Cronbach's alpha** (4) that assess the overall consistency of a scale. Cronbach's alpha is calculated from the correlations between scores on individual items. Values of this measure above 0.80 are considered excellent, and below 0.50 unacceptable. Low values for internal consistency indicate that some of the individual items may be measuring different characteristics.

Creating New Scales

When an investigator needs to measure a characteristic for which there is no standard questionnaire or interview approach, it may be necessary to develop a new instrument or scale. The task can range from the creation of a single new question about a minor variable in a small study to developing and testing a new multi-item scale for measuring the primary outcome for a multicenter investigation. At the simplest end of this spectrum, the investigator may use good judgment and basic principles of writing to develop an item that can be pretested to make sure it is clear and produces appropriate answers. At the other extreme, developing a new instrument to measure an important concept may need a systematic approach that can take years from initial draft to final product.

The latter process often begins by generating potential items for the instrument from interviews with individuals and **focus groups** (small groups of people who are relevant to the research question and who are invited to spend 1 or 2 hours discussing specific topics pertaining to the study with a group leader). An instrument is then drafted, followed by critical review by peers, mentors, and experts. The investigator then proceeds with the iterative sequence of pretesting, revising, shortening, and validating that is described in the next section (and illustrated by Example 15.1).

EXAMPLE 15.1 Development of a New Multi-Item Instrument

The National Eye Institute Visual Function Questionnaire exemplifies the painstaking development and testing of a multi-item instrument. Mangione and colleagues devoted several years to creating and testing the scale because it was intended to serve as a primary measurement of outcome of many studies of eye disease (5, 6). They began by interviewing patients with eye diseases about the ways that the conditions affected their lives. Then they organized focus groups of patients with the diseases and analyzed transcripts of these sessions to choose relevant questions and response options. They produced and pretested a long questionnaire that was administered to hundreds of participants in several studies. They used data from these studies to identify items that made the largest contribution to variation in scores from person to person and to shorten the questionnaire from 51 to 25 items.

Because the creation and validation of new multi-item instruments is time-consuming, it should generally only be undertaken for variables that are central to a study, and when existing measures are inadequate or inappropriate for the people who will be included in the study.

■ STEPS IN ASSEMBLING THE INSTRUMENTS FOR THE STUDY

Make a List of Variables

Before designing an interview or questionnaire instrument, write a detailed list of the information to be collected and concepts to be measured in the study. Consider listing the role of each

item (e.g., predictors, outcomes, and potential confounders) in answering the main research questions.

Prefer Existing Measures, if Suitable

Assemble a file of questions or instruments that are available for measuring each variable. When there are several alternative methods, create an electronic file for each variable to be measured and then find and file copies of candidate questions or instruments for each item. It is important to use the best possible instruments to measure the main predictors and outcomes of a study, so most of the effort of collecting alternative instruments should focus on these **major variables**.

Start by collecting instruments from other investigators who have conducted studies that included measurements of interest. Existing questionnaires and information on their validity, internal consistency, and reliability can be found in the methods section of published reports, and by searching the Web for key terms such as "health outcomes questionnaires."

Borrowing instruments from other studies has the advantage of saving development time and allowing results to be compared across studies. It is ideal to use existing instruments without modification. However, if some of the items are inappropriate (as may occur when a questionnaire developed for one cultural group is applied to a different setting), it may be necessary to delete, change, or add a few items.

If an established instrument is too long, it may be useful to contact those who developed the instrument to see if they have shorter versions. Deleting items from established scales risks changing the meaning of scores and endangering comparisons of the findings with results from studies that used the intact scale. Shortening a scale can also diminish its reproducibility or its sensitivity to detect changes. However, it is sometimes acceptable to delete sections or "subscales" that are not essential to the study while leaving other parts intact.

Compose a New Instrument, if Necessary

The first draft of the instrument should have a broad reach, including more questions about the topic than will eventually be included in the instrument. The investigator should read the first draft carefully, attempting to answer each question as if he were a respondent and trying to imagine ways to misinterpret questions. The goal is to identify words or phrases that might be confusing or misunderstood, to find abstract words or jargon that could be translated into simpler, concrete terms, to notice complex questions that can be split into two or more questions. Colleagues and experts in questionnaire design should be asked to review the instrument, considering the content of the items as well as clarity.

Revise and Shorten the Set of Instruments for the Study

Studies usually collect more data than will be analyzed. Long interviews, questionnaires, and examinations may tire respondents and thereby decrease the accuracy and reproducibility of their responses. It is usually best to resist the temptation to include additional questions or measures "just in case" they might produce interesting data. Questions that are not essential to answering the main research question increase the amount of effort involved in obtaining, entering, cleaning, and analyzing data. Time devoted to unnecessary or marginally valuable data can detract from other efforts and decrease the overall quality and productivity of the study.

To decide if a concept is essential, the investigator can think ahead to analyzing and reporting the results of the study. Sketching out the final tables will help to ensure that all needed variables are included and to identify those that are less important. Once that is done, here's a maxim for deciding which items to include: **When in doubt, leave it out**.

Pretest

Pretest the instrument clarity and timing. For key measurements, large pilot studies may be valuable to find out whether each question produces an adequate range of responses and to test the validity and reproducibility of the instrument.

Validate

Questionnaires and interviews can be assessed for validity (an aspect of accuracy) and for reproducibility (precision) in the same fashion as any other type of measurement (Chapter 4). The process begins with choosing questions that have **face validity**, the subjective but important judgment that the items assess the characteristics of interest, and continues with efforts to establish **content validity** and **construct validity**. Whenever feasible, new instruments can then be compared with established **gold standard** approaches to measuring the condition of interest. Ultimately, the **predictive validity** of an instrument can be assessed by correlating measurements with future outcomes.

If an instrument is intended to measure change, then its responsiveness can be tested by applying it to patients before and after receiving treatments considered effective by other measures. For example, a new instrument designed to measure quality of life in people with impaired visual acuity might include questions that have face validity ("Are you able to read a newspaper without glasses or contact lenses?"). Answers could be compared with the responses to an existing validated instrument (Example 15.1) among patients with severe cataracts and among those with normal eye examinations. The responsiveness of the instrument to change could be tested by comparing responses of patients with cataracts before and after surgery. However, the process of validating new instruments is time-consuming and expensive, and worthwhile only if existing instruments are inadequate for the research question or population to be studied.

■ ADMINISTERING THE INSTRUMENTS

Questionnaires Versus Interviews

There are two basic approaches to collecting data about attitudes, behaviors, knowledge, health, and personal history. Questionnaires are instruments that respondents fill out by themselves, and interviews are those that are administered verbally by an interviewer. Each approach has advantages and disadvantages.

Questionnaires are generally a more efficient and uniform way to administer simple questions, such as age or habits of tobacco use. Questionnaires are less expensive than interviews because they require less research staff time, and they are more easily standardized. **Interviews** are usually better for collecting answers to complicated questions that require explanation or guidance, and interviewers can make sure that responses are complete. Interviews may be necessary when participants have variable ability to read and understand questions. However, interviews are more costly and time-consuming, and the responses may be influenced by the relationship between the interviewer and respondent.

Both types of instruments can be standardized, but interviews are inevitably administered at least a little differently each time. Both methods of collecting information are susceptible to errors caused by imperfect memory; both are also affected, though not necessarily to the same degree, by the respondent's tendency to give socially acceptable answers.

Interviewing

The skill of the interviewer can have a substantial impact on the quality of the responses. **Standardizing** the interview procedure from one interview to the next is the key to

maximizing reproducibility, with uniform wording of questions and uniform nonverbal signals during the interview. Interviewers must strive to avoid introducing their own biases into the responses by changing the words or the tone of their voice. For the interviewer to comfortably read the questions verbatim, the interview should be written in language that resembles common speech. Questions that sound unnatural or stilted when said aloud will encourage interviewers to improvise their own, more natural but less standardized way of asking the question.

Sometimes it is necessary to follow-up on a respondent's answers to encourage him to give an appropriate answer or to clarify the meaning of a response. This "**probing**" can also be standardized by writing standard phrases in the margins or beneath the text of each question. To a question about how many cups of coffee respondents drink on a typical day, some respondents might respond, "I'm not sure; it's different from day to day." The instrument could include the follow-up probe: "Do the best you can; tell me about how many you drink on a typical day."

Interviews can be conducted in person or over the telephone. **Computer-assisted telephone interviewing (CATI)** is a telephone surveying technique in which the interviewer follows a script and the computer facilitates the collection and editing of data. **Interactive voice response (IVR)** systems replace the interviewer with computer-generated questions that collect subject responses by telephone keypad or voice recognition (7). **In-peson** interviews, however, may be necessary if the study requires direct observation of participants or physical examinations, or if potential participants do not have telephones (e.g., the homeless).

Methods of Administering Questionnaires

Questionnaires can be given to subjects in person or administered through the mail, by e-mail, or through a website. Distributing questionnaires in person allows the researcher to explain the instructions before the participant starts answering the questions. When the research requires the participant to visit the research site for examinations, questionnaires can also be sent in advance of an appointment and answers checked for completeness before the participant leaves.

E-mailed questionnaires have several advantages over those sent by paper mail. Although they can only be sent to participants who have access to and familiarity with the Internet, questionnaires sent by e-mail are an easy way to provide data that can be directly entered into databases.

Questionnaires on websites or with **handheld electronic devices** have come into widespread use as efficient and inexpensive approaches to collecting health survey information (8). These approaches can produce very clean data because answers can be automatically checked for missing and out-of-range values, the errors pointed out to the respondent, and the responses accepted only after the errors are corrected.

■ CONSIDER DIRECT MEASUREMENTS

Advances in measurement instruments and biological assays are creating alternatives to questionnaires and interviews for measuring many common conditions and exposures. For example, direct measurement of physical activity by wearing small accelerometers yields more objective and precise estimation of total activity, patterns of actigraphy, and energy expenditure than do questionnaires about physical activity (9). Sensors worn at night can more accurately measure the amount and quality of sleep (10). Measurement of blood levels of nutrients such as vitamin D provides a more accurate measurement of the exposure to the nutrient than asking about consumption of foods containing vitamin D. Investigators should be alert for new technologies, often enabled by wireless electronic devices, that directly measure characteristics previously assessed only indirectly by questionnaires and interviews.

SUMMARY

1. For many clinical studies, the quality of the results depends on the quality and appropriateness of the **questionnaires** and **interviews**. Investigators should make sure the **instruments** are as **valid** and **reproducible** as possible before the study begins.
2. **Open-ended questions** allow subjects to answer without limitations imposed by the investigator, and **closed-ended questions** are easier to answer and analyze. The response options to a closed-ended question should be **exhaustive** and **mutually exclusive**.
3. Questions should be **clear**, **simple**, **neutral**, and **appropriate** for the population that will be studied. Investigators should examine potential questions from the viewpoint of potential participants, looking for **ambiguous terms** and common pitfalls such as **double-barreled questions**, **hidden assumptions**, and **answer options that do not match the question**.
4. Questionnaires should be **easy to read**, and interview questions should be comfortable to read out loud. The **format** should fit the method for electronic data entry and be spacious and uncluttered.
5. To measure **abstract variables** such as attitudes or health status, questions can be combined into **multi-item scales** to produce a total score. Such scores assume that the questions measure a single characteristic and that the responses are **internally consistent**.
6. An investigator should search out and use **existing instruments** that are known to produce valid and reliable results. When it is necessary to **modify existing measures** or **devise a new one**, the investigator should start by collecting existing measures to be used as potential models and sources of ideas.
7. The whole set of instruments to be used in a study should be **pretested** and timed before the study begins. For new instruments, small initial pretests can improve the clarity of questions and instructions; later, larger pilot studies can test and refine the new instrument's **range**, **reproducibility**, and **validity**.
8. **Self-administered questionnaires** are more economical than interviews, they are more readily standardized, and the added privacy can enhance the validity of the responses. **Interviews**, on the other hand, can ensure more complete responses and enhance validity through improved understanding.
9. Administration of instruments by **computer-assisted telephone interviewing**, **e-mail**, portable **electronic devices**, or on the study **website** can enhance the efficiency of a study.

APPENDIX 15
An Example of a Questionnaire About Smoking

The following items are taken from a self-administered paper questionnaire used in our Study of Osteoporotic Fractures. Note that the branching questions are followed by arrows that direct the subject to the next appropriate question and that the format is uncluttered with the responses consistently lined up on the left of each next area. For a link to an online version of this example, see www.epibiostat.ucsf.edu/dcr/.

1. Have you smoked at least 100 cigarettes in your entire life?

☐ → Yes

☐ No

2. About how old were you when you smoked your first cigarette?

☐☐ years old

3. On the average over the entire time since you started smoking, about how many cigarettes did you smoke per day?

☐☐ cigarettes per day

4. Have you smoked any cigarettes in the past week?

☐ → Yes

5. About how many cigarettes per day did you smoke in the past week?

☐☐ cigarettes per day

Please skip to next page, question #7

☐ No

6. How old were you when you stopped smoking?

☐☐ years old

Please go to question #7

7. Have you ever lived for at least a year in the same household with someone who smoked cigarettes regularly?

☐ Yes → 8. For about how many years, in total, have you lived with someone who smoked cigarettes regularly at the time?

☐☐ years

9. On the average over the entire time you lived with people who smoked, about how many cigarettes a day were smoked while you were at home?

☐☐ cigarettes per day

10. Do you now live in the same household with someone who smokes cigarettes regularly?

☐ Yes

☐ No

☐ No

11 etc.

REFERENCES

1. Iarossi G. *The power of survey design: a user guide for managing surveys, interpreting results, and influencing respondents*. Washington, DC: World Bank, 2006. Available at: https://openknowledge.worldbank.org/bitstream/handle/10986/6975/350340The0Powe1n0REV01OFFICIAL0USE1.pdf?sequence=1, accessed 03/11/13.
2. McDowell I. *Measuring health: a guide to rating scales and questionnaires*, 3rd ed. New York: Oxford University Press, 2006.
3. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*, 4th ed. New York: Oxford University Press, 2009.
4. Bland JM, Altman DG. Cronbach's alpha. *BMJ* 1997;314:572.
5. Mangione CM, Berry S, Spritzer K, et al. Identifying the content area for the 51-item National Eye Institute Visual Function Questionnaire: results from focus groups with visually impaired persons. *Arch Ophthalmol* 1998;116:227–233.
6. Mangione CM, Lee PP, Pitts J, et al. Psychometric properties of the National Eye Institute Visual Function Questionnaire (NEI-VFQ). NEI-VFQ Field Test Investigators. *Arch Ophthalmol* 1998;116:1496–1504.
7. Kobak KA, Greist JH, Jefferson JW, et al. Computer assessment of depression and anxiety over the phone using interactive voice response. *MD Comput* 1999;16:64–68.
8. Dillman DA, Smyth JD, Christian LM. *Internet, mail, and mixed-mode surveys: the tailored design method*, 3rd ed. Hoboken, NJ: Wiley, 2008.
9. Mackey DC, Manini TM, Schoeller DA, et al. Validation of an armband to measure daily energy expenditure in older adults. *J Gerontol A Biol Sci Med Sci* 2011;66:1108–1113.
10. Girshik J, Fritschi L, Heyworth J, et al. Validation of self-reported sleep against actigraphy. *J Epidemiol* 2012;22:462–468.

CHAPTER 16

Data Management

Michael A. Kohn, Thomas B. Newman, and Stephen B. Hulley

We have seen that undertaking a clinical research project requires choosing a study design, defining the population, and specifying the predictor and outcome variables. Ultimately, most information about the subjects and variables will reside in a computer **database** that will be used to store, update, and monitor the data, as well as format the data for statistical analysis. The study database may also store **administrative data**, such as call logs, visit schedules, and reimbursement records. Simple study databases consisting of individual data tables can be maintained using **spreadsheet** or statistical software. More complex databases containing multiple inter-related data tables require **database management software**.

Data management for a clinical research study involves defining the **data tables**, developing the **data entry** system, and **querying** the data for **monitoring** and **analysis**. In large clinical trials, especially trials preparatory to application for regulatory approval of a drug or device, the specialists who create data entry forms, manage and monitor the data collection process, and format and extract the data for analysis are referred to as **clinical data managers** (1). Large pharmaceutical companies running multiple clinical trials devote significant resources and personnel to clinical data management. Although the scale is generally much smaller, beginning investigators also need to attend carefully to data management issues.

DATA TABLES

All computer databases consist of one or more data tables in which the **rows** correspond to individual **records** (which may represent subjects, events, or transactions) and the **columns** correspond to **fields** (attributes of the records). For example, the simplest study databases consist of a single table in which each row corresponds to an individual study subject and each column corresponds to a subject-specific attribute such as name, date of birth, sex, and predictor or outcome status. In general, the first column is a unique **subject identification number** ("subjectID"). Using a unique subject identifier that has no meaning external to the study database simplifies the process of "de-linking" study data from personal identifiers for purposes of maintaining subject confidentiality. If the database contains additional tables with records corresponding to examinations, laboratory results, or telephone calls, then the first column in each of these tables should be a unique record identifier such as ExamID, LabResultID, or CallID. The unique record identifier for a data table is also called the table's **primary key**.

Figure 16.1 shows a simplified data table for a hypothetical cohort study (inspired by a real study [2]) of the association between neonatal jaundice and IQ score at age 5. Each row in the table corresponds to a study subject, and each column corresponds to an attribute of that subject. The dichotomous predictor is whether or not the subject had "Jaundice," and the continuous outcome is "IQ," which is the subject's IQ at age 5.

If the study data are limited to a **single table** such as the table in Figure 16.1, they are easily accommodated in a spreadsheet or statistical package. We often refer to a database consisting

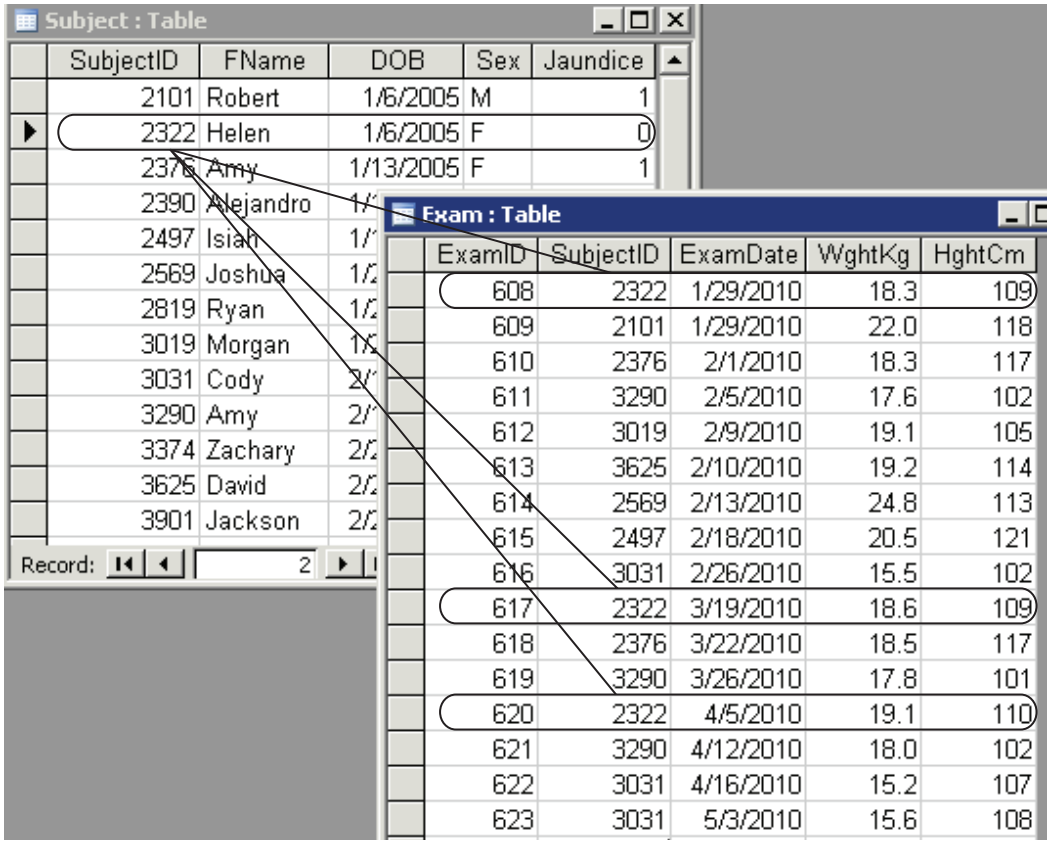
SubjectID	FName	DOB	Sex	Jaundice	ExamDate	WghtKg	HghtCm	IQ
2101	Robert	1/6/2005	M	1	1/29/2010	23.9	118	104
2322	Helen	1/6/2005	F	0	1/29/2010	18.3	109	94
2376	Amy	1/13/2005	F	1	3/22/2010	18.5	117	85
2390	Alejandro	1/14/2005	M	0				
2497	Isiah	1/18/2005	M	0	2/18/2010	20.5	121	74
2569	Joshua	1/23/2005	M	1	2/13/2010	24.8	113	115
2819	Ryan	1/26/2005	M	0				
3019	Morgan	1/29/2005	F	0	2/9/2010	19.1	105	105
3031	Cody	2/15/2005	M	0	4/16/2010	15.2	107	132
3290	Amy	2/16/2005	F	1	4/12/2010	18.0	102	125
3374	Zachary	2/21/2005	M	1				
3625	David	2/22/2005	M	1	2/10/2010	19.2	114	134
3901	Jackson	2/28/2005	M	0				

■ **FIGURE 16.1** Simplified data table for a cohort study of the association between neonatal jaundice and IQ score at age 5. The dichotomous predictor is "Jaundice," defined here as whether total bilirubin rose to 25 mg/dL or more in the first 2 days after birth, and the continuous outcome is "IQ," the subject's IQ score at age 5. Subjects 2390, 2819, 3374, and 3901 were not examined at age 5.

of a single, two-dimensional table as a "**flat file**." Many statistical packages have added features to accommodate more than one table, but at their core, most remain flat-file databases.

The need to include more than one table in a study database (and move from spreadsheet or statistical software to data management software) arises if the study tracks multiple lab results, medications, or other repeated measurements per study subject. A single data table with one row per study subject cannot accommodate a large and variable number of repeated measurements. The database should store medications, lab results, or other repeated measurements in separate tables distinct from the table of study subjects. A row in one of these separate tables corresponds to an individual measurement including, for example, the type of measurement, the measurement date/time, and the result or value of the measurement. One field in the row must include the subject identification number to link the measurement back to subject-specific fields. In this "**multi-table relational database**," the relationship between the table of subjects and the tables of measurements is termed **one-to-many**. Strictly speaking, the term relational has little to do with the between-table relationships. In fact, relation is the formal term from mathematical set theory for a data table (3, 4).

Although the subjects in our infant jaundice study received the IQ exam only once at age 5, most of them had other examinations during which, along with other measurements, height and weight were assessed. The height and weight data were used to calculate body mass index (BMI) and growth curve percentiles. (See "Extracting Data [Queries]" later in this chapter.) The best way to accommodate these data is in a separate table of examinations in which each row corresponds to a discrete examination, and the columns represent examination date, examination results, and the subject identification number to link back to information in the subject table such as sex, date of birth (DOB), and whether or not the child had neonatal jaundice (Figure 16.2). In this two-table database structure, querying the examination table for all exams performed within a particular time period requires searching a single exam date column. A change to a subject-specific field like date of birth is made in one place, and consistency is preserved. Fields holding personal identifiers such as name and date of birth appear only in the subject table, The other table(s) link back to this information via the subjectID. The database can still accommodate subjects (such as Alejandro, Ryan, Zachary, and Jackson) who have no exams.



Subject : Table

SubjectID	FName	DOB	Sex	Jaundice
2101	Robert	1/6/2005	M	1
2322	Helen	1/6/2005	F	0
2376	Amy	1/13/2005	F	1
2390	Alejandro	1/		
2497	Isiah	1/		
2569	Joshua	1/		
2819	Ryan	1/		
3019	Morgan	1/		
3031	Cody	2/		
3290	Amy	2/		
3374	Zachary	2/		
3625	David	2/		
3901	Jackson	2/		

Record: 2

Exam : Table

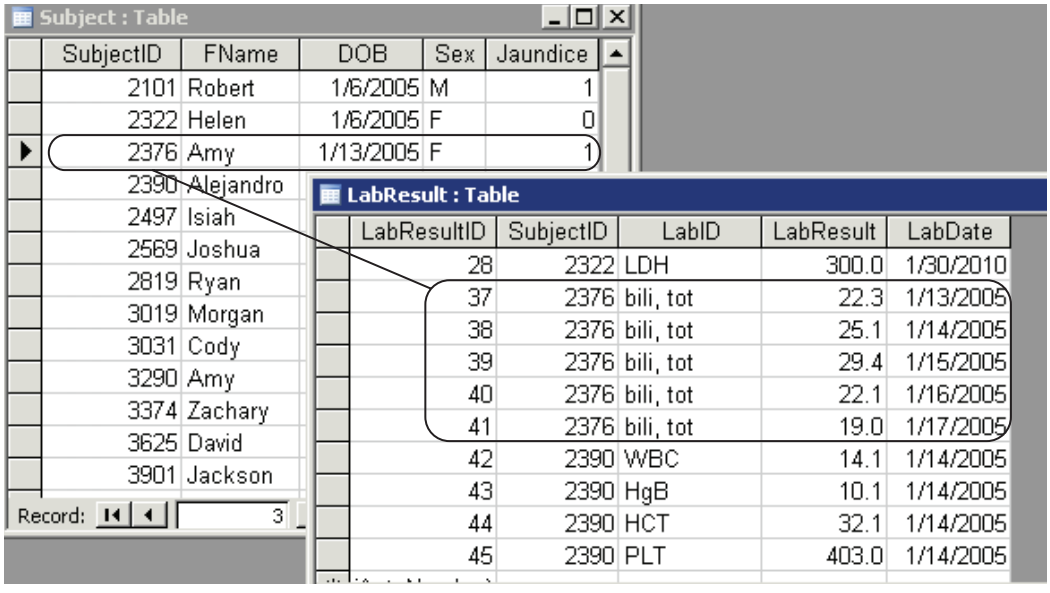
ExamID	SubjectID	ExamDate	WghtKg	HghtCm
608	2322	1/29/2010	18.3	109
609	2101	1/29/2010	22.0	118
610	2376	2/1/2010	18.3	117
611	3290	2/5/2010	17.6	102
612	3019	2/9/2010	19.1	105
613	3625	2/10/2010	19.2	114
614	2569	2/13/2010	24.8	113
615	2497	2/18/2010	20.5	121
616	3031	2/26/2010	15.5	102
617	2322	3/19/2010	18.6	109
618	2376	3/22/2010	18.5	117
619	3290	3/26/2010	17.8	101
620	2322	4/5/2010	19.1	110
621	3290	4/12/2010	18.0	102
622	3031	4/16/2010	15.2	107
623	3031	5/3/2010	15.6	108

■ **FIGURE 16.2** The two-table infant jaundice study database has a table of study subjects in which each row corresponds to a single study subject and a table of examinations in which each row corresponds to a particular examination. For example, subject 2322 is identified as Helen, date of birth 1/6/2005, in the first table, and is shown with data from three exams in the anonymous second table. Since a subject can have multiple examinations, the relationship between the two tables is one-to-many. The SubjectID field in the exam table links the exam-specific data to the subject-specific data.

Detailed tracking of lab results also requires a separate table. Neonatal jaundice is presented here as a dichotomous subject-specific field. If the investigators need the entire trajectory of bilirubin levels after birth, then the database should include a separate lab result table with one record per lab result and fields for date/time of lab test, lab test type (total bilirubin), test result (bilirubin level), and subjectID for linking back to the subject-specific information (Figure 16.3).

A study's administrative data such as call logs, visit schedules, and reimbursement records also require multiple separate tables. In the infant jaundice study, multiple calls were made to the parents of each study subject. It would be difficult or impossible to track these calls in a data table with one row per study subject. Instead, a separate table had one row per call with a subjectID field linking back to the study subject about whom the call was made.

Structuring the database with multiple related tables, instead of trying to accommodate the data in a very wide and complex single table, is called **normalization**. Some data managers refer to normalization as converting from one or a few "short–fat" tables to many "tall–skinny" tables (1). Normalization eliminates redundant storage and the opportunity for inconsistencies. Relational database software can be set to maintain **referential integrity**, meaning that it will not allow creation of an exam, lab result, or call record for a subject who does not already exist in the subject table. Similarly, it prevents deletion of a subject unless all of that subject's exams, lab results, and calls have also been deleted.



■ **FIGURE 16.3** The linkage between the table of subjects and the table of lab results. The lab results capture the trajectory of Amy's total bilirubin over her first 5 days after birth.

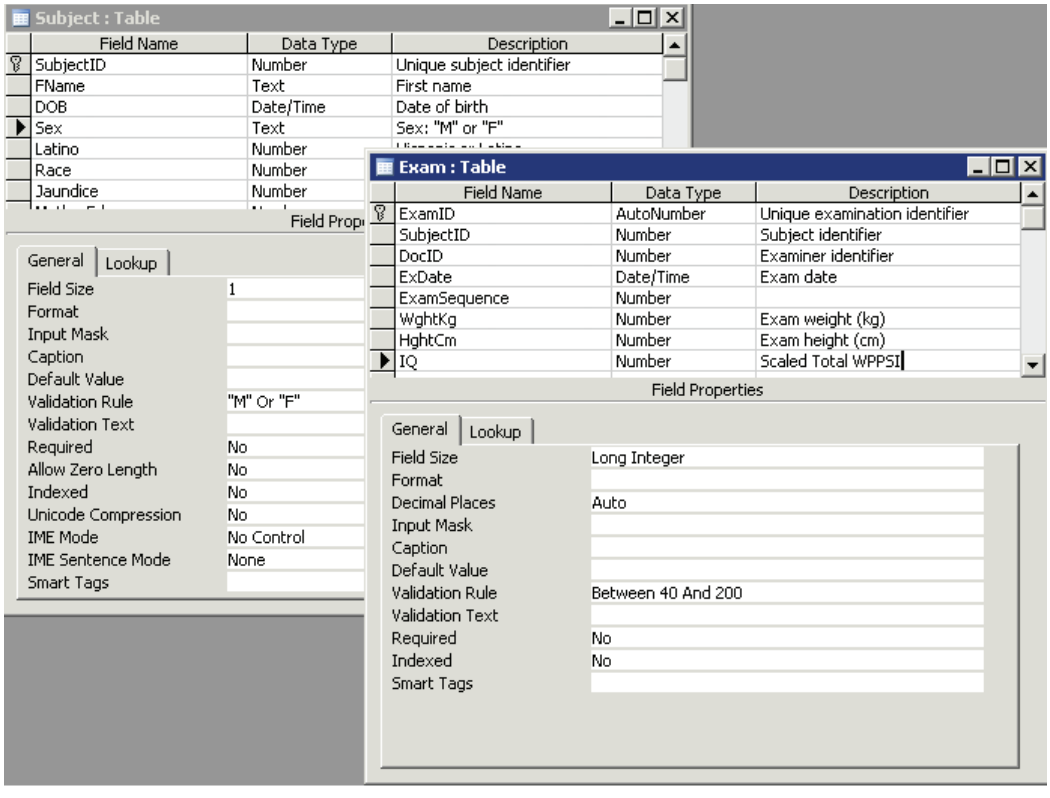
Data Dictionaries, Data Types, and Domains

So far we have seen tables only in the "datasheet" view. Each column or field has a name and, implicitly, a data type and a definition. In the "Subject" table of Figure 16.2, "FName" is a text field that contains the subject's first name; "DOB" is a date field that contains the subject's birth date, and "Jaundice" is a yes/no field that indicates whether the bilirubin exceeded 25 mg/dL in the first 2 days after birth. In the "Exam" table, "WghtKg" is a real-number weight in kilograms and "IQ" is an integer IQ score. The **data dictionary** makes these column definitions explicit. Figure 16.4 shows the subject and exam tables in table design (or "data dictionary") view. Note that the data dictionary is itself a table with rows representing fields and columns for field name, data type, and field description. Since the data dictionary is a table of information about the database itself, it is referred to as **metadata**. Although Figure 16.4 displays two data dictionaries, one for the "Subject" table and one for the "Exam" table, the entire database can be viewed as having a single data dictionary rather than one dictionary for each table. For each field in the database, the single data dictionary requires specification of the field's table name in addition to the field name, field type, field description, and range of allowed values.

Each field also has a **domain** or range of allowed values. For example, the allowed values for the "Sex" field are "M" and "F". The software will not allow entry of any other value in this field. Similarly the "IQ" field allows only integers between 40 and 200. Data managers for clinical trials generally refer to validation rules as "edit checks" (1). Creating validation rules to define allowed values affords some protection against data entry errors. Some of the data types come with automatic validation rules. For example, the database management software will always reject a date of April 31.

Variable Names

Most spreadsheet, statistical, and database management programs allow long column headings or variable names. Philosophies and naming conventions abound. We recommend variable names that are short enough to type quickly, but long enough to be self-explanatory. Although they are often allowed by the software, we recommend avoiding spaces and special characters



■ **FIGURE 16.4** The table of study subjects ("Subject") and the table of measurements ("Exam") in "data dictionary" view. Each variable or field has a name, a data type, a description, and a domain or set of allowed values.

in variable names. We distinguish separate words in a variable name by using aptly named "InterCaps," but others may prefer using an underscore character. It is generally better to use a variable name that describes the field rather than its location on the data collection form (e.g., "EverSmokedCigarettes" or "EverSmo," instead of "Question1"). Most software packages allow users to designate a longer, more descriptive, and easier to read **variable label** to use on data entry forms and reports instead of the compact variable name.

Common Data Elements

Several funding and regulatory organizations have launched initiatives to develop common data elements for study databases in specific areas of clinical research. These organizations include government agencies such as the National Institute for Neurologic Disorders and Stroke (5), the National Cancer Institute (6), the United States Food and Drug Administration (7), and the European Medicines Agency and nongovernmental, nonprofit associations such as the Clinical Data Interchange Standards Consortium (CDISC) (8).

The rationale is that research studies in the same clinical area often need to collect the same measurements. Standardizing record structures, field names/definitions, data types/formats, and data collection forms (case report forms) will eliminate the problem of "reinventing the wheel" as often occurs in new research studies (5) and enable sharing and combining data across multiple separate studies. This entails establishing a data dictionary and a set of data collection instruments with accompanying instructions that all investigators in a particular area of research are encouraged to use. Part of thorough scholarship in one's chosen research area is awareness of existing data standards.

DATA ENTRY

Whether the study database consists of one or many tables and whether it uses spreadsheet, statistical, or database management software, a mechanism for **populating the data tables** (entering the data) is required.

Keyboard Transcription

Historically, the common method for populating a study database has been to first collect data on **paper forms**. In clinical trials, a paper data collection form corresponding to a specific subject is commonly called a **case report form** or **CRF**. The investigator or a member of the research team may fill out the paper form or, in some cases, the subject himself fills it out. Study personnel can then transcribe the data via keyboard from the paper forms into the computer tables. Transcription can occur directly into the data tables (e.g., the response to question 3 on subject 10 goes into the cell at row 10, column 3) or via on-screen forms designed to make data entry easier and including automatic data validation checks. Transcription should occur as shortly as possible after the data collection, so that the subject and interviewer or data collector are still available if responses are found to be missing or out of range. Also, as discussed later in this chapter, monitoring for data problems (e.g., outlier values) and preliminary analyses can only occur once the data are in the computer database.

If transcribing from paper forms, the investigator may consider **double data entry** to ensure the fidelity of the transcription. The database program compares the two values entered for each variable and presents a list of values that do not match. Discrepant entries are then checked on the original forms and corrected. Double data entry identifies data entry errors at the cost of doubling the time required for transcription. An alternative is to double-enter a random sample of the data. If the error rate is acceptably low, double data entry is unlikely to be worth the effort and cost for the remaining data.

Distributed Data Entry

If data collection occurs at multiple sites, the sites can e-mail or fax paper forms to a central location for transcription into the computer database, but this practice is increasingly rare. More commonly, the data are transcribed at the sites directly into the study database via online forms. If Internet connectivity is a problem, data are stored on a local computer at the site and transmitted online or via a portable memory device such as a USB drive. Government regulations require that electronic health information be either de-identified or transmitted securely (e.g., encrypted and password-protected).

Electronic Data Capture

Primary data collection onto paper will always have its place in clinical research; a fast and user-friendly way to capture data on a nonvolatile medium is using pen and paper. However, hand writing data onto a paper form is increasingly rare. In general, research studies should collect data primarily using **online forms**. In clinical trials, electronic forms are called electronic case report forms **(eCRFs)**. Data entry via online forms has many advantages:

- The data are **keyed directly** into the data tables without a second transcription step, removing that source of error.
- The computer form can include **validation checks** and provide immediate feedback when an entered value is out of range.
- The computer form can also incorporate **skip logic**. For example, a question about packs per day appears only if the subject answered "yes" to a question about cigarette smoking.
- The form may be viewed and data entered on **portable**, **wireless devices** such as a tablet (iPad), smartphone, or notebook computer.

When using online forms for electronic data capture, it sometimes makes sense to print out a paper record of the data immediately after collection. This is analogous to printing out a receipt after a transaction at an automated teller machine. The printout is a paper "snapshot" of the record immediately after data collection and may be used as the original or source document if a paper version is required.

Coded Responses Versus Free Text

Defining a variable or field in a data table includes specifying its range of allowed values. For subsequent analysis, it is preferable to limit responses to a range of coded values rather than allowing free text responses. This is the same as the distinction made in Chapter 15 between "closed-ended" and "open-ended" questions. If the range of possible responses is unclear, initial data collection during the pretesting of the study can allow free text responses that will subsequently be used to develop coded response options.

The set of response options to a question should be **exhaustive** (all possible options are provided) and **mutually exclusive** (no two options can both be correct). A set of mutually exclusive response options can always be made collectively exhaustive by adding an "other" response. Online data collection forms provide three possible formats for displaying the mutually exclusive and collectively exhaustive response options: drop-down list, pick list (field list), or option group (Figure 16.5). These formats will be familiar to any research subject or data entry person who has worked with an online form. Note that the drop-down list saves screen space but will not work if the screen form will be printed to paper for data collection, because the response options will not be visible.

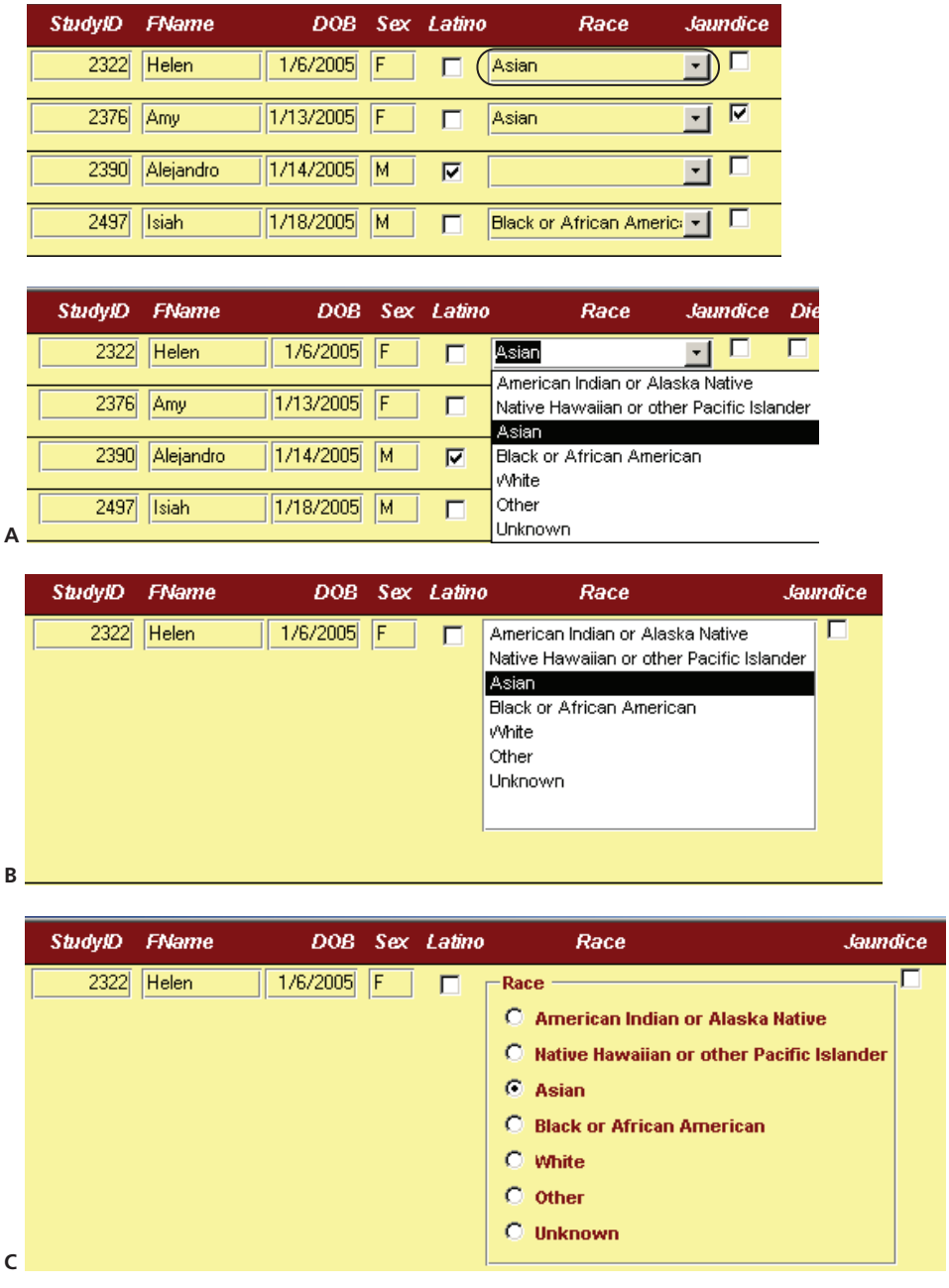
A question with a set of mutually exclusive responses corresponds to a single field in the data table. In contrast, the responses to an "All that apply" question are not mutually exclusive. They correspond to as many yes/no fields as there are possible responses. By convention, response options for "All that apply" questions use square check boxes rather than the round radio buttons used for option groups with mutually exclusive responses. As discussed in Chapter 15, we discourage "All that apply" questions and prefer to require a yes or no response to each item. Otherwise an unmarked response could either mean "does not apply" or "not answered." In coding yes/no (dichotomous) variables, make 0 represent *no* or *absent*, and 1 represent *yes* or *present*. With this coding, the average value of the variable is interpretable as the proportion with the attribute.

Importing Measurements and Laboratory Results

Much study information, such as baseline demographic information in the hospital registration system, lab results in the laboratory's computer system, and measurements made by dual energy x-ray absorptiometry (DEXA) scanners and Holter monitors, is already in digital electronic format. Where possible, these data should be imported directly into the study database to avoid the labor and potential transcription errors involved in re-entering data. For example, in the study of infant jaundice, the demographic data and contact information are obtained from the hospital database. Computer systems can almost always produce tab-delimited or fixed-column-width text files that the database software can import. In clinical trials, this type of batch-uploaded information is referred to as "non-CRF (case report form) data" (1).

Data Management Software

Now that we have discussed data tables and data entry, we can make the distinction between the study database's back end and front end. The **back end** consists of the data tables themselves. The **front end** or "interface" consists of the online forms used for entering, viewing, and editing the data. Table 16.1 lists some software applications used in data management for clinical research. Simple study databases consisting of a single data table can use spreadsheet or



■ **FIGURE 16.5** Formats for entering from a mutually exclusive, collectively exhaustive list of responses. The drop-down list (A; dropped down in lower panel) saves screen space but will not work if the screen form will be printed to paper for data collection. Both the pick list (which is just a drop-down list that is permanently dropped down; B) and the option group (C) require more screen space, but will work if printed.

statistical software for the back-end data table and the study personnel can enter data directly into the data table's cells, obviating the need for front-end data collection forms. More complex study databases consisting of multiple data tables require **relational database** software to maintain the back-end data tables. If the data are collected first on paper forms, entering the data will require transcription into online forms.

TABLE 16.1 SOME SOFTWARE USED IN RESEARCH DATA MANAGEMENT

- Spreadsheet
  - Microsoft Excel
  - Google Drive Spreadsheet*
  - Apache OpenOffice Calc*
- Statistical Analysis
  - Statistical Analysis System (SAS)
  - Statistical Package for the Social Sciences (SPSS)
  - Stata
  - R*
  - EpiInfo* (for Windows only)
- Integrated Desktop Database Systems
  - Microsoft Access (for Windows only)
  - Filemaker Pro
- Relational Database Systems
  - Oracle
  - SQL Server
  - MySQL*
  - PostgreSQL*
- Integrated Web-Based Platforms for Research Data Management
  - Research Electronic Data Capture* (REDCap—academic only, hosted by investigator's institution)
  - QuesGen (primarily academic, vendor hosted)
  - MediData RAVE (primarily non-academic corporate, vendor hosted)
  - Oracle InForm (non-academic corporate, company hosted)
  - Datalabs EDC (corporate, vendor hosted)
  - OnCore
  - OpenClinica
- Online Survey Tools
  - SurveyMonkey
  - Zoomerang
  - Qualtrics

* Free

As discussed in Chapter 15, several tools, including SurveyMonkey, Zoomerang, and Qualtrics, exist for developing online surveys that will be e-mailed to study participants or posted on the study website. All of these tools provide multiple question format options, skip logic, and the capability to aggregate, report on, and export survey results.

Some **statistical packages**, such as SAS, have developed data entry modules. **Integrated desktop database** programs, such as Microsoft Access and Filemaker Pro, also provide extensive tools for the development of on-screen forms.

Research studies increasingly use integrated, Web-enabled, research data management platforms. **REDCap** (Research Electronic Data Capture) is a Web-based research data collection system developed by an academic consortium based at Vanderbilt University. It enables researchers to build data entry forms, surveys, and surveys with attached data entry forms. REDCap is made available to academic investigators only and must be hosted at the investigator's institution. This is an outstanding "do-it-yourself" tool for beginning academic investigators that allows rapid development of surveys and on-screen data collection forms. It also provides access to a repository of downloadable data collection instruments. As with all

do-it-yourself Web development tools, options for customization and advanced functionality are limited. A REDCap database consists of a single table containing one row for each of a fixed number of user-defined "events" per study subject. It does not permit detailed tracking of a large and variable number of repeated measurements per study subject, such as lab results, vital signs, medications, or call logs. REDCap also cannot do sophisticated data validation, querying (see later in this chapter), or reporting, but does make export into statistical packages easy.

Full-featured, Web-based research data management platforms such as **QuesGen**, **MediData RAVE**, or **Oracle InForm** can accommodate complex data structures and provide sophisticated data validation, querying, and reporting. The companies that provide these tools also provide support and configuration assistance. While there may be some additional cost involved, these solutions are worth considering when the do-it-yourself tools lack the sophistication to meet the study's requirements.

EXTRACTING DATA (QUERIES)

Once the database has been created and data entered, the investigator will want to **organize**, **sort**, **filter**, and **view** ("query") the data. **Queries** are used for monitoring data entry, reporting study progress, and ultimately analyzing the results. The standard language for manipulating data in a **relational database** is called Structured Query Language or **SQL** (pronounced "**sequel**"). All relational database software systems use one or another variant of SQL, but most provide a graphical interface for building queries that makes it unnecessary for the clinical researcher to learn SQL.

A query can **join** data from two or more tables, display only selected fields, and filter for records that meet certain criteria. Queries can also calculate values based on raw data fields from the tables. Figure 16.6 shows the results of a query on our infant jaundice database that filters for boys examined in February and calculates age in months (from birth date to date of exam) and BMI (from weight and height). The query also uses a sophisticated table-lookup function to calculate growth curve percentile values for the child's BMI. Note that the result of a query that joins two tables, displays only certain fields, selects rows based on special criteria, and calculates certain values still looks like a table in datasheet view. One of the tenets of the relational database model is that operations on tables produce table-like results. The data in Figure 16.6 are easily exported to a statistical analysis package. Note that no personal identifiers are included in the query.

Identifying and Correcting Errors in the Data

The first step toward avoiding errors in the data is testing the data collection and management system as part of the overall pretesting for the study. The entire system (data tables, data entry

Boys Examined in February (With BMI Percentile) : Select Query

SubjectID	Sex	ExamDate	AgeMonths	WghtKg	HghtCm	BMIcalc	BMIPerc
2497	M	2/18/2010	61	20.5	121	14.0	8
2569	M	2/13/2010	60	24.8	113	19.4	99
3031	M	2/26/2010	59	15.5	102	14.9	33
3625	M	2/10/2010	59	19.2	114	14.7	26
4430	M	2/23/2010	59	35.0	100	35.0	100
5305	M	2/23/2010	60	20.5	116	15.2	43
5310	M	2/24/2010	60	19.6	115	14.8	28

Record: 5 of 7

FIGURE 16.6 A query in datasheet view that filters for boys examined in February and calculates age in months (from birth date to date of exam) as well as body mass index (BMI) from weight and height. The query also uses a sophisticated table-lookup function to calculate growth curve percentile values for the child's BMI. For SubjectID 4430, the 100th percentile value associated with the BMI of 35.0 should trigger investigation of the outlier as a possible data entry error.

forms, and queries) should be tested using dummy data. For clinical trials that will be used in an FDA submission, this is a regulatory requirement under Code of Federal Regulations, Chapter 21, Part 11 (21 CFR 11) (9).

We have discussed ways to enhance the fidelity of keyboard transcription or electronic data capture once data collection begins. Values that are outside the permissible range should not get past the data entry process. However, the database should also be queried for missing values and outliers (extreme values that are nevertheless within the range of allowed values). For example, a weight of 35 kg might be within the range of allowed values for a 5-year-old, but if it is 5 kg greater than any other weight in the data set, it bears investigation. Many data entry systems are incapable of doing cross-field validation, which means that the data tables may contain field values that are within the allowed ranges but inconsistent with one another. For example, it would be highly unlikely for a 35 kg 5-year-old to have a height of 100 cm. While the weight and height values are both within the allowed ranges, the weight (extremely high for a 5-year-old) is inconsistent with the height (extremely low for a 5-year-old). Such an inconsistency can be suspected using a query like the one depicted in Figure 16.6.

Missing values, outliers, inconsistencies, and other data problems are identified using queries and communicated to the study staff, who can respond to them by checking original source documents, interviewing the participant, or repeating the measurement. If the study relies on paper source documents, any resulting changes to the data should be highlighted (e.g., in red ink), dated, and signed. As discussed later in this chapter, electronic databases should maintain an audit log of all data changes.

If data are collected by several investigators from different locations, means and medians should be compared across investigators and sites. Substantial differences by investigator or site can indicate systematic differences in measurement or data collection.

Data editing and cleaning should give higher priority to more important variables. For example, in a randomized trial, the most important variable is the outcome, so missing data and errors should be minimized. In contrast, errors in other variables, such as the date of a visit, may not substantially affect the results of analyses. Data editing is an iterative process; after errors are identified and corrected, editing procedures should be repeated until very few important errors are identified. At this point, for some studies, the edited database may be declared final or "**locked**," so that no further changes are permitted (1).

■ ANALYSIS OF THE DATA

Analyzing the data often requires creating new, derived variables based on the raw field values in the data set. For example, continuous variables may be made dichotomous (e.g., BMI > 25 defined as overweight), new categories created (specific drugs grouped as antibiotics), and calculations made (years of smoking × number of packs of cigarettes per day = pack years). Missing data should be handled consistently. "Don't know" may be recoded as a special category, combined with "no," or excluded as missing. If the study uses database software, queries can be used to derive the new variables prior to export to a statistical analysis package. This is especially important for variables like the percentiles in Figure 16.6 that require complex programming or a separate "look-up" table. Alternatively, derivation of the new fields can occur in the statistical package itself. Many investigators are more familiar with statistical packages than database programs and prefer to calculate derived variables after export.

■ CONFIDENTIALITY AND SECURITY

If research study subjects are also clinic or hospital patients, their identifying information is protected under the Privacy Rule of the Health Insurance Portability and Accountability Act (**HIPAA**) (10); that said, regardless of whether the subjects are also patients, the investigator is obligated both ethically and legally to protect their confidentiality. The database should assign

each subject a unique subject identifier (subjectID) that has no meaning external to the study database (i.e., the subjectID should not incorporate the subject's name, initials, birth date, or medical record number). Any database fields that do contain personal identifiers should be deleted prior to sharing the data. If the database uses multiple tables, the personal identifiers can be kept in a separate table. Study databases that contain personal identifiers must be maintained on secure servers accessible only to authorized members of the research team, each of whom will have a user ID and password. Dedicated Web-based research data management platforms such as REDCap and QuesGen allow designation of fields containing subject identifiers. Different user roles can allow or prohibit exporting, changing, or even viewing these specially designated fields.

The database system should **audit** all **data entry and editing**. Auditing allows determination of when a data element was changed, who made the change, and what change was made. For new drug trials, this is a regulatory requirement (9). Dedicated Web-based research platforms such as REDCap, QuesGen, and MediData RAVE automatically provide user validation and auditing.

The study database must be **backed up** regularly and **stored off-site**. Periodically the back-up procedure should be tested by restoring a backed-up copy of the data. As with user validation and auditing, hosted platforms like REDCap, QuesGen, and MediData RAVE automatically provide backups and data security. At the end of the study the original data, data dictionary, final database, and the study analyses should be **archived** for future use. Such archives can be revisited in future years, allowing the investigator to respond to questions about the integrity of the data or analyses, perform further analyses to address new research questions, and share data with other investigators.

SUMMARY

1. The study **database** consists of one or more data tables in which the **rows** correspond to **records** (e.g., study subjects) and the **columns** correspond to **fields** (attributes of the records).
2. Identifying study subjects with a unique **subjectID** that has no meaning external to the study database enables the "de-linking" of study data from personal identifiers for purposes of maintaining **confidentiality**. Databases that contain **personal identifiers** must be stored on secure servers, with access restricted and audited.
3. Accommodating a variable number of **repeated measurements** per study subject, such as lab results or medications, requires **normalization** of the measurement data into separate tables in which each row corresponds to a **measurement** rather than an individual study subject.
4. The study database may also store **administrative data** such as **call logs**, **exam schedules**, and **reimbursement records**.
5. The **data dictionary** specifies the **name**, **data type**, **description**, and **range of allowed values** for all the fields in the database.
6. The **data entry system** is the means by which the data tables are populated; **electronic data capture** via online forms is replacing transcription from paper forms for data entry.
7. A **spreadsheet** or **statistical package** is adequate only for the simplest **study databases**; complex databases require the creation of a **relational database** using **database management software** based on Structured Query Language (**SQL**).
8. **Database queries** sort and filter the data as well as calculate values based on raw data fields. Queries are used to **monitor** data entry, provide **reports** on study progress, and format the results for **analysis**.
9. Loss of the database must be prevented by regular **backups** and **off-site storage**, and by **archiving** copies of key versions of the database for future use.

REFERENCES

1. Prokscha S. *Practical guide to clinical data management*, 3rd ed. Boca Raton: CRC Press, 2012.
2. Newman TB, Liljestrand P, Jeremy RJ, et al. Outcomes among newborns with total serum bilirubin levels of 25 mg per deciliter or more. *N Engl J Med* 2006;354(18):1889–1900.
3. Codd EF. A relational model of data for large shared data banks. *Communications of the ACM* 1970;13(6):377–387.
4. Date CJ. *An introduction to database systems*, 7th ed. Reading, Mass: Addison-Wesley, 2000.
5. Grinnon ST, Miller K, Marler JR, et al. National Institute of Neurological Disorders and Stroke common data element project—approach and methods. *Clin Trials* 2012;9(3):322–329.
6. NCI. *The National Cancer Institute Cancer Data Standards Registry and Repository*. 2012. Available from: https://cabig.nci.nih.gov/concepts/caDSR/, accessed 9/29/12.
7. FDA. Driving biomedical innovation: initiatives to improve products for patients. October, 2011. Available from: http://www.fda.gov/downloads/AboutFDA/ReportsManualsForms/Reports/UCM274464.pdf, accessed 1/29/13.
8. CDISC. The Clinical Data Interchange Standards Consortium Study data tabulation model. 2012. Available from: http://www.cdisc.org/sdtm, accessed 1/29/2013.
9. DHHS. Guidance for industry: computerized systems used in clinical trials. May, 2007. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM070266.pdf, accessed 1/29/2013.
10. DHHS. Protecting personal health information in research: understanding the HIPAA Privacy Rule. 2003. Available from: http://privacyruleandresearch.nih.gov/pr_02.asp, accessed 1/29/2013.

CHAPTER 17

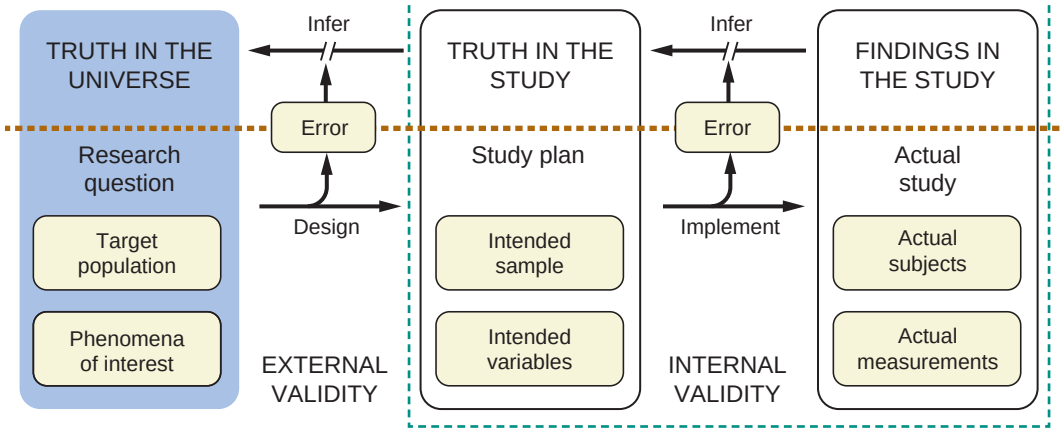
Implementing the Study and Quality Control

Deborah G. Grady and Stephen B. Hulley

Most of this book has dealt with the left-hand side of the clinical research model, addressing matters of design (Figure 17.1). In this chapter we turn to the right hand, **implementation** side. Even the best of plans thoughtfully assembled in the armchair may work out differently in practice. Skilled research staff may be unavailable, study space less than optimal, participants less willing to enroll than anticipated, the intervention poorly tolerated, and the measurements challenging. The conclusions of a well-designed study can be marred by ignorance, carelessness, lack of training and standardization, and other errors in finalizing and implementing the protocol.

Successful study implementation begins with **assembling resources** including space, staff, and financial management for **study start-up**. The next task is to **finalize the protocol** through a process of **pretesting** recruitment, measurement, and intervention plans in an effort to avoid the need for **protocol revisions** after data collection has begun. The study is then carried out with a systematic approach to **quality control** of **clinical** and **lab procedures** and of **data management**, following the FDA-endorsed principles of **Good Clinical Practice (GCP)**.

Some of the strategies in this chapter pertain to major studies with large research teams distributed across multiple centers that are led by senior investigators. However, the information is also relevant to beginning investigators who may be co-investigators in these large studies, or PI on a smaller study.



■ **FIGURE 17.1** This chapter focusses on the area within the dashed green line: implementing a research project.

■ ASSEMBLING RESOURCES

Space

It is possible to conduct some clinical research studies completely online, using Web-based interactive systems, mailed interventions (such as drugs or devices), remote monitoring, home visits for measurement, and online data entry. However, most research studies still require physical space to conduct study visits and make measurements. This space should be accessible, attractive, and sufficient. Failing to successfully negotiate for space early in the study planning process can result in difficulty enrolling participants, poor adherence to study visits, incomplete data, and unhappy staff. **Clinical research space** must be easily accessible to participants and have adequate available parking. The space should be welcoming, comfortable, and spacious enough to accommodate staff, measurement equipment, and storage of study drug and study-related files. If there will be a physical examination, provision for privacy and hand washing must be available. If the participants must go to other places for tests (such as the hospital laboratory or radiology department) these should be easily accessible. In some studies, such as those that enroll sick patients or deliver interventions that could be dangerous, access to cardiopulmonary resuscitation teams and equipment may be required.

Many academic medical centers have **clinical research centers** that provide fully equipped research space staffed by experienced research staff. Clinical research centers often include the ability to make specialized measurements (such as caloric intake, bone density, and insulin clamp studies), and may provide access to other services (such as participant recruitment, database management, and statistical analysis). These centers provide an excellent option for carrying out clinical and translational research, but generally require separate application and review procedures, and reimbursement for services.

The Research Team

Research teams range in size from small—often just the investigator and a part-time research assistant—to multiple full-time staff for large studies. Regardless of size, all research teams must accomplish similar activities and fill similar roles, which are described in Table 17.1. Often, one person carries out several of these activities. However, some of these duties require special expertise, such as statistical programming and analyses. Some team members, such as the financial and human resources managers, are generally employed by the university or medical center, and provided by the investigator's department or unit. Regardless of the size of the study team, the **principal investigator (PI)** must make sure that each of the functions described in Table 17.1 is carried out.

After deciding on the number of team members and the distribution of duties, the next step is to work with a departmental administrator to find qualified and experienced **job applicants**. This can be difficult, because formal training for some research team members is variable, and job requirements vary from one study to the next. For example, the crucial position of project director may be filled by a person with a background in nursing, pharmacy, public health, laboratory services, or pharmaceutical research, and the duties of this position can vary widely.

Most universities and medical centers have formal methods for posting job openings, but other avenues, such as newspaper and Web-based advertisements, can be useful. The safest approach is to find staff of known competence; for example, someone working for a colleague whose project has ended. It is also common to negotiate with colleagues to hire their experienced staff part-time. Some academic medical centers or units within the medical center support a pool of experienced research coordinators and other staff who can be hired part-time.

TABLE 17.1 FUNCTIONAL ROLES FOR MEMBERS OF A RESEARCH TEAM*

ROLE	FUNCTION	COMMENT
Principal investigator	Ultimately responsible for the design, funding, staffing, conduct, and quality of the study, and for reporting findings	
Project director/ clinic coordinator	Provides day-to-day management of all study activities	Experienced, responsible, meticulous, and with good interpersonal and organizational skills
Recruiter	Ensures that the desired number of eligible participants are enrolled	Knowledgeable and experienced with a range of recruitment techniques
Research assistant/clinic staff	Carries out study visit procedures and makes measurements	Physical examination or other specialized procedures may require special licenses or certification
Quality control coordinator	Ensures that all staff follow standard operating procedures (SOPs), and oversees quality control	Observes study procedures to ensure adherence to SOPs, may supervise audit by external groups such as U.S. Food and Drug Administration (FDA)
Data manager	Designs, tests, and implements the data entry, editing, and management system	
Programmer/ analyst	Produces study reports that describe recruitment, adherence, and data quality; conducts data analyses	Works under the supervision of the principal investigator (PI) and statistician
Statistician	Collaborates on study design, estimates sample size and power, designs analysis plan and data and safety monitoring guidelines, interprets findings	Often plays a major role in overall study design, conduct, interim monitoring, data analyses, and presentation of results
Administrative assistant	Provides administrative support, sets up meetings, and so on	
Financial manager	Prepares budget and manages expenditures	Provides projections to help manage budget
Human resources manager	Assists in preparing job descriptions, hiring, evaluations	Helps manage personnel issues and problems

*In small studies, one person may take on several of these roles; others, such as the financial and human resources managers, are usually provided by the department and shared with other faculty.

Leadership and Team-Building

The quality of a study that involves more than one person on the research team begins with the integrity and **leadership of the PI**. The PI should ensure that all staff are properly trained and certified to carry out their duties. He should clearly convey the message that protection of human subjects, maintenance of privacy, completeness and accuracy of data, and fair presentation of research findings are paramount. He cannot watch every measurement made by colleagues and staff, but if he creates a sense that he is broadly aware of all study activities and feels strongly about human subjects' protection and the quality of the data, most people will respond in kind. It is helpful to meet with each member of the team from time to time, expressing appreciation and discussing problems and solutions. A good leader is adept at delegating authority appropriately and at the same time setting up a hierarchical system of supervision that ensures sufficient oversight of all aspects of the study.

From the outset of the planning phase, the investigator should lead regular **staff meetings** with all members of the research team. Meetings should have the agenda distributed in advance, with progress reports by listed individuals who have been given responsibility for specific areas of the study. These meetings provide an opportunity to discover and solve problems, and to involve everyone in the process of developing the project and conducting the research. Staff meetings are enhanced by scientific discussions and updates related to the project. Regular staff meetings are a great source of morale and interest in the goals of the study and provide "on-the-job" education and training.

Most research-oriented universities and medical centers provide a wide range of **institutional resources** for conducting clinical research. These include human resources and financial management services, consultation services, and centralized clinical research centers that provide space and experienced research staff. Many universities also have core laboratories where specialized measurements can be performed, centralized space and equipment for storage of biologic specimens or images, centralized database management services, professional recruitment centers, expertise regarding U.S. Food and Drug Administration (FDA) and other regulatory issues, and libraries of study forms and documents. This infrastructure may not be readily apparent in a large sprawling institution, and investigators should seek to become familiar with their local resources before trying to do it themselves.

Study Start-Up

At the beginning of the study, the PI must finalize the budget, develop and sign any contracts that are involved, define staff positions, hire and train staff, obtain institutional review board (IRB) approval, write the operations manual, develop and test forms and questionnaires, develop and test the database, and plan participant recruitment. This period of study activity before the first participant is enrolled is referred to as **study start-up**, and requires intensive effort. Adequate time and planning for study start-up are important to the conduct of a high-quality study.

Adequate funding for conducting the study is crucial. The **budget** will have been prepared at the time the proposal is submitted for funding, well in advance of starting the study (Chapter 19). Most universities and medical centers employ staff with financial expertise to assist in the development of budgets (the **preaward manager**). It is a good idea to get to know this person well, to respect his or her stress level around deadline times by meeting timetable goals, and to thoroughly understand regulations related to various sources of funding.

In general, the rules for spending NIH and other public funds are considerably more restrictive than for industry or foundation funding. The total amount of the budget usually cannot be increased if the work turns out to be more costly than predicted, and shifting money across categories of expense (e.g., personnel, equipment, supplies, travel) or substantial reductions in the percent effort of key personnel generally require approval by the sponsor. Universities and medical centers typically employ financial personnel whose main responsibility is to ensure that funds available to an investigator through grants and contracts are spent appropriately. This **postaward manager** should prepare regular reports and projections that allow the investigator to make adjustments in the budget to make the best use of the available finances during the life of the study, ensuring that the budget will not be overdrawn at the end of the study. Having a modest surplus at the end of the study can be a good thing, as sponsors often approve "**no-cost extensions**" that allow the use of the surplus funds after the formal end of the study period to complete or extend the work described in the scope of the award.

The budget for a study supported by a pharmaceutical company is part of a contract that incorporates the protocol and a clear delineation of the tasks to be carried out by the investigator and the sponsor. **Contracts** are legal documents that obligate the investigator to activities and describe the timing and amount of payment in return for specified "**deliverables**," such as meeting recruitment milestones and submitting progress reports. University or medical center

lawyers are needed to help develop such contracts and ensure that they protect the investigator's intellectual property rights, access to data, publication rights, and so forth. However, lawyers are generally unfamiliar with the tasks required to complete a specific study, and input from the investigator is crucial, especially with regard to the scope of work and deliverables.

Institutional Review Board Approval

The **IRB** must approve the study protocol, consent form, and recruitment materials before recruitment can begin (Chapter 14). Investigators should be familiar with the requirements of their local IRB and the time required to obtain approval. IRB staff are generally very helpful in these matters, and should be contacted early on to discuss any procedural issues and design decisions that affect study participants.

Operations Manual and Forms Development

The study protocol is commonly expanded to create the **operations manual**, which includes the protocol, information on study organization and policies, and a detailed version of the methods section of the study protocol (Appendix 17A). It specifies exactly how to recruit and enroll study participants, and describes all activities that occur at each visit—how randomization and blinding will be achieved, how each variable will be measured, quality control procedures, data management practices, the statistical analysis plan, and the plan for data and safety monitoring (Chapter 11). It should also include all of the questionnaires and forms that will be used in the study, with instructions on contacting the study participants, carrying out interviews, completing and coding study forms, entering and editing data, and collecting and processing specimens. An operations manual is essential for research carried out by several individuals, particularly when there is collaboration among investigators in more than one location. Even when a single investigator does all the work himself, written operational definitions help reduce random variation and changes in measurement technique over time.

Design of the **data collection forms** will have an important influence on the quality of the data and the success of the study (Chapter 16). Before the first participant is recruited, the forms should be pretested. Any entry on a form that involves judgment requires explicit operational definitions that should be summarized briefly on the form itself and set out in more detail in the operations manual. The items should be coherent and their sequence clearly formatted with skip patterns (see Appendix 15). Pretesting will ensure clarity of meaning and ease of use. Labeling each page with the date, name, and ID number of the subject and staff safeguards the integrity of the data. Web-based digital forms, handheld computers, personal digital assistants and other devices for collecting data must be pretested during study start-up, and directions for their use included in the operations manual.

Database Design

Before the first participant is recruited, the database that will be used to enter, edit, store, monitor, and analyze the data must be created and tested. Depending on the type of database that will be used and the scope of the study, development and testing of the data entry and management system can require weeks to months after staff with the appropriate skills have been identified, hired, and trained. Many academic medical centers provide services to help investigators develop an appropriate database and provide widely used database software programs. For very large studies, professional database design and management services are available, but it's good to get advice on these options from trusted in-house technical experts and senior advisors.

Even for small studies, time spent at the outset creating a database that will house the study data is usually well spent (Chapter 16). Investigators eager to begin a study and start recording data sometimes record data only on paper forms or in a spreadsheet such as Microsoft Excel,

rather than an actual database program. This approach, while easier initially, ends up costing much more in time and effort later, when it is time to analyze the data. The advantage of setting up a database early is that it allows the investigator to consider at the outset what values are acceptable for each variable and disallow or generate alerts for out of range, illogical, and missing values. High quality data entry and management systems improve quality control at the time of data collection or data entry and reduce time that will need to be spent later on data cleaning. But the most important value of a high quality data systemss is to avoid discovering late in a study that there are a large number of missing, out of range or illogical values that cannot be corrected.

Recruitment

Approaches to successfully recruiting the goal number of study participants are described in Chapter 3. We want to emphasize here that timely recruitment is the most difficult aspect of many studies. Adequate time, staff, resources, funding and expertise are essential, and should be planned well in advance of study start-up.

■ FINALIZING THE PROTOCOL

Pretests and Dress Rehearsals

Pretests and pilot studies are designed to evaluate the feasibility, efficiency, and cost of study methods; the reproducibility and accuracy of measurements; likely recruitment rates; and (sometimes) outcome rates and effect sizes. The nature and scale of pretests and pilot studies depend on the study design and the needs of the study. For most studies, a series of pretests or a small pilot study serves very well, but for large, expensive studies a full-scale pilot study may be appropriate. It may be desirable to spend up to 10% of the eventual cost of the study to make sure that recruitment strategies will work, measurements are appropriate, and sample size estimates are realistic.

Pretests are evaluations of specific questionnaires, measures, or procedures that can be carried out by study staff to assess their functionality, appropriateness, and feasibility. For example, pretesting the data entry and database management system is generally done by having study staff complete forms with missing, out-of-range, or illogical data; entering these data; and testing to ensure that the data editing system identifies the errors.

Before the study begins, it is a good idea to test plans for clinic visits and other study procedures in a full-scale **dress rehearsal**. The purpose is to iron out problems with the final set of instruments and procedures. What appears to be a smooth, problem-free protocol on paper usually reveals logistic and substantive problems in practice, and the dress rehearsal will generate improvements in the approach. The PI himself can serve as a **mock subject** to experience the study and the research team from that viewpoint.

Minor Protocol Revisions Once Data Collection Has Begun

No matter how carefully the study is designed and the procedures pretested, problems inevitably appear once the study has begun. The general rule is to make as few changes as possible at this stage. Sometimes, however, protocol modifications can strengthen the study.

The decision as to whether a **minor change** will improve the integrity of the study is often a trade-off between the benefit that results from the improved methodology and the disadvantages of altering the uniformity of the study methods, spending time and money to change the system, and creating confusion for some members of the team. Decisions that simply involve making an **operational definition** more specific are relatively easy. For example, in a study that excludes persons with alcohol abuse, can a person who has been abstinent for several years be included? This decision should be made in consultation with co-investigators, but with

adequate communication through memos and the operations manual to ensure that it is applied uniformly by all staff for the remainder of the study. Often minor adjustments of this sort do not require IRB approval, particularly if they do not involve changing the protocol that has been approved by the IRB, but the PI should ask an IRB staff member if there is any uncertainty. Any change to the protocol, informed consent form, operations manual, or other study documents should be identified by giving the revised document a new version number, and approaches should be in place to make sure the latest version of each document is in use.

Substantive Protocol Revisions Once Data Collection Has Begun

Major changes in the study protocol, such as including different kinds of participants or changing the intervention or outcome, are a serious problem. Although there may be good reasons for making these changes, they must be undertaken with a view to analyzing and reporting the data separately if this will lead to a more appropriate interpretation of the findings. The judgments involved are illustrated by two examples from the Raloxifene Use for The Heart (RUTH) trial, a multicenter clinical trial of the effect of treatment with raloxifene on coronary events in 10,101 women at high risk for coronary heart disease events. The initial definition of the primary outcome was the occurrence of nonfatal myocardial infarction (MI) or coronary death. Early in the trial, it was noted that the rate of this outcome was lower than expected, probably because new clinical co-interventions such as thrombolysis and percutaneous angioplasty lowered the risk for MI. After careful consideration, the RUTH Executive Committee decided to change the primary outcome to include acute coronary syndromes other than MI. This change was made early in the trial; appropriate information had been collected on potential cardiac events to determine if these met the new criteria for acute coronary syndrome, allowing the study database to be searched for acute coronary syndrome events that had occurred before the change was made (1).

Also early in the RUTH trial, emerging results from the Multiple Outcomes of Raloxifene Evaluation (MORE) trial showed that the relative risk of breast cancer was markedly reduced by treatment with raloxifene (2). These results were not conclusive, since the number of breast cancers was small, and there were concerns about generalizability since all women enrolled in MORE had osteoporosis. To determine if raloxifene would also reduce the risk of breast cancer in another population—older women without osteoporosis—the RUTH Executive Committee decided to add breast cancer as a second primary outcome (1).

Each of these changes was major, requiring a protocol amendment, approval of the IRB at each clinical site, approval of the FDA, and revision of a large number of forms and study documents. These are examples of substantive revisions that enhanced feasibility or the information content of the study without compromising its overall integrity. Tinkering with the protocol is not always so successful. Substantive revisions should only be undertaken after weighing the pros and cons with members of the research team and appropriate advisors such as the Data and Safety Monitoring Board, sponsor or funding agency. The investigator must then deal with the potential impact of the change when he analyzes data and draws the study conclusions.

Closeout

At some point in all longitudinal studies and clinical trials, follow-up of participants stops. The period during which participants complete their last visit in the study is often called "**closeout**." Closeout of clinical studies presents several issues that deserve careful planning (3). At a minimum, at the closeout visit staff should thank participants for their time and effort and inform them that their participation was critical to the success of the study. In addition, closeout may include the following activities:

- Participants (and their physicians) should generally be informed of the results of clinically relevant laboratory tests or other measurements that were performed during the study, either in person at the last visit (with a copy in writing) or later by mail.

- In a blinded clinical trial, participants may be told their treatment status, either at the last visit or by mail at the time all participants have completed the trial and the main data analyses are complete or the main manuscript based on study results is published.
- A copy of the main manuscript based on the study results and a press release or other description of the findings written in lay language should generally be mailed to participants (and their physicians) at the time of presentation or publication, with a phone number for participants who have questions.
- After all participants have completed the study, they may be invited to a reception during which the PI thanks them, discusses the results of the study, and answers questions.

QUALITY CONTROL DURING THE STUDY

Good Clinical Practice

A crucial aspect of clinical research is the approach to ensuring that all aspects of the study are of the highest quality. Guidelines for high-quality research, called **Good Clinical Practice (GCP)**, were developed to apply specifically to clinical trials that test drugs requiring approval by the FDA or other regulatory agencies, and are defined as "an international ethical and scientific quality standard for designing, conducting, recording, and reporting trials that involve the participation of human subjects. Compliance with this standard provides public assurance that the rights, safety, and wellbeing of trial subjects are protected" (4).

These principles are increasingly applied to clinical trials sponsored by federal and other public agencies, and to research designs other than trials (Table 17.2). GCP requirements are described in detail in the FDA Code of Federal Regulations Title 21 (4, 5). The International Conference on Harmonization (6) provides quality control guidelines used by regulatory agencies in Europe, the United States, and Japan.

GCP is best implemented by **standard operating procedures (SOPs)** for all study-related activities. The study protocol, operations manual, statistical analysis plan and Data and Safety Monitoring plan can be considered SOPs, but often do not cover areas such as how staff are trained and certified, how the database is developed and tested, or how study files are maintained, kept confidential, and backed up. Many academic medical centers have staff who specialize in processes for meeting GCP guidelines and can provide various templates and models for SOPs. The related topic of ethical conduct of research is addressed in Chapter 14, and in this chapter we focus on quality control of study procedures and data management.

Quality Control for Clinical Procedures

It is a good idea to assign one member of the research team to be the **quality control coordinator** who is responsible for implementing appropriate quality control techniques for all

TABLE 17.2 ASPECTS OF THE CONDUCT OF CLINICAL RESEARCH THAT ARE COVERED BY GOOD CLINICAL PRACTICES

- The design is supported by preclinical, animal, and other data as appropriate.
- The study is conducted according to ethical research principles.
- A written protocol is carefully followed.
- Investigators and those providing clinical care are trained and qualified.
- All clinical and laboratory procedures meet quality standards.
- Data are reliable and accurate.
- Complete and accurate records are maintained.
- Statistical methods are prespecified and carefully followed.
- The results are clearly and fairly reported.

aspects of the study, supervising staff training and certification, and monitoring the use of quality control procedures during the study. The goal is to detect possible problems before they occur and prevent them. The quality control coordinator may also be responsible for preparing for and acting as the contact person for audits by the IRB, FDA, study sponsor, or NIH. Quality control begins during the planning phase and continues throughout the study (Table 17.3).

- **The operations manual.** The operations manual is a very important aspect of quality control (Appendix 17A). To illustrate, consider measuring height in a study where change in height will be used as a predictor of osteoporosis. Since measurement of height is a partially subjective outcome for which there is no feasible gold standard, the operations manual should give specific instructions for the type of measurement device to be used (brand and model of stadiometer), as well as instructions on preparing the participant for the measurement (remove shoes), positioning the patient on the measurement device, and making the measurement.
- **Calibration, training, and certification.** Measurement devices (scales, stadiometers, imaging equipment, laboratory equipment, etc.) should be professionally calibrated before beginning the study and periodically during the study. Standardized training of study staff is essential to high-quality research. All staff involved in the study should receive appropriate training before the study begins, and be certified as to competence with regard to key procedures and measurements. With regard to measurement of height, for example, members of the team can be trained in each aspect of the measurement and required to obtain satisfactory measurements on mock participants whose height is known. The certification procedure should be supplemented during the study by scheduled recertifications and a log of training, certification, and recertification should be maintained at the study site.
- **Performance review.** Supervisors should review the way clinical procedures are carried out by periodically sitting in on representative clinic visits or telephone calls. After obtaining the study participant's permission, the supervisor can be quietly present for at least one complete example of every kind of interview and technical procedure each member of his research team performs. This may seem awkward at first, but it soon becomes comfortable.

TABLE 17.3 QUALITY CONTROL OF CLINICAL PROCEDURES*

Steps that precede the study	Develop a manual of operations
	Define recruitment strategies
	Create operational definitions of measurements
	Create standardized instruments and forms
	Create quality control systems
	Create systems for blinding participants and investigators
	Appoint quality control coordinator
	Train the research team and document this
	Certify the research team and document this
Steps during the study	Provide steady and caring leadership
	Hold regular staff meetings
	Create special procedures for drug interventions
	Recertify the research team
	Undertake periodic performance review
	Compare measurements across technicians and over time

*Clinical procedures include blood pressure measurement, structured interview, chart review, and so on.

It is helpful to use a standardized **checklist** (provided in advance and based on the protocol and operations manual) during these observations. Afterward, communication between the supervisor and the research team member can be facilitated by reviewing the checklist and resolving any quality control issues that were noted in a positive and nonpejorative fashion. The timing and results of performance reviews should be recorded in training logs.

Involving **peers** from the research team as reviewers is useful for building morale and teamwork, as well as for ensuring the consistent application of standardized approaches among members of the team who do the same thing. One advantage of using peers as observers in this system is that all members of the research team acquire a sense of ownership of the quality control process. Another advantage is that the observer often learns as much from observing someone else's performance as the person at the receiving end of the review procedure.

- **Periodic reports.** It is important to **tabulate data** on the technical quality of the clinical procedures and measurements at regular intervals. This can give clues to the presence of missing, inaccurate, or variable measurements. Differences among the members of a blood pressure screening team in the mean levels observed over the past 2 months, for example, can lead to the discovery of differences in their measurement techniques. Similarly, a gradual change over a period of months in the standard deviation of sets of readings can indicate a change in the technique for making the measurement. Periodic reports should also address the success of recruitment, the timeliness of data entry, the proportion of missing and out-of-range variables, the time to address data queries, and the success of follow-up and adherence to the intervention.
- **Special procedures for drug interventions.** Clinical trials that use drugs, particularly those that are blinded, require special attention to the quality control of labeling, drug delivery, and storage; dispensing the medication; and collecting and disposing of unused medication. Providing the correct drug and dosage is ensured by carefully planning with the manufacturer or research pharmacy regarding the nature of the drug distribution approach, by overseeing its implementation, and occasionally by testing the composition of the blinded study medications to make sure they contain the correct constituents. Drug studies also require clear procedures and logs for tracking receipt of study medication, storage, distribution, and return by participants.

Quality Control for Laboratory Procedures

The quality of laboratory procedures can be controlled using many of the approaches described in Table 17.3 for clinical procedures. In addition, the fact that specimens are being removed from the participants (creating the possibility of mislabeling) and the technical nature of laboratory tests lead to several special strategies:

- **Attention to labeling.** When a participant's blood specimen is mistakenly labeled with another individual's name, it may be impossible to correct or even discover the error later. The only solution is prevention, **avoiding mislabeling and transposition errors** by carefully checking the participant's name and number when labeling each specimen. Computer printouts of labels for blood tubes and records speed the process of labeling and avoid the mistakes that can occur when numbers are handwritten. A good procedure when transferring serum from one tube to another is to label the new tube in advance and hold the two tubes next to each other, reading one out loud while checking the other; this can also be automated with scannable **bar codes**.
- **Blinding.** The task of blinding the observer is easy when it comes to measurements on specimens, and it is always a good idea to label specimens so that the technician has no knowledge of the study group or the values of other key variables. Even for apparently objective procedures, like an automated blood glucose determination, this precaution reduces opportunities for bias and provides a stronger methods section when reporting the results. However, blinding laboratory staff means that there must be clear procedures for reporting

abnormal results to a member of the staff who is qualified to review the results and decide if the participant should be notified or other action should be taken. In clinical trials, there must also be strategies in place for (sometimes emergent) unblinding if laboratory measures indicate abnormalities that might be associated with the trial intervention and require immediate action.

- **Blinded duplicates, standard pools and consensus measures.** When specimens or images are sent to a central laboratory for chemical analysis or interpretation, it may be desirable to send blinded duplicates—a second specimen from a random subset of participants given a separate and fictitious ID number—through the same system. This strategy gives a measure of the precision of the laboratory technique. Another approach for serum specimens that can be stored frozen is to prepare a pool of serum at the outset and periodically send aliquots through the system that are blindly labeled with fictitious ID numbers. Measurements carried out on the serum pool at the outset, using the best available technique, establish its values; the pool is then used as a gold standard during the study, providing estimates of accuracy and precision. A third approach, for measurements that have inherent variability such as a Pap test or mammography readings, is to involve two independent, blinded readers. If both agree within predefined limits, the result is established. Discordant results may be resolved by discussion and consensus, or the opinion of a third reader.
- **Commercial laboratory contracts.** In some studies, biologic measures made on blood, sera, cells, or tissue are made under contract to commercial laboratories. The lab must be appropriately licensed and certified and a copy of these certifications should be on file in the study office. Commercial labs should provide data on the reproducibility of their measurements, such as coefficients of variation, guarantee timely service and provide standardized procedures for handling coded specimens, notifying investigators of abnormal results, and transferring data to the main database.

Quality Control for Data Management

The investigator should set up and pretest the data management system before the study begins. This includes designing the forms for recording measurements; choosing computer hardware and software for data entry, editing and management; designing the data editing parameters for missing, out-of-range, and illogical entries; testing the data management system; and planning dummy tabulations to ensure that the appropriate variables are collected (Table 17.4).

TABLE 17.4 QUALITY CONTROL OF DATA MANAGEMENT: STEPS THAT PRECEDE THE STUDY

Be parsimonious: collect only needed variables
Select appropriate computer hardware and software for database management
Program the database to flag missing, out-of-range, and illogical values
Test the database using missing, out-of-range, and illogical values
Plan analyses and test with dummy tabulations
Design paper or electronic forms that are:
Self-explanatory
Coherent (e.g., multiple-choice options are exhaustive and mutually exclusive)
Clearly formatted for data entry with arrows directing skip patterns
Printed in lower case using capitals, underlining, and bold font for emphasis
Esthetic and easy to read
Pretested and validated (see Chapter 15)
Labeled on every page with date, name, ID number, and/or bar code

- **Missing data.** Missing data can be disastrous if they affect a large proportion of the measurements, and even a few missing values can sometimes bias the conclusions. A study of the long-term sequelae of an operation that has a delayed mortality rate of 5%, for example, could seriously underestimate this complication if 10% of the participants were lost to follow-up and if death were a common reason for losing them. Erroneous conclusions due to missing data can sometimes be corrected after the fact—in this case by an intense effort to track down the missing participants—but often the measurement cannot be replaced. There are statistical techniques for **imputing missing values** based on other information from baseline or other follow-up visits or from mean values among other participants. Although these techniques are useful, particularly for multivariate analysis in which the accumulation of missing data across a number of predictor variables could otherwise lead to large proportions of participants unavailable for analysis, they do not guarantee conclusions free of nonresponse bias if there are substantial numbers of missing observations.

  The only good solution is to design and carry out the study in ways that avoid missing data; for example, by having a member of the research team check forms for completeness before the participant leaves the clinic, designing electronic data entry interfaces that do not allow skipped entries, and designing the database so that missing data are immediately flagged for attention by study staff (Table 17.5). Missing clinical measurements should be addressed while the participant is still in the clinic when it is relatively easy to correct errors that are discovered.
- **Inaccurate and imprecise data.** This is an insidious problem that often remains undiscovered, particularly when more than one person is involved in making the measurements. In the worst case, the investigator designs the study and leaves the collection of the data to his research assistants. When he returns to analyze the data, some of the measurements may be seriously biased by the consistent use of an inappropriate technique. This problem is particularly severe when the errors in the data cannot be detected after the fact. The investigator will assume that the variables mean what he intended them to mean, and, ignorant of the problem, may draw conclusions from his study that are wrong.

  Staff training and certification, periodic performance reviews, and regular evaluation of differences in mean or range of data generated by different staff members can help identify or prevent these problems. **Computerized editing** plays an important role, using data entry and management systems programmed to flag or not to allow submission of forms with missing, inconsistent, and out-of-range values. A standardized procedure should be in place for changing original data on any data form. Generally this should be done as soon after data collection as possible, and with a process that includes marking through the original entry (not erasing it), signing and dating the change. Similar processes should be included in

TABLE 17.5 QUALITY CONTROL OF DATA MANAGEMENT: STEPS DURING THE STUDY

Flag or check for omissions and major errors while participant is still in the clinic
No errors or transpositions in ID number, name code, or date on each page.
All the correct forms for the specified visit have been filled out.
No missing entries or faulty skip patterns.
Entries are legible.
Values of key variables are within permissible range.
Values of key variables are consistent with each other (e.g., age and birth date).
Carry out periodic frequency distributions and variance measures to discover aberrant values
Create other periodic tabulations to discover errors (see Appendix 17B)

electronic data entry and editing systems. This provides an electronic "**audit trail**" to justify changes in data and prevent fraud.

Periodic tabulation and inspection of frequency distributions of important study variables at regular intervals allows the investigator to assess the completeness and quality of the data at a time when correction of past errors may still be possible (e.g., by contacting the participant by email or phone, or requesting that the participant return to the study offices), and when further errors in the remainder of the study can be prevented. A useful list of topics for quality control reports is provided in Appendix 17B.

- **Fraudulent data.** Clinical investigators who lead research teams have to keep in mind the possibility of an unscrupulous colleague or employee who chooses fabrication of study information as the easiest way to get the job done. Approaches to guarding against such a disastrous event include taking great care in choosing colleagues and staff, developing a strong relationship with them so that ethical behavior is explicitly understood and rigorously followed by all, being alert to the possibility of fraud when data are examined, and making unscheduled checks of the primary source of the data to be sure that they are real.

Collaborative Multicenter Studies

Many research questions require larger numbers of participants than are available in a single center, and these are often addressed in collaborative studies carried out by research teams that work in several locations. Sometimes these are all in the same city or state, and a single investigator can oversee all the research teams. Often, however, collaborative studies are carried out by investigators in cities thousands of miles apart with separate funding, administrative, and regulatory structures.

Multicenter studies of this sort require special steps to ensure that all centers are using the same study procedures and producing comparable data that can be combined in the analysis of the results. A **coordinating center** establishes a communication network; coordinates the development of the operations manual, forms, and other standardized quality control aspects of the trial; trains staff at each center who will make the measurements; and oversees data management, analysis, and publication. Collaborative studies generally use distributed electronic data entry systems connected through the Internet.

There is also a need for establishing a governance system with a **steering committee** made up of the PIs and representatives of the funding institution, and with various subcommittees. One **subcommittee** needs to be responsible for **quality control** issues, developing the standardization procedures and the systems for training, certification, and performance review of study staff. These tend to be complicated and expensive, requiring **centralized training** for relevant staff from each center, **site visits** for performance review, and data audits by coordinating center staff and peers (Appendix 17B). Other subcommittees generally include groups that oversee **recruitment** and **clinical activities**, a group that reviews and approves **publications and presentations**, and one that considers proposed ancillary studies.

In a multicenter study, changes in operational definitions and other study methods often result from questions raised by a clinical center that are answered by the relevant study staff or committee and posted on the study website in a running list to make sure that everyone involved in the study is aware of the changes. If a significant number of changes accumulate, dated revised pages in the operations manual and other study documents should be prepared that include these changes. Small single-site studies can follow a simpler pattern, making notes about changes that are dated and retained in the operations manual.

A Final Thought

A common error in research is the tendency to collect **too much data**. The fact that the baseline period is the only chance to measure baseline variables leads to a desire to include everything that might conceivably be of interest, and there is a tendency to have more follow-up visits and

collect more data at them than is useful. Investigators tend to collect far more data than they will ever analyze or publish.

One problem with this approach is the time and costs required by measuring less important things; participants become tired and annoyed, and the quality of more important measurements deteriorates. Another problem is the added size and complexity of the database, which makes quality control and data analysis more difficult. It is wise to question the need for every variable that will be collected and to eliminate many that are optional. Including a few intentional redundancies can improve the validity of important variables, but parsimony is the general rule.

SUMMARY

1. Successful study implementation begins with **assembling resources** including **space**, **staff**, and **funding** for the study and its **start-up**, all of which require strong **leadership** by the PI.
2. **Study start up** requires managing the **budget**, obtaining **IRB approval** and finalizing the **protocol** and **operations manual** through a process of **pretesting** the appropriateness and feasibility of plans for **recruitment**, **interventions**, **predictor** and **outcome variable measurements**, **forms**, and the **database**; the goal is to minimize the need for subsequent protocol revisions once data collection has begun.
3. **Minor protocol revisions** after the study has begun, such as adding an item to a questionnaire or modifying an operational definition, are relatively **easily accomplished**, though IRB approval may sometimes be required and data analysis may be affected.
4. **Major protocol revisions** after the study has begun, such as a change in the nature of the intervention, inclusion criteria, or primary outcome, have **major implications** and should be undertaken reluctantly and with the approval of key bodies such as the DSMB, IRB, and funding institution.
5. There is a need for **closeout** precedures to properly inform participants of study findings and to manage transition of and implications for their care.
6. **Quality control** during the study should be assured with a systematic approach under the supervision of a **quality control coordinator**, following the principles of **Good Clinical Practice (GCP)**, and including:
   a. **Standard operating procedures (SOPs)** with an **operations manual**; **staff training**, **certification**, and **performance review**; **periodic reports** (on recruitment, visit adherence, and measurements); and regular **team meetings**.
   b. Quality control for **laboratory procedures—blinding** and systematically **labeling** specimens taken from study participants, and using **standard pools**, **blinded duplicates and consensus measures**.
   c. Quality control of **data management**—designing forms and electronic systems to enable oversight of the **completeness**, **accuracy**, and **integrity** of collecting, entering, editing, and analyzing the data.
7. **Collaborative multicenter studies** create **subcommittees** and other distributed systems for managing the study and quality control.

APPENDIX 17A
Example of an Operations Manual Table of Contents[1]

[1]N.B. This is a model for a large multicenter trial. The manual of operations for a small study can be less elaborate.

Chapter 11. Data and Safety Monitoring Guidelines
Appendices
- Letters to participants, primary providers, and so on
- Questionnaires, forms
- Details on procedures, criteria, and so on
- Recruitment materials (advertisements, fliers, letters, etc.)

APPENDIX 17B
Quality Control Tables and Checklists

I. Tabulations for monitoring performance characteristics[2]
  A. Clinic characteristics
    1. Recruitment
      a. Number of participants screened for enrollment; number excluded and tabulation of reasons for exclusion
      b. Cumulative graph of number recruited compared with that required to achieve recruitment goal
    2. Follow-up
      a. Number of completed follow-up examinations for each expected visit; number seen within specified time frame
      b. Measures of adherence to the study intervention, visits and measures
      c. Number of dropouts and participants who cannot be located for follow-up
    3. Data quantity and quality
      a. Number of forms completed, number that generated edit messages, number of unanswered edit queries, time to resolution of queries
      b. Number of forms missing, number or proportion of missing variables
    4. Protocol adherence
      a. Number of ineligible participants enrolled
      b. Summary of data on pill counts and other adherence measures by treatment group
  B. Data center characteristics
    1. Number of forms received and number awaiting data entry
    2. Cumulative list of coding and protocol changes
    3. Timetable indicating completed and unfinished tasks
  C. Central laboratory characteristics
    1. Number of samples received and number analyzed
    2. Number of samples inadequately identified, lost, or destroyed
    3. Number of samples requiring reanalysis and tabulation of reasons
    4. Mean and variance of blind duplicate differences, and secular trend analyses based on repeat determinations of known standards
  D. Reading center characteristics
    1. Number of records received and read
    2. Number of records received that were improperly labeled or had other deficiencies (tabulate deficiencies)
    3. Analyses of repeat readings as a check on reproducibility of readings and as a means of monitoring for time shifts in the reading process
II. Site visit components
  A. Site visit to clinical center
    1. Private meeting of the site visitors with the PI
    2. Meeting of the site visitors with members of the clinic staff
    3. Inspection of examining and record storage facilities

[2]Tables should contain results for the entire study period, and, when appropriate, for the time period covered since production of the last report. Rates and comparisons among staff and participating units should be provided when appropriate.

4. Comparison of data contained on randomly selected data forms with those contained in the computer data file
5. Review of file of data forms and related records to assess completeness and security against loss or misuse
6. Observation of clinic personnel carrying out specified procedures
7. Check of operations manuals, forms, and other documents on file at the clinic to assess whether they are up-to-date
8. Observation or verbal walk through of certain procedures (e.g., the series of examinations needed to determine participant eligibility)
9. Conversations with actual study participants during or after enrollment as a check on the informed consent process
10. Private conversations with key support personnel to assess their practices and philosophy with regard to data collection
11. Private meeting with the PI concerning identified problems

B. Site visit to data center
1. Review of methods for inventorying data received from clinics
2. Review of methods for data management and verification
3. Assessment of the adequacy of methods for filing and storing paper records received from clinics, including the security of the storage area and methods for protecting records against loss or unauthorized use
4. Review of available computing resources
5. Review of method of randomization and of safeguards to protect against breakdowns in the randomization process
6. Review of data editing procedures and audit trails
7. Review of computer data file structure and methods for maintaining the analysis database
8. Review of programming methods both for data management and analysis, including an assessment of program documentation
9. Comparison of information contained on original study forms with that in the computer data file
10. Review of methods for generating analysis data files and related data reports
11. Review of methods for backing up the main data file
12. Review of master file of key study documents, such as handbooks, manuals, data forms, minutes of study committees, and so on, for completeness

REFERENCES

1. Mosca L, Barrett-Connor E, Wenger NK, et al. Design and methods of the Raloxifene Use for The Heart (RUTH) Study. *Am J Cardiol* 2001;88:392–395.
2. MORE Investigators. The effect of raloxifene on risk of breast cancer in postmenopausal women: results from the MORE randomized trial. Multiple outcomes of raloxifene evaluation. *JAMA* 1999;281:2189–2197.
3. Shepherd R, Macer JL, Grady D. Planning for closeout–From day one. *Contemp Clin Trials* 2008;29:136–139
4. http://www.fda.gov/downloads/Drugs/Guidances/ucm073122.pdf
5. FDA Regulations Relating to Good Clinical Practice and Clinical Trials. Available at: www.fda.gov/ScienceResearch/SpecialTopics/RunningClinicalTrials/ucm114928.htm
6. Information about Good Clinical Practices in the European Medicines Agency International Conference on Harmonization. Available at: http://www.ich.org or at http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000035.jsp&murl=menus/regulations/regulations.jsp&mid=WC0b01ac0580027645&jsenabled=true

CHAPTER 18

Community and International Studies

Norman Hearst and Thomas Novotny

Most clinical research takes place in university medical centers or other large medical institutions. Such sites offer many advantages for conducting research, including the obvious one of having experienced researchers. An established culture, reputation, and infrastructure for research all facilitate the work of everyone from novice investigator to tenured professor. Success breeds more success, thus concentrating clinical research in centers of excellence. This chapter, in contrast, deals with research that takes place outside of such centers.

We define **community research** as research that takes place outside the usual medical center setting and that is designed to meet the needs of the communities where it is conducted. **International research**, particularly in poor countries, can involve many of the same challenges of responding to local needs and establishing a research infrastructure where none existed before. Further, such research requires understanding of numerous political, bureaucratic, and cultural complexities that arise in international research. Community and international research both often involve **collaboration** between local investigators and colleagues from an established research center. Such collaborations are critical in solving longstanding or emerging global or local health problems, and they can be extraordinary opportunities for personal growth and mutual learning. However, these collaborations can be **challenging** because of **physical distances** separating the investigators, **cultural differences** involving the participants, **political issues** involving local and national institutions, and **funding constraints** at both the donor and recipient levels.

WHY COMMUNITY AND INTERNATIONAL RESEARCH?

Collaborative research is often the only way to address research questions that have to do with special settings, new and re-emerging diseases, or specific populations. Research in academic medical centers tends to focus on clinical or basic science priorities that may be quite different from local community needs, and,even more different from global health problems that affect large segments of the world's population. These global problems require collective efforts for humanitarian reasons and because national, state, or local borders do not insulate communities from the effects of such problems. The "**10/90 gap**" in health research, in which 90% of the global burden of disease receives only 10% of global research investment (1), is ample justification for more collaborative research that addresses the enormous health problems of low- and middle-income countries (**LMICs**). As such, there exists a need to bolster the research **capacity** in LMICs and communities through international and national collaboration. This includes careful attention to developing institutional review processes and human subjects' protections. Furthermore, participation in the research process has benefits for a community and for researchers that go beyond the value of the information collected in a particular study. Lasting relationships, a sense of pride, and perhaps even economic development may result from community research that is done with care and concern for the public good.

Local Questions

Many research questions require answers available only through **local research**. National or state level data from central sources may not accurately reflect local disease burdens or the distribution of risk factors in the local community. Interventions, especially those designed to change behavior, may not have the same effect in different settings. For example, the public health effectiveness of condom promotion as an HIV/AIDS prevention strategy is quite different in the United States than it is in Africa (2). Finding approaches that fit local needs requires local research methods (Table 18.1).

Biologic data on the pathophysiology of disease and the effectiveness of treatments are usually generalizable to a wide variety of populations and cultures. However, there may be racial, cultural, or genetic differences or differences based on regional disease etiology that require local research. For example, the efficacy of antihypertensive drugs may be different in patients of African and European descent (3); the causative agents and patterns of antimicrobial sensitivity for pneumonia are different in Bolivia and Boston; and the perception of health, health care, and illness can differ significantly across different communities (4).

Greater Generalizability

Community research is sometimes useful for producing results that are more **generalizable**. For example, patients with back pain who are seen at referral hospitals are very different from patients who present with back pain to primary care providers. Studies of the natural history of back pain or response to treatment at a tertiary care center therefore may be of limited use for clinical practice in the community.

Partly in response to this problem, several **practice-based research networks** have been organized in which physicians from community settings work together to study research questions of mutual interest (5). An example is the response to treatment of patients with carpal tunnel syndrome in primary care practices (6). Most patients improved with conservative therapy; few required referral to specialists or sophisticated diagnostic tests. Previous studies had recommended early surgical intervention for carpal tunnel syndrome based on studies of patients treated at a major referral center.

Issues of generalizability are also important in **international research**. Research findings from one country will not always apply in another. But while results generalize best to where the research was done, they may also be relevant for migrant populations that originated in the country of the research. Such displaced populations are of ever increasing importance in a globalized world that had 214 million international migrants as of 2010 (7). Globalization now necessitates a broader perspective on disease risk and also on collaborative research approaches to address diseases that cross national borders so easily.

Building Local Capacity

Clinical research should not be the exclusive property of academic medical centers. The priorities of researchers in these sites are bound to reflect the priorities of funders, the issues they

TABLE 18.1 EXAMPLES OF RESEARCH QUESTIONS REQUIRING LOCAL RESEARCH

What are the rates of child car seat and seat belt use in a low-income neighborhood of Chicago?
What are the patterns of antimicrobial resistance of tuberculosis isolates in Uganda?
What is the impact of a worksite-based STI prevention campaign for migrant farm workers in Texas?
What proportion of coronary heart disease among women in Brazil is associated with cigarette smoking?

encounter in their daily practice, and what they believe to be of general scientific or economic importance. Conducting research in community and international settings ensures that questions of local importance must be prioritized (8).

The value of **community participation** in research goes beyond the specific information collected in each study. Conducting research has a positive ripple effect by raising local scholarly standards and by encouraging creativity and independent thinking. Each project builds skills and confidence that allow local researchers to see themselves as participants in the scientific process, not just as consumers of knowledge produced elsewhere. This in turn encourages more research. Furthermore, participating in research can bring intellectual and financial resources to a community and help to encourage local empowerment and self-sufficiency.

COMMUNITY RESEARCH

In theory, starting up community research is the same process as for any other research endeavor. The general approach outlined in this book applies just as well in a small town in rural America or Nepal as it does in San Francisco or London. In practice, the greatest challenge is finding experienced colleagues or mentors with whom to interact and learn. Such help may not be available locally. This often leads to an important early decision for would-be community or international investigators: to work alone or in collaboration with more established investigators based elsewhere.

Starting on Your Own

Getting started in research without the help of a more experienced colleague is like teaching oneself how to swim: It is not impossible, but it is difficult and sometimes fraught with unforeseen dangers. Often, however, it is the only option. Following a few rules may make the process easier.

- **Start simple.** It is seldom a good idea to begin research in a community with a randomized controlled trial. Small pilot descriptive studies producing useful local data may make more sense—it is better to achieve a small success than a large failure. More ambitious projects can be saved for later and can draw on the pilot data you generated previously. For example, a descriptive study of condom use among young men in Uganda conducted by a novice local researcher served as a first step toward a larger intervention trial on HIV/AIDS prevention in that community (9).
- **Think about the local comparative advantages.** What questions can an investigator answer in his local setting better than anywhere else? This usually means leaving the development of new laboratory techniques and treatments to academic medical centers and pharmaceutical research organizations. It is often best for a young investigator to focus on health problems or populations that are unusual elsewhere, but common in the local community.
- **Network.** As discussed in Chapter 2, networking is important for any investigator. A new investigator should make whatever contact he can with scientists elsewhere who are addressing similar research questions. If formal collaborators are not available, it may at least be possible to find someone to give feedback on a draft of a research protocol, a questionnaire, or a manuscript through e-mail and telephone. Attending a scientific conference in one's field of interest is a good way to make such contacts and referring to a senior colleague's work can be a good way to initiate such a contact.

Collaborative Research

Because it is difficult to get started on one's own, a good way to begin research in a community is often in collaboration with more experienced researchers based elsewhere, especially if those

investigators have established trust, contacts, and methodologies in the target country. There are two main models for such collaboration: top-down and bottom-up (10).

The **top-down** model refers to studies that originate in an academic center and involve community investigators in the recruitment of patients and the conduct of the study. This occurs, for example, in large multicenter trials that invite hospitals and clinics to enroll patients into an established research protocol. This approach has the advantages that come with built-in senior collaborators who are usually responsible for designing the study and obtaining the necessary resources and clearances to carry it out.

In the **bottom-up** model, established investigators provide guidance and technical assistance to local investigators and communities developing their own research agendas. Some academic medical centers offer training programs for community investigators or international researchers. If one can gain access to such a program or establish an equivalent relationship, this can be ideal for building local research capacity, especially when such a partnership is sustained on a long-term basis. However, establishing an institutional relationship of this type is not easy. Most funding agencies are more interested in sponsoring specific research projects than in devoting resources to building local research capacity and collaborations. Even when funding to cover training and travel expenses is available, experienced investigators may prefer to spend their time conducting their own research rather than helping others get started. Still, the value of collaborative community-based participatory research (**CBPR**), in which the community participates fully in all aspects of the research, cannot be overemphasized in terms of satisfaction, importance, and relevance to the local community (11).

Community researchers need to take advantage of the potential incentives they can offer to more established investigators with whom they would like to work. In the top-down model, the most important thing they can offer is access to subjects. In the bottom-up model, the incentives can include the intrinsic scientific merit of a study in the community, co-authorship of resulting publications, and the satisfaction of building a collaborative relationship and helping a community develop research capacity.

To start a new research program, the ideal option may be to form a long-term partnership with an established research institution. **Memoranda of Understanding (MOUs)** can be signed by collaborating agencies so that written evidence of communication and agreements can be provided to potential funders. Having this collaboration established in advance can save time and frustration. Collaboration under such a structure can include a combination of top-down and bottom-up projects. It must be remembered, however, that good research collaboration is fundamentally between individual investigators. An academic institution may provide the climate, structure, and resources that support individual collaboration, but the individuals themselves must provide the cultural sensitivity, mutual respect, hard work, and long-term commitment to make it work.

INTERNATIONAL RESEARCH

International research often involves collaboration between groups with different levels of experience and resources and thus is subject to many of the same issues as community research. However, international research brings **additional challenges**. The issues described in the following section are especially important.

Barriers of Distance, Language, and Culture

Without a thorough understanding of a community's cultural perspectives, many researchers find that even the best laid plans fail despite careful planning and advanced technologies. To avoid failure, researchers must understand the cultural perceptions of disease in the communities where they intend to work and develop culturally sound approaches to their collaborative research. Because of the **distances** involved, opportunities for face-to-face **communication**

between international colleagues are limited. If at all possible, colleagues on both sides should make at least one **site visit** to each other's institutions. International conferences may sometimes provide additional opportunities to meet, but such opportunities are likely to be rare. Fortunately, **e-mail**, the Internet, and **Skype** have made international communication easier, faster, and less expensive. Good communication is possible at any distance, but it requires effort on both sides. The most modern methods of communication are of no help if they are not used regularly. Lack of frequent communication and prompt response to queries made on either side is a sign that a long-distance collaboration may be in trouble.

Language differences are often superimposed on the communication barriers caused by distance. If the first language spoken by investigators at all sites is not the same, it is important that there be a language that everyone can use (usually that language is English). Expecting that all interactions are to be in English, however, places investigators in many countries at a disadvantage. Foreign investigators who do not speak the local language are unlikely to have more than a superficial understanding of the country's culture and cannot participate fully in many key aspects of a study, including questionnaire development and validation. They will not be able to conduct conversations with study subjects and research assistants. This communication is especially important in studies with behavioral components.

Even when linguistic barriers are overcome, **cultural** differences can cause serious misunderstandings between investigators and their subjects or among investigators. Literal, word-for-word translations of questionnaires may have different meanings, be culturally inappropriate, or omit key local factors. Institutional norms may be different. For example, in some settings, a foreign collaborator's department chief who had little direct involvement in a study might expect to be first author of the resulting publication. Such issues should be anticipated and clearly laid out in advance, as part of the process of institutional development for the project. Patience, goodwill, and flexibility on all sides can usually surmount problems of this type. For larger projects, it may be advisable to include an anthropologist, ethicist, or other expert on cultural issues as part of the research team.

Frequent, clear, and open communications and prompt clarification of any questions or confusion are essential. When dealing with cultural and language differences, it is better to be repetitive and risk stating the obvious than to make incorrect assumptions about what the other person thinks or is saying. Written affiliation agreements that spell out mutual responsibilities and obligations may help clarify issues such as data ownership, authorship order, publication rights, and decisions regarding the framing of research results. Development of such agreements requires the personal and careful attention of collaborators from both sides.

Issues of Funding

Because of economic inequities, collaborations between institutions in rich and poor countries are generally only possible with **funding** originating from the rich country or, less often, from other rich countries or international organizations. An increasing number of large donor organizations are active in global health research, but their support is often limited to a very specific research agenda with strict requirements for measurable results. Much bilateral donor funding tends to flow through the institution in the rich country, reinforcing the subordinate position of institutions in LMICs. As in any situation with an **unequal balance of power**, this creates ethical challenges. When investigators from rich countries control the purse strings, it is not uncommon for them to treat their counterparts in poor countries more like employees than colleagues. International donors and funding agencies need to be especially careful to discourage this and instead to promote true joint governance of collaborative activities (8).

Different practices of **financial management** are another potential area for conflict among research consortium members. Institutions in rich countries may attempt to impose accounting standards that are difficult or impossible to meet locally. Institutions in LMICs may load budgets with computers and other equipment that they expect to keep after the study is over. While

this is understandable given their needs and lack of alternative funding sources, it is important that any subsidies beyond the actual cost of conducting the research be clearly negotiated and that accounting practices be put in place to meet the needs of the funding agencies. Conversely, high institutional overheads and investigator salaries often create an inequitable situation with the majority of funding for collaborative research staying in the donor country even when most of the work is done in the partner country.

Donor country institutions and donors should pay particular attention to building the research **administration capacity** of local partners. This could mean providing administrative and budgetary training or using consultants in the field to help with local administrative tasks. A requirement for international partners is to obtain a D-U-N-S Number, a unique nine digit identification number for each physical location of institutions applying for contracts or grants from the US Federal Government (http://fedgov.dnb.com/webform). Efforts invested in developing administrative capacity may pay off in improved responsiveness to deadlines, more efficient reporting, avoiding unnecessary conflict, and building a solid infrastructure for future research.

Ethical Issues

International research raises ethical issues that must be faced squarely. All the general ethical issues for research apply (Chapter 14). Because international research may present particular risks for violations of human subject protections, it requires additional considerations and safeguards.

What, for example, is the **appropriate comparison group** when testing new treatments in an LMIC where conventional treatment is unavailable? Placebo controls are unethical when other effective treatments are the standard of care elsewhere. But what is the "standard of care" in a community where most people are too poor to afford proven treatments that may be available in many countries? On the one hand, it may not be possible for investigators to provide state-of-the-art treatment to every participant in a study. On the other hand, allowing placebo controls simply because of inadequate access to drugs and medical care is unethical and has been challenged by many intergovernmental groups and patient advocacy organizations. For example, studies of less expensive oral antiretroviral treatments to prevent mother-to-child transmission of HIV done in countries where most women did not have access to a proven existing treatment regimen demonstrate some of these issues (12, 13).

A related issue has to do with **testing treatments that are unlikely to be economically accessible to the population of the host country**. Are such studies ethical, even if they follow all the usual rules? For example, would it be ethical to study a new drug for Type II diabetes in a LMIC where this drug would probably be unaffordable? These questions do not have simple answers. Established international conventions governing ethical research, such as the Declaration of Helsinki, have been challenged and are subject to multiple interpretations (14, 15).

A key test may be to consider why the study is being conducted in an LMIC in the first place. If the true goal is to gather information to help the people of that country, this should weigh in the study's favor and it should be designed accordingly. Ideally, the goal of research should be sustainable change and added value for the host country (16). If, on the other hand, the goal is expediency or to avoid obstacles to doing a study in a rich country, the study should be subject to all ethical requirements that would apply in the sponsoring country, including the important requirement of distributive justice (see Chapter 14).

For this and other reasons, studies in poor countries that are directed or funded from elsewhere should be approved by **ethical review boards in both countries**. But while such approval is necessary, it does not guarantee that a study is ethical. Systems for ethical review of research in many poor countries are weak or nonexistent and can sometimes be manipulated by local investigators or politicians. Conversely, review boards in rich countries are sometimes ignorant of or insensitive to the special issues involved in international research. Official approval does not remove the final responsibility for the ethical conduct of research from the investigators themselves.

Another important ethical concern is the **treatment of collaborators** from partner LMICs. Several issues must be agreed upon in advance. Who owns the data that will be generated? Who needs whose permission to conduct and publish analyses? Will local investigators get the support they need to prepare manuscripts for international publication without having to pay for this by giving up first authorship? How long a commitment is being made on both sides? A large trial in several poor countries of voluntary counseling and testing to prevent HIV infection abruptly dropped its collaborating site in Indonesia (17). According to the investigators, this was because the outcome variable of interest turned out to be less common at that site than projected in the study's power calculations. Even though this decision made practical sense, it was perceived by the Indonesians as a breach of faith.

Other ethical issues may have to do with **local economic and political realities**. For example, a planned clinical trial of pre-exposure HIV prophylaxis with tenofovir for commercial sex workers was cancelled even though it had been cleared by multi-national ethical review boards (18). The intended study subjects were concerned that they might end up with no source of medical care for problems related to HIV infection or drug effects and were not willing to participate without guarantees of lifetime health insurance. The prime minister of the country intervened to stop the trial.

Finally, an explicit goal of all international collaboration should be to **increase local research capacity**. What skills and equipment will the project leave behind when completed? What training activities will take place for project staff? Will local researchers participate in international conferences? Will this be only for high-level local investigators who already have many such opportunities, or will junior colleagues have a chance as well? Will the local researchers be true collaborators and principal authors of publications, or are they simply being hired to collect data? Scientists in poor countries should ask and expect clear answers to these questions. As summarized in Table 18.2, **good communication and long-term commitment** are recurring themes in successful international collaborative research.

The World Health Organization recently published a set of case studies dealing with ethical issues in global health research (19) to help investigators, ethics review committee members, health authorities, and others to play their respective roles in the ethical conduct of research.

TABLE 18.2 STRATEGIES TO IMPROVE INTERNATIONAL COLLABORATIVE RESEARCH

Scientists in low- and middle-income countries (LMICs)
Choose collaborators carefully
Learn English (or other language of collaborators)
Become familiar with the international scientific literature in the area of study
Be sure that collaboration will build local research capacity
Clarify administrative and scientific expectations in advance
Scientists in upper-income countries
Choose collaborators carefully
Learn the local language and culture
Be sensitive to local ethical issues
Encourage local collaboration in all aspects of the research process
Clarify administrative and scientific expectations in advance
Funding agencies
Set funding priorities based on public health need
Encourage true collaboration rather than a purely "top-down" model
Recognize the importance of building local research capacity
Make subsidies for local equipment and infrastructure explicit
Be sure that overhead and high salaries in the upper-income country do not take too much of the budget

Much can be learned from the mistakes and successes of others, but with goodwill on the part of funders, donor-country partners, and officials on both sides of the research partnerships, ethical principles can be assured in international research and capacity for such research strengthened globally.

Risks and Frustrations

Researchers from rich countries who contemplate becoming involved in international research need to start with a realistic appreciation of the difficulties and risks involved. Launching such work is usually a long, slow process. **Bureaucratic obstacles** are common on both ends. In countries that lack infrastructure and political stability, years of work can be vulnerable to major disruption from natural or manmade **catastrophes**. In extreme cases, these situations can threaten the safety of project staff or investigators. For example, important collaborative HIV/AIDS research programs that had been built over many years were completely destroyed by civil wars in Rwanda and the Congo. Less catastrophic and more common challenges are the daily hardships and **health risks** that expatriate researchers may face, ranging from unsafe water and malaria to smog, common crime, and traffic accidents.

Another frustration for researchers in LMICs is the difficulty in **applying their findings**. Even when new strategies for preventing or treating disease can be successfully developed and proven to be effective, lack of political will and resources often thwarts their widespread application in host countries. Researchers need to be realistic in their expectations, gear their work toward investigating strategies that would be feasible to implement if found effective, and be prepared to act as advocates for improving the health of the populations they study.

The Rewards

Despite the difficulties, the need for more health research in many parts of the world is overwhelming. By participating in international research, an investigator in a donor country can sometimes have a far greater and more immediate **impact on public health** than would be possible by staying within the walls of academia. This impact comes not only from the research itself but also from what is sometimes called **global health diplomacy**. In fact, health is now seen as a major driving force for foreign policy priorities (20). Health diplomacy may be practiced through collaborative research on global health challenges such as HIV/AIDS, malaria, TB, maternal and child health, and health systems strengthening. Health and politics have always been intertwined, but in a globalized world, there is a growing need for collaborative actions to address major trans-border health issues; international research is part of this global effort. The chance to have meaningful involvement and make a real contribution to global health is a privilege that can **enrich careers** and our personal lives. All stand to gain through increased collaboration and expanded research opportunities.

■ SUMMARY

1. **Community and international research** is necessary to discover **regional differences** in such things as the **epidemiology** of a disease, and the **cultural** and other local **factors** that determine which **interventions** will be **effective**.
2. **Local participation** in clinical research can have secondary benefits to the region such as enhanced levels of **scholarship** and **self-sufficiency**.
3. Although the theoretical and ethical issues involved in community and international research are broadly applicable, practical issues such as acquiring funding and mentoring are **more difficult** in a community or international setting; tips for success include **starting small**, thinking of **local advantages**, and **networking**.
4. **Collaboration** between academic medical centers and community researchers can follow a **top-down model** (community investigators conduct studies that originate from the

academic center) or a **bottom-up model** (investigators from the academic center help community investigators conduct research that they themselves originate).

5. **International research** involves many of the same issues as community research with additional **challenges**, particularly in low- and middle-income countries (**LMICs**), that are related to **communication** and **language**, **cultural differences**, **funding**, **unequal power structures**, and **financial and administrative practices**.
6. International research has its own set of **ethical issues**, including testing **treatments that may be unaffordable** in the LMICs, use of **placebos** in **vulnerable populations**, and the status and treatment of **collaborators**.
7. Overcoming the challenges in international research can bring the rewards of **helping people in need**, **of being part of a larger global health community**, and **of enriching one's cultural experiences**.

REFERENCES

1. Unite for Sight. The importance of global health research: closing the 10/90 gap. Available at: http://www.uniteforsight.org/global-impact-lab/global-health-research#_ftnref12, accessed 9/23/12.
2. Hearst N, Chen S. Condom promotion for AIDS prevention in the developing world: is it working? *Studies in Family Planning* 2004;35(1):39–47.
3. Drugs for hypertension. *Med Lett Drugs Ther* 1999;41:23–28.
4. Griffith BN, Lovett GD, Pyle DN, et al. Self-rated health in rural Appalachia: health perceptions are incongruent with health status and health behaviors. *BMC Public Health* 2011;11:229. doi:10.1186/1471-2458-11-229.
5. Nutting PA, Beasley JW, Werner JJ. Practice-based research networks answer primary care questions. *JAMA* 1999;281:686–688.
6. Miller RS, Ivenson DC, Fried RA, et al. Carpal tunnel syndrome in primary care: a report from ASPN. *J Fam Pract* 1994;38:337–344.
7. United Nations Department of Economic and Social Affairs (UN DESA). Trends in international migrant stock: the 2008 revision. Available at: *http://esa.un.org/migration/index.asp?panel=1*, accessed 1/12/2013.
8. Lee K, Mills A. Strengthening governance for global health research: the countries that most need health research should decide what should be funded. *BMJ* 2009;2000:775–776.
9. Kajubi P, Kamya MR, Kamya S, et al. Increasing condom use without reducing HIV risk: results of a controlled community trial in Uganda. *Journal of AIDS* 2005;40(1):77–82.
10. Hearst N, Mandel J. A research agenda for AIDS prevention in the developing world. *AIDS* 1997;11(Suppl 1):S1–4.
11. Minkler M and Wallerstein N, eds. (2008). *Community-Based Participatory Research for Health: From Process to Outcomes*. ISBN 978-0-470-26043-2. Jossey-Bass
12. Lurie P, Wolfe SM. Unethical trials of interventions to reduce perinatal transmission of the human immunodeficiency virus in developing countries. *N Engl J Med* 1997;337:853–856.
13. Perinatal HIV Intervention Research in Developing Countries Workshop Participants. Science, ethics, and the future of research into maternal-infant transmission of HIV-1. *Lancet* 1999;353:832–835.
14. Brennan TA. Proposed revisions to the Declaration of Helsinki: will they weaken the ethical principles underlying human research? *N Engl J Med* 1999;341:527–531.
15. Levine RJ. The need to revise the Declaration of Helsinki. *N Engl J Med* 1999;341:531–534.
16. Taylor D, Taylor CE. *Just and lasting change: when communities own their futures*. Baltimore: JHU Press, 2002.
17. Kamenga MC, Sweat MD, De Zoysa I, et al. The voluntary HIV-1 counseling and testing efficacy study: design and methods. *AIDS and Behavior* 2000;4:5–14.
18. Page-Shafer K, Saphonn V, Sun LP, et al. HIV prevention research in a resource-limited setting: the experience of planning a trial in Cambodia. *Lancet* 2005;366(9495):1499–1503.
19. Cash R, Wikler D, Saxena A, et al. *Casebook on ethical issues in international health research*. Geneva: World Health Organization, 2009.
20. Katz R, Kornblet S, Arnold G, et al. Defining health diplomacy: changing demands in the era of globalization. *The Milbnk Quarterly* 2011;89(3):503–523.

CHAPTER 19

Writing a Proposal for Funding Research

Steven R. Cummings, Deborah G. Grady, and Stephen B. Hulley

The **protocol** is the detailed written plan of the study. Writing the protocol forces the investigator to organize, clarify, and refine all the elements of the study, and this enhances the scientific rigor and the efficiency of the project. Even if the investigator does not require funding for a study, a protocol is necessary for guiding the work and for obtaining ethical approval from the institutional review board (IRB). A **proposal** is a document written for the purpose of obtaining funds from granting agencies. It includes descriptions of the study's aims, significance, research approach, human subjects concerns, and the budget and other administrative and supporting information that is required by the specific agency.

This chapter will describe **how to write** a proposal that will **get funded**. It focuses on original research proposals using the format suggested by the National Institutes of Health (NIH), but proposals to most other funding agencies (such as the Department of Veterans Affairs, Centers for Disease Control, Agency for Healthcare Research and Quality, and private foundations) generally require a similar format. Excellent advice on writing an application, preparing budgets, and submitting proposals is available on the NIH website (http://grants.nih.gov/grants/writing_application.htm).

WRITING PROPOSALS

The task of preparing a proposal generally requires several months of organizing, writing, and revising. The following steps can help the project get off to a good start.

- **Decide where the proposal will be submitted.** Every funding agency has its own unique areas of interests, processes, and requirements for proposals. Therefore, the investigator should start by deciding where the proposal will be submitted, determining the limit on amounts of funding, and obtaining **specific guidelines** about how to craft the proposal and deadlines for that particular agency. The NIH is a good place to start, at http://grants.nih.gov/grants/oer.htm. Areas of interest can be identified through the websites of individual institutes that describe their priorities. Additional information about current areas of interest can be obtained by **talking** with scientific administrators at the NIH Institutes, whose contact information and areas of responsibility are listed on NIH Funding Opportunity Announcements and institute websites.
- **Organize a team and designate a leader.** Most proposals are written by a team of several people who will eventually carry out the study. This team may be small (just the investigator and his mentor) or large (including collaborators, a biostatistician, a fiscal administrator, research assistants, and support staff). It is important that this team include or have access to the main expertise needed for designing the study.

  One member of the team must assume the responsibility for leading the effort. Generally this individual is the **principal investigator (PI)**, who will have the ultimate authority and accountability for the study. The PI must exert steady leadership during proposal development, delegating responsibilities for writing and other tasks, setting deadlines, conducting periodic meetings of the team, ensuring that all necessary tasks are completed on time, and personally taking charge of the quality of the proposal.

The PI is often an experienced scientist whose knowledge and wisdom are useful for design decisions and whose track record with previous studies increases the likelihood of a successful study and, therefore, of funding. That said, the NIH encourages **new investigators** to apply for grants as PIs, has special funding opportunities for them, and often gives preference to funding their proposals (http://grants.nih.gov/grants/new_investigators/). The NIH definition of "new investigator" is a scientist who has not yet been the PI of an NIH research grant. But **first-time PIs** are most likely to be funded if they already have some experience carrying out research—under the guidance of a senior scientist and with funding provided by that individual, by a career development award, or by small institutional or foundation grants. A track record in **publishing**, including some first authorships, is essential to provide evidence that the new investigator has the potential to be a successful independent scientist, and is prepared and able to lead the research.

A first-time PI should include **co-investigators** on the grant application who have a track record of successful research in the area of interest to provide guidance about the conduct of the study and to improve the chances of a favorable review. Sometimes this can be accomplished by the **multiple-PI** mechanism; NIH allows more than one PI on proposals if the PIs bring different, but complementary expertise and their distinct roles and responsibilities are clearly defined (http://grants.nih.gov/grants/multi_pi/overview.htm).

- **Follow the guidelines of the funding agency.** All funding sources provide written **guidelines** that the investigator must carefully study before starting to write the proposal. This information includes the types of research that will be funded and detailed instructions for organizing the proposal, page limits, the amount of money that can be requested, the timeline, and elements that must be included in the proposal.

  However, these guidelines do not contain *all* the important information that the investigator needs to know about the operations and preferences of the funding agencies. Early in the development of the proposal it is a good idea to **discuss the plan** with an individual at the agency who can clarify what the agency prefers (such as the scope and detail required in the proposal) and comment on whether the agency is interested in the planned research area. The NIH, other federal agencies and private foundations have scientific administrators ("**project officers**") whose job is to help investigators design their proposals to be more responsive to the agency's funding priorities. It can be very helpful to contact the project officer responsible for the relevant research area by e-mail or telephone to clarify the agency guidelines, interests, and review procedures. Subsequently, meeting with the project officer at a scientific conference that happens to be convenient or while traveling near the agency headquarters is a good way to establish a working relationship that promotes fundable proposals.

  It is useful to make a **checklist** of the details that are required, and to review the checklist repeatedly before submitting the proposal. Rejection of an otherwise excellent proposal for lack of adherence to specified details is a frustrating and avoidable experience. University grant managers generally have checklists that they review before submission of a proposal.

- **Establish a timetable and meet periodically.** A schedule for completing the writing tasks keeps gentle pressure on team members to meet their obligations on time. In addition to addressing the scientific components specified by the funding agency, the **timetable** should take into account the administrative requirements of the institution where the research will take place. Universities often require a time-consuming review of the budget and subcontracts before a proposal can be submitted to the funding agency, so the *real* deadline to complete a proposal may be several days or even weeks before the agency deadline. Leaving these details to the end can precipitate a last-minute crisis that damages an otherwise well-done proposal.

  A timetable generally works best if it specifies deadlines for written products and if each individual participates in setting his own assignments. The timetable should be reviewed at periodic meetings or conference calls of the writing team to check that the tasks are on schedule and the deadlines are still realistic.

- **Find model proposals.** It is extremely helpful to borrow recent **successful proposals** to the agency from which funding is being sought. Successful applications illustrate in a concrete way the format and content of a good proposal. The investigator can find inspiration for new ideas from the model and design and write a proposal that is clearer, more logical, and more persuasive. It is also a good idea to borrow examples of written criticisms that have been provided by the agency for previous successful or unsuccessful proposals. This will illustrate the key points that are important to the scientists who will be reviewing the proposal. These examples can often be obtained from colleagues or the Office of Sponsored Research at the investigator's institution.
- **Work from an outline.** Begin by setting out the proposal in outline form (Table 19.1). This provides a starting point for writing and is useful for organizing the tasks that need to be done.

TABLE 19.1 MAIN ELEMENTS OF A PROPOSAL, BASED ON THE NIH MODEL

- Title
- Project summary or abstract
- Administrative parts
  - Budget and budget justification
  - Biosketches of investigators
  - Facilities and resources
- Specific aims
- Research strategy
  - Significance
  - Innovation
  - Approach
    - Overview
    - Justification (rationale for the planned research and preliminary data)
    - Study subjects
      - Selection criteria
      - Design for sampling
      - Plans for recruitment
      - Plans to optimize adherence and complete follow-up
    - Study procedures (if applicable)
      - Randomization
      - Blinding
    - Measurements
      - Main predictor variables (intervention, if a clinical trial)
      - Potential confounding variables
      - Outcome variables
    - Statistics
      - Approach to statistical analyses
      - Hypotheses, sample size, and power
    - Content and timing of study visits
    - Data management and quality control
    - Timetable and organizational chart
    - Limitations and alternative approaches
- Human subjects
- References
- Appendices and collaborative agreements

If several people will be working on the grant, the outline helps in **assigning responsibilities** for writing parts of the proposal. One of the most common road-blocks to creating an outline is the feeling that an entire research plan must be worked out before starting to write the first sentence. The investigator should put this notion aside and let his thoughts flow onto paper, creating the raw material for editing, refining, and getting specific advice from colleagues.

- **Review and revise repeatedly.** Writing a proposal is an **iterative process**; there are usually many versions, each reflecting new ideas, advice, and additional data. Beginning early in the process of writing the proposal, drafts should be critically reviewed by colleagues who are familiar with the subject matter and the funding agency. Particular attention should go to the significance and innovativeness of the research, the validity of the design and methods, and the clarity of the writing. It is better to have sharp and detailed criticism before the proposal is submitted than to have the project rejected because of failure to anticipate and address problems. When the proposal is nearly ready for submission, the final step is to review it carefully for internal consistency; format; adherence to agency guidelines; and formatting, grammatical, and typographical errors. Sloppy writing implies sloppy work and incompetent leadership, and significantly detracts from otherwise good ideas.

■ ELEMENTS OF A PROPOSAL FOR A MAJOR GRANT

The elements of a proposal for a major research grant such as an NIH R01 are set out in Table 19.1. Applications for other types of NIH grants and contracts, and from other funding institutions, may require less information or a different format, and the investigator should pay careful attention to the guidelines of the agency that will receive the proposal.

The Beginning

The **title** should be descriptive and concise. It provides the first impression and a lasting reminder of the overall research goal and design of the study. For example, this title - "A randomized trial of MRI-guided high frequency ultrasound vs. sham ultrasound for treating symptomatic fibroids." - succinctly summarizes the research question and study design. Avoid unnecessary and empty phrases like "A study to determine the. . . ."

The **project summary** or **abstract** is a concise summary of the protocol that should begin with the research aims and rationale, then set out the design and methods, and conclude with a statement of the impact of potential findings of the study. The abstract should be informative to persons working in the same or related fields, and understandable to a scientifically literate lay reader. Most agencies require that the abstract be kept within a limited number of words, so it is best to use efficient and descriptive terms. The abstract should go through enough revisions to ensure that it is first rate. This will be the only page read by some reviewers, and a convenient reminder of the specifics of the proposal for everyone else. It must therefore stand on its own, incorporate all the main features of the proposed study, and persuasively describe the strengths and potential impacts.

The Administrative Parts

Almost all agencies require an administrative section that includes a budget and a description of the qualifications of personnel, the resources of the investigator's institution, and access to equipment, space, and expertise.

The **budget** section is generally organized according to guidelines from the funding institution. The NIH, for example, has a prescribed format that requires a detailed budget for the first 12-month period and a summary budget for the entire proposed project period (usually 2–5 years). The detailed 12-month budget includes the following categories of expenses: personnel (including names and positions of all persons involved in the project, the percent of time each will devote to the project, and the dollar amounts of salary and fringe benefits for each

individual); consultant costs; equipment; supplies; travel; patient care costs; alterations and renovations; consortium/contractual costs; and other expenses (e.g., the costs of telephones, mail, conference calls, copying, illustration, publication, books, fee-for-service contracts, etc.).

The budget should not be left until the last minute. Many elements require time (e.g., to get good estimates of the cost of space, equipment, and personnel). Universities generally employ knowledgeable administrators whose job is to help investigators prepare budgets and the other administrative parts of a proposal. The best approach is to notify this administrator as soon as possible about the plan to submit a proposal, and schedule regular meetings or calls with him to review progress and the timeline for finishing the administrative sections. An administrator can begin working as soon as the outline of the proposal is formulated, recommending the amounts for budget items and helping to ensure that the investigator does not overlook important expenses. Institutions have regulations that must be followed and deadlines to meet, and an experienced administrator can help the investigator anticipate his institution's rules, pitfalls, and potential delays. The administrator can also be very helpful in drafting the text of the sections on budget justification and resources, and in collecting the biosketches, appendices, and other supporting materials for the proposal.

The need for the amounts requested for each item of the budget must be fully explained in a **budget justification**. Salaries will generally comprise most of the overall cost of a typical clinical research project, so it is important to document the need and specific responsibilities for each person to justify the requested percent effort. Complete but concise job descriptions for the investigators and other members of the research team should leave no doubt in the reviewers' minds that the estimated effort of each individual is essential to the success of the project.

Reviewers are often concerned about the percentages of time committed by key members of the research team. Occasionally, proposals may be criticized because key personnel have only a very small commitment of time listed in the budget and a large number of other commitments implying that they may not be able to devote the necessary energy to the proposed study. More often, the reviewers may balk at percentages that are inflated beyond the requirements of the job description.

Even the best-planned **budgets** will **change** as the needs of the study change or there are unexpected expenses and savings. In general, once the grant is awarded the investigator is allowed to spend money in different ways from those specified in the budget, provided that the changes are modest and the expenditures are related to the aims of the study. When the investigator wants to move money across categories or to make a substantial change (up or down) in the effort of key investigators, he may need to get approval from the funding agency. Agencies generally approve reasonable requests for rebudgeting so long as the investigator is not asking for an increase in total funds.

The NIH requires a biosketch for all investigators and consultants who will receive funding from the grant. **Biosketches** are four-page resumes that follow a specified format that includes a personal statement about how the investigator's experience makes him well suited for conducting this study, and lists education and training, positions and employment, honors, a limited number of relevant publications, and relevant research grants and contracts.

The section of the proposal on **resources** available to the project may include computer and technical equipment, access to specialized imaging or measurement devices, office and laboratory space, and resources available to facilitate participant recruitment, data collection and management, and specimen storage. The resources section often draws on "boilerplate"—descriptions from previous proposals, or from material supplied by the investigator's institution, center, or laboratory.

Specific Aims

The specific aims are statements of the research question(s) using concrete terms to specify the desired outcome. This section of an NIH proposal must be **concise** because it is limited to

a single page. And because this is the page that many reviewers pay most attention to, it should be written carefully and revised repeatedly as the proposal is developed.

A common pattern is to begin with two to three short paragraphs that summarize the background information: the research question and why it's important, the studies that have been done and how they haven't solved the problem, and the approach that is planned for answering the question in the proposed study. This is followed by a concise statement of the specific aims, expressed as tangible descriptive objectives and, whenever possible, as testable **hypotheses**.

The aims are presented in a logical sequence that the investigator tailors to the study he plans. He may begin with cross-sectional aims for the baseline period followed by aims related to follow-up findings. Or he may begin with aims that address pathophysiologic mechanisms and end with aims that address clinical or public health outcomes. A pattern that works especially well for career development awards (termed "mixed methods research") begins with qualitative aims that may use focus groups to design a key instrument or intervention, followed by quantitative aims with predictors, outcomes and hypothesis tests. Yet another pattern is to start with the most important aim to highlight it; the sequence of aims often serves as an **outline** for organizing later sections of the proposal, so this has the advantage of giving the primary aim first place in all other sections of the proposal, such as sample size and power.

The Specific Aims section often ends with a short final paragraph that concisely sums up the potential **impact** of the study findings on knowledge of health and disease, clinical practice, public health, or future research. The goal is to make a compelling case that will lead review committee members who were not primary or secondary reviewers (and who may only have read this one page in the proposal) to support an outstanding score.

Research Strategy

The current NIH format limits most types of proposals to 12 pages for presenting the **research strategy**, in three sections:

- The **significance** section, typically two to three pages, describes how the study findings would advance scientific understanding, address an important problem or a barrier to progress in the field, improve clinical practice or public health, or influence policy. This section can briefly state the magnitude of the problem, summarize what has already been accomplished, define problems with current knowledge, and show how the proposed study will advance the field.
- The **innovations** section, typically one to two pages, points out ways the proposed study differs from prior research on the topic. It can emphasize the potential to document new mechanisms of disease, new measurement methods, different or larger populations, new treatment or prevention methods, or new approaches to analyzing the data. The NIH guidelines focus on how the research will shift current research or clinical practice paradigms by using innovative concepts, methods, or interventions. That said, many funded clinical studies result in only incremental improvements and refinements in concept, methods, or interventions. Our advice is to describe the novel features of the research *accurately*, without overstating claims that the study will change paradigms or use wholly innovative methods.
- The **approach section** (formerly termed "methods") is typically seven to nine pages long. It provides the details of study design and conduct, and receives close scrutiny from reviewers. NIH guidelines suggest that the approach section be organized by specific aims, and that it include the components and approximate sequence in Table 19.1. This section generally starts with a concise overview of the approach, sometimes accompanied by a schematic diagram or table, to orient the reader (Table 19.2). The overview should clearly state the study design and give a brief description of the study participants, the main measurements, any intervention, length of follow-up, and main outcome(s).

TABLE 19.2 STUDY TIMELINE FOR A RANDOMIZED TRIAL OF THE EFFECT OF TESTOSTERONE ADMINISTRATION ON RISK FACTORS FOR HEART DISEASE, PROSTATE CANCER, AND FRACTURES

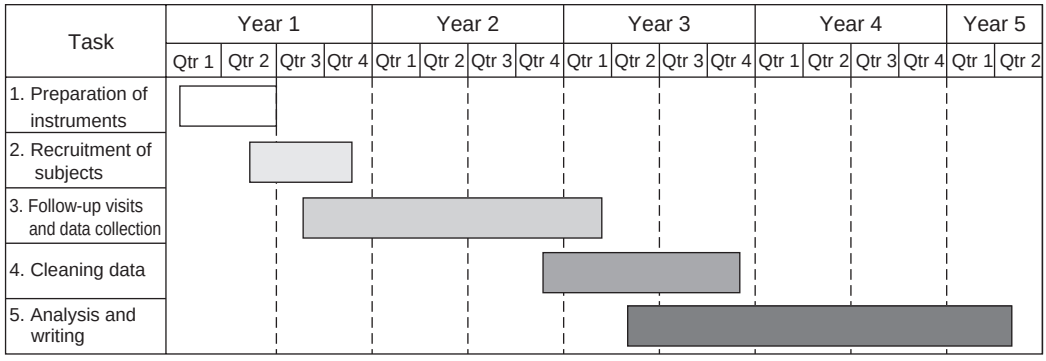
	SCREENING VISIT	RANDOMIZATI ON VISIT	3 MONTHS	6 MONTHS	12 MONTHS
Medical history	X	–	–	–	X
Blood pressure	X	X	X	X	X
Prostate examination	X	–	–	–	X
Prostate specific antigen	X	–	–	–	X
Blood lipid levels	–	X	X	X	X
Markers of inflammation	–	X	–	–	X
Bone density	–	X	–	–	X
Markers of bone turnover	–	X	X	–	X
Handgrip strength	–	X	X	X	X
Adverse events	–	–	X	X	X

The approach section typically includes a brief rationale for the research supported by **preliminary data**—previous research by the investigator and his team indicating that the proposed study will be successful. Emphasis should be placed on the importance of the previous work and on the reasons it should be continued or extended. Results of pilot studies that support the importance of the research question and the **feasibility** of the study are important to many types of proposals, especially when the research team has limited previous experience with the proposed methods, when the question is novel, and when there may be doubts about the feasibility of the proposed procedures or recruitment of participants. This is an opportunity to show that the investigator and his team have the specific experience and expertise necessary to conduct the study.

Other specific components of the approach section have been discussed earlier. The **study subjects** section (Chapter 3) should clearly define and provide a rationale for inclusion and exclusion criteria and specify the sampling method. It is important to describe how the study participants will be recruited and to assure the reviewers that the investigators are capable of enrolling the desired number of study participants. Plans for optimizing adherence to the study intervention (if applicable) and study visits should be provided.

The approach section should include a description of important **study procedures**, such as randomization and blinding. **Study measurements** (Chapter 4) should clearly describe how predictor, outcome, and potential confounding variables will be measured and at what point in the study these measurements will be made, as well as how interventions will be applied, and how the main outcome will be ascertained and measured.

The **statistics** section usually begins with the **plans for analysis**, organized by specific aim. The plan can be set out in the logical sequence; for example, first the descriptive tabulations and then the approach to analyzing associations among variables. This is followed by a discussion of **sample size and power** (Chapters 5 and 6) that should begin with a statement of the null hypothesis for the aim that will determine the sample size for the study. Estimates of sample size and power rely on assumptions about the magnitude of associations that are likely to be detected, and the precision of the measurements that will be made. These assumptions must be justified by referencing published literature or preliminary work that supports these judgments. It is often useful to include a table or figure showing how variations in the effect size, power, or other assumptions influence the sample size to demonstrate that the investigator has made reasonable choices. Most NIH review panels attach considerable importance to the statistical section, so it is a good idea to involve a statistician in writing this component.



■ **FIGURE 19.1** A hypothetical timetable.

It is helpful to include a table that lists **study visits** or participant contacts, the timing of visits, and what procedures and measurements will occur at each visit. Such a table provides a concise overview of all study activities (Table 19.2). Descriptions of **data management** and **quality control** (Chapters 16 and 17) should address how the study data will be collected, stored and edited, along with plans for maximizing data quality and security.

The proposal must provide a realistic work plan and **timetable**, including dates when each major phase of the study will be started and completed (Figure 19.1). Similar timetables can be prepared for staffing patterns and other components of the project. For large studies, an **organizational chart** describing the research team can indicate levels of authority and accountability, lines of reporting, and show how the team will function.

While not a required section, it can be helpful to include a discussion of the **limitations** of the proposed research and **alternative approaches**. Rather than ignore potential flaws, an investigator may decide to address them explicitly, discussing the advantages and disadvantages of the various trade-offs in reaching the chosen plan. Pointing out important challenges and potential solutions can turn potential criticisms of the application into strengths. It is a mistake to overemphasize these problems, however, for this may lead a reviewer to focus disproportionately on the weaker aspects of the proposal. The goal is to reassure the reviewer that the investigator has anticipated all of the important potential problems and has a realistic and thoughtful approach to dealing with them.

Final Components of a Major Proposal

The **human subjects** section is devoted to the ethical issues raised by the study, addressing issues of safety, privacy, and confidentiality. This section indicates the specific plans to inform potential participants of the risks and benefits, and to obtain their consent to participate (Chapter 14). It also describes the inclusion of women, children, and participants from minority groups, as required of NIH proposals, and justifies exclusion of any of these groups.

The **references** send a message about the investigator's familiarity with the field. They should be comprehensive but parsimonious and up-to-date—not an exhaustive and unselected list. Each reference should be cited accurately; errors in these citations or misinterpretation of the work will be viewed negatively by reviewers who are familiar with the field of research.

For some types of proposals, **appendices** can be useful for providing detailed technical and supporting material mentioned briefly in the text. (However, to avoid the use of appendices to circumvent page limits for proposals, NIH strictly limits their use.) Appendices may include data collection instruments (such as questionnaires) and clinical protocols, and up to three manuscripts and abstracts that have been accepted but not yet published. Primary and secondary reviewers are the only review committee members who receive the appendices. Therefore, everything important must be succinctly summarized in the main proposal.

The proposed use and value of each **consultant** should be described, accompanied by a signed letter of agreement from the individual and a copy of his biosketch. (Investigators who will receive salary support from the grant do not provide letters, because they are officially part of the proposal.) Other **letters of support** should also be included, such as those from persons who will provide access to equipment or resources. An explanation of the programmatic and administrative arrangements between the applicant organization and **collaborating institutions** and laboratories should be included, accompanied by letters of commitment from responsible officials addressed to the investigator.

CHARACTERISTICS OF GOOD PROPOSALS

A good proposal for research funding has several attributes. First is the **scientific quality** of the research strategy: It must be based on a good research question, use a design and approach that are rigorous and feasible, and have a research team with the experience, skill, and commitment to carry it out. Second is **clarity of presentation**; a proposal that is concise and engaging, well organized, thoughtfully written, attractively presented, and free of errors reassures the reader that the conduct of research is likely to be of similar high quality.

Reviewers are often overwhelmed by a large stack of proposals, so the merits of the project must stand out in a way that will not be missed even with a quick and cursory reading. A clear **outline** that follows the specific aims, short sections with meaningful **subheadings**, and the use of **tables and figures** to break up lengthy stretches of text can guide the reviewer's understanding of the important features of the proposal. Current NIH guidelines recommend starting paragraphs with a **topic sentence** in bold type that makes the key point, allowing harried reviewers to understand the essential elements of the proposal by quickly scanning topic sentences. It is important to consider the diverse points of view and expertise of the reviewers, including enough detail to convince an expert reviewer of the significance and sophistication of the proposed work while still engaging the larger number of reviewers unfamiliar with the area of investigation.

Most reviewers are put off by overstatement and other heavy-handed forms of grantsmanship. Proposals that exaggerate the importance of the project or overestimate what it can accomplish will generate skepticism. Writing with enthusiasm is good, but the investigator should be realistic about the limitations of the project. Reviewers are adept at identifying potential problems in design or feasibility.

A final round of scientific review by skilled scientists who have not been involved in developing the proposal, at a point in time when substantial changes are still possible, can be extraordinarily helpful to the proposal as well as a rewarding collegial experience. It is also useful to have someone with excellent writing skills supplement word-processing spell- and grammar-check programs with advice on style and clarity.

FINDING SUPPORT FOR RESEARCH

Investigators should be alert to opportunities to conduct good research without formal proposals for funding. For example, a beginning researcher may personally analyze data sets that have been collected by others, or receive small amounts of staff time from a senior scientist or his department to conduct small studies. Conducting research without funding of formal proposals is quicker and simpler but has the disadvantage that the projects are necessarily limited in scope. Furthermore, academic institutions often base decisions about career advancement in part on a scientist's track record of garnering funding for research.

There are four main cetegories of funds for medical research:

- **The government** (notably NIH, but also the Department of Veterans Affairs, Centers for Disease Control and Prevention [CDC], Agency for Healthcare Research and Quality [AHRQ], Patient Centered Outcomes Research Institute [PCORI], Department of Defense [DOD], and many other federal, state, and county agencies);

- **Foundations, professional societies** such as the American Heart Association (AHA) and the American Cancer Society (ACS)), and **individual donors**;
- **Corporations** (notably pharmaceutical and device manufacturing companies); and
- **Intramural resources** (e.g., the investigator's university).

Getting support from these sources is a complex and competitive process that favors investigators with **experience** and **tenacity**, and beginning investigators are well advised to find a mentor with these characteristics. In the following sections, we focus on several of the most important funding sources.

NIH Grants and Contracts

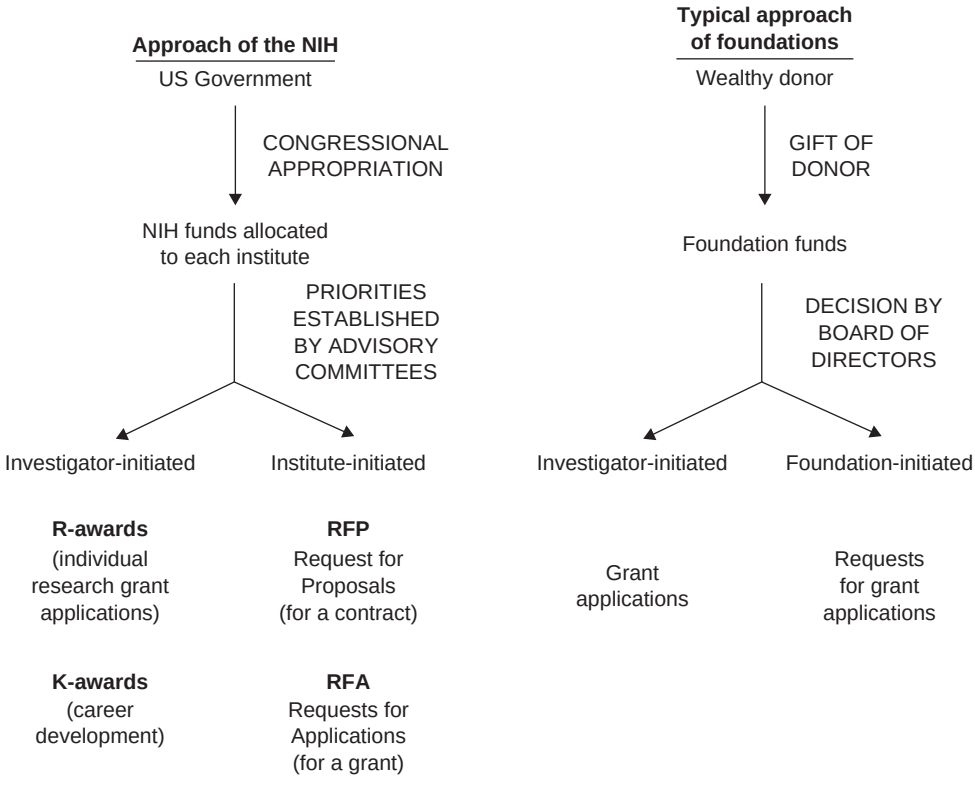
The NIH offers many types of grants and contracts. The "**R**" **awards** (R01 and smaller R03 and R21 awards) support research projects conceived by the investigator on a topic of his choosing or written in response to a publicized request by one of the institutes at NIH (see www.nimh.nih.gov/research-funding/grants/research-grants-r.shtml). The "**K**" **awards** (K23, K01, K08, K24 and locally awarded K12 and KL2 awards) are an excellent resource that provide salary support for training and career development of junior investigators, as well as modest support for research (see www.grants.nih.gov/training/careerdevelopmentawards.htm/).

Institute-initiated proposals are designed to stimulate research in areas designated by NIH advisory committees, and take the form of either Requests for Proposals (**RFPs**) or Requests for Applications (**RFAs**). In response to an RFP, the investigator contracts to perform certain research activities determined by the NIH. Under an RFA, the investigator conducts research in a topic area defined by the NIH, with a specific research question and study plan he has proposed. RFPs use the **contract** mechanism to reimburse the investigator's institution for the costs involved in achieving the planned objectives, and RFAs use the **grant** mechanism to support activities that are more open-ended.

After submitting a proposal, the application goes through a **review process** that includes an initial administrative review by NIH staff, **peer review** by a group of scientists, recommendations about funding by the institute advisory council, and the final decision about funding by the institute director. Grant applications are usually reviewed by one of many NIH "**study sections**," groups of scientific reviewers with a specific area of research expertise drawn from research institutions around the country. A list of the study sections and their current membership is available on the NIH website.

The NIH process for reviewing and funding proposals is described at cms.csr.nih.gov. When an investigator submits a grant application, it is assigned by the NIH Center for Scientific Review (CSR) to a particular study section (Figure 19.2). Proposals are assigned to a primary and two or more secondary reviewers who each provide a separate rating from 1 to 9 for **significance**, **innovation**, **approach**, **investigators**, and **environment**, and then an overall rating of the likely **impact** of the study. A score of "1" indicates an exceptionally strong application with essentially no weaknesses, and a "9" is an application with serious substantive weaknesses and very few strengths. The assigned reviewers' ratings are revealed to the study section, and proposals with scores in the upper half are discussed with the entire committee; the remainder are "triaged" (not discussed), with a few deferred to the next cycle 4 months later pending clarification of points that were unclear. After discussion, the assigned reviewers again propose ratings (the scores may have changed as a result of the discussion), and all committee members then provide scores by secret ballot. These are averaged, multiplied by 10 to yield an overall score from 10 (best) to 90 (worst), and used by each institute to prioritize funding decisions.

The investigator should decide in advance, with advice from senior colleagues, which study section would be the best choice to review the proposal. Study sections vary a great deal not only in topic area but also in the expertise of the reviewers and in the quality of competing applications. Although assignment to a study section cannot be fully controlled, the investigator



■ **FIGURE 19.2** Overview of NIH and foundation funding sources and mechanisms.

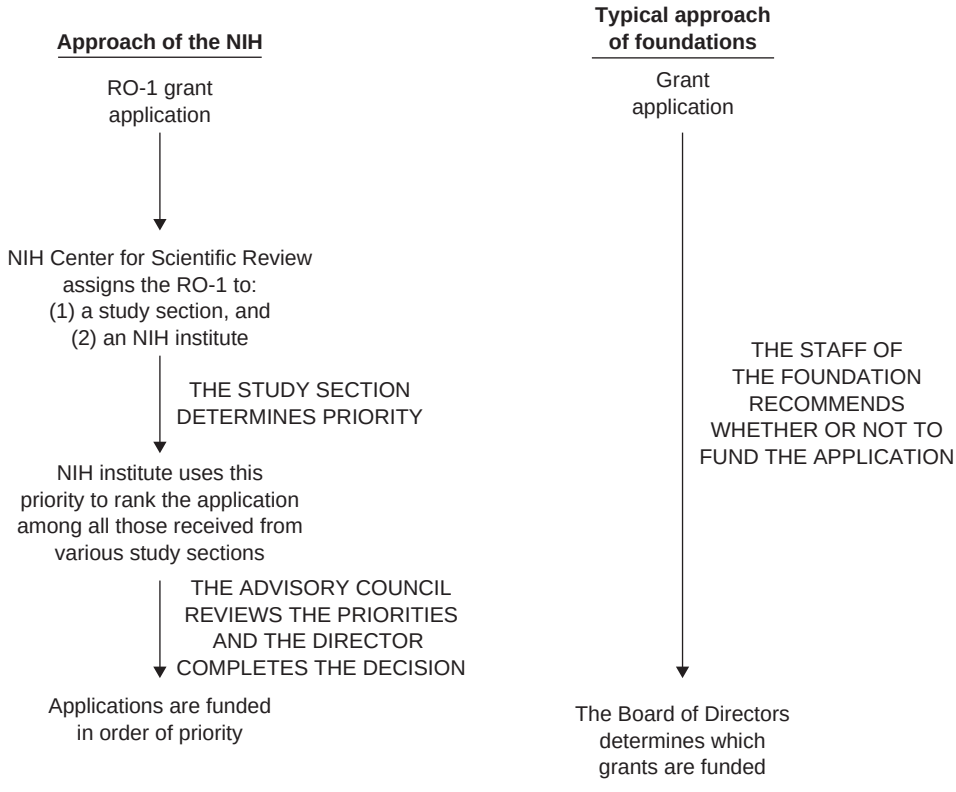
may be able to influence assignment by involving the project officer he has been working with in steering the application.

In addition to assigning each grant application to a particular study section, the CSR also assigns it to a particular **institute** (or center) at NIH. Each institute then funds the grants assigned to it, in order of priority score tempered by an advisory council review and sometimes over-ridden by the institute (Figure 19.3). Proposals from new investigators who have not yet received NIH research funding are funded at somewhat more lenient scores and percentile cutoffs than those from established investigators. If an application is of interest to more than one institute, the institutes sometimes arrange to share funding.

After an application has been reviewed, the investigator receives written notification of the study section's action. This **summary statement** includes the score and detailed comments and criticisms from the committee members who reviewed the application.

NIH applications that are not funded, as is often the case for the first submission, can be revised and resubmitted only once. If the reviewers' criticisms and scores suggest that the application can be made more attractive to the committee, then a revised version may have an excellent chance of obtaining funding when it is resubmitted. (It may be more difficult to raise reviewers' enthusiasm if they indicate that the proposal lacks innovation or significance.) Project officers from the relevant institute usually attend the study section meetings and it is important to discuss the review with one of them soon after the meeting because the written comments have usually been drafted before the meeting and may not reflect issues study section members raised that led to revisions in the scores.

An investigator need not automatically make all the changes suggested by reviewers, but he should adopt revisions that will satisfy the reviewer's criticisms wherever possible and justify any decision not to do so. NIH limits responses to reviews to a single page introduction describing the changes that have been made in the revised proposal. A good format for



■ **FIGURE 19.3** NIH and foundation procedures for reviewing grant applications.

the introduction is to succinctly summarize each major criticism from the summary statement in bold or italic font and address it with a concise statement of the consequent change in the proposal. To help reviewers focus on these revisions, changes should be marked, for example by a vertical line in the left hand margin of the text.

Grants from Foundations and Professional Societies

Private foundations (such as The Robert Wood Johnson Foundation) generally restrict their funding to specific areas of interest. Some disease-based foundations and **professional societies** (such as the American Heart Association and American Cancer Society) also sponsor research programs, many of which are designed to support junior investigators. The total amount of research support is far smaller than that provided by NIH, and most foundations have the goal of using this money to fund projects of merit that have topics or methods that are unlikely to be funded by NIH. A few foundations offer career development awards that are focused on specific areas such as quality of health care. **The Foundation Center** (http://fdncenter.org/) maintains a searchable directory of foundations and contact information, along with advice about how to write effective proposals to foundations. Decisions about funding follow procedures that vary from one foundation to another, but usually **respond rapidly** to relatively **short proposals** (Figure 19.3). The decisions are often made by an executive process rather than by peer review; typically, the staff of the foundation makes a recommendation that is ratified by a board of directors.

To determine whether a foundation might be interested in a particular proposal, an investigator should consult with his mentors and check the **foundation's website**. The website will generally describe the goals and purposes of the foundation and often list projects that have recently been funded. If it appears that the foundation might be an appropriate source

of support, it is best to contact the appropriate staff member of the foundation to describe the project, determine potential interest, and obtain guidance about how to submit a proposal. Many foundations ask that investigators send a letter describing the background and principal goals of the project, the qualifications of the investigators, and the approximate duration and costs of the research. If the letter is of sufficient interest, the foundation may request a more detailed proposal.

Research Support from Industry

Corporations that make drugs and devices are a major source of funding, especially for randomized trials of new treatments. Large companies generally accept applications for investigator-initiated research that may include small studies about the effects or mechanisms of action of a treatment, or epidemiologic studies about conditions of interest to the company. They will often supply the drug or device and a matching placebo for a clinical trial proposed by an investigator that is of interest to the company. They may also provide small grants to support educational programs in areas of their interest. However, by far the largest form of industry support for clinical research is through contracts to PIs of clinical sites for enrolling participants into **multicenter trials** that test new drugs and devices. These large trials are sometimes designed and managed by an academic coordinating center, but usually they are run by the corporate sponsor, often through a contract with a clinical research organization (CRO).

Requests for support for research or educational programs, or to participate as a site in a trial, generally begin by contacting the regional representative of the company. If the company is interested in the topic, the investigator may be asked to submit a relatively short application and complete a budget and other forms. Companies often give preference to requests from "**opinion leaders**," clinicians or investigators who are well known, have been involved in research or consultation with the company, and whose views may influence how other clinicians prescribe drugs or use devices. Therefore, a young investigator seeking industry support should generally get the help of a well-known mentor in contacting the company and submitting the application.

Contracts for enrolling participants in clinical trials generally pay clinical site PIs a fixed fee for each participant enrolled in a multi-site trial and the trial closes enrollment when the desired study-wide goal has been met. An investigator may enroll enough participants to receive funding that exceeds his costs, in which case he may retain the surplus as a long-term unrestricted account, but he will lose money if he recruits too few participants to pay the staff and overhead expenses for the trial. Before deciding to participate as a site in these multi-site trials, the investigator should be certain that the contract can be approved by the administrative offices and institutional review board of his institution in time to enroll enough participants before recruitment closes.

Funding from industry, particularly from marketing departments, is often channeled into topics and activities intended to increase sales of the company's product. The findings in industry-managed trials are generally analyzed by company statisticians and manuscripts are sometimes drafted by their medical writers. A number of site PIs are generally selected to be co-authors of peer-reviewed publications. Federal regulations require that authors have access to data (including the right to have analyses performed of study-wide data), make substantial contributions to manuscripts, and assume responsibility for the conclusions; we encourage site PI's to seek authorship roles for themselves and their co-investigators, and, if successful, to fulfill these **authorship requirements**. Ideally, analysis plans, manuscripts, and presentations from multicenter studies should be reviewed and approved by a publications committee that has written guidelines and a majority of members who are not employees of the sponsor.

An **advantage** of corporate support is that it is the only practical way to address some research questions. There would be no other source of funds, for example, for testing a new antibiotic that is not yet on the market. Another advantage is the relative speed with which

this source of funding can be acquired; decisions about small investigator-initiated proposals are made within a few months and drug companies are often eager to sign up qualified investigators to participate in multicenter clinical trials. Scientists at the company generally have extensive expertise about the treatment and about research methodology that can be useful in planning analyses and interpreting the results. Additionally, most pharmaceutical companies place a high premium on maintaining a reputation for integrity which enhances their dealings with the FDA and their stature with the public. The research expertise, statistical support, and financial resources they provide can improve the quality of the research.

Intramural Support

Universities typically have local research funds for their own investigators. Grants from these intramural funds are generally limited to relatively small amounts, but they are usually available much more **quickly** (weeks to months) and to a **higher proportion** of applicants than grants from the NIH or private foundations. Intramural funds may be restricted to special purposes, such as pilot studies that may lead to external funding, or the purchase of equipment. Such funds are often earmarked for junior faculty and provide a unique opportunity for a new investigator to acquire the experience of leading a funded project.

SUMMARY

1. A **proposal** is an expanded version of the detailed written plan of a study (the **protocol**) that is used to request funding and also contains budgetary, administrative and supporting information required by the funding agency.
2. An investigator who is working on a research proposal should begin by getting advice from senior colleagues about the **research question** he will pursue and the **choice of funding agency**. The next steps are to study that agency's written **guidelines** and to **contact a scientific administrator** in the agency for advice.
3. The process of writing a proposal, which often takes much longer than expected, includes organizing a **team** with the necessary expertise; designating a **principal investigator (PI)**; **outlining a proposal** to conform strictly to **agency guidelines**; establishing a **timetable** for written products; finding a **model proposal**; and reviewing progress at regular **meetings**. The proposal should be **reviewed** by knowledgeable colleagues, revised often, and polished at the end with attention to detail.
4. The major elements of a proposal include the **abstract** (summary), the **administrative parts** centered around the budget, budget justification, biosketches and resources, the very important **specific aims**, and the **research strategy** with its **significance**, **innovations**, and **approach** sections including **previous research** by the investigator.
5. A **good proposal** requires not only a **good research question**, **study plan**, and **research team**, but also a **clear presentation**: The proposal must communicate clearly and concisely, following a logical outline and indicating the advantages and disadvantages of **trade-offs** in the study plan. The **merits** of the proposal should stand out using **subheadings**, **tables and diagrams** so that they will not be missed by a busy reviewer.
6. There are four main sources of support for clinical research:
   a. The **NIH** and other governmental sources are the **largest** providers of support, using a complex system of peer and administrative review that moves slowly but funds a wide array of **grants and contracts** for **research**, and for **career development**.
   b. **Foundations and societies** are often interested in promising research questions that escape NIH funding, and have review procedures that are **quicker** but more **parochial** than those of NIH.

c. **Manufacturers of drugs and devices** are a very **large source** of support that is mostly channeled to company-run studies of **new drugs** and **medical devices**; however, corporations value partnerships with leading scientists and support some investigator-initiated research.
d. **Intramural funds** from the investigator's institution tend to have **favorable funding rates** for getting small amounts of money **quickly**, and are an excellent first step for **pilot studies** and **new investigators**.

Exercises

Chapter 1 Getting Started: The Anatomy and Physiology of Clinical Research

1. Appendix 1 provides an outline of the Early Limited Formula ("ELF") study carried out in two academic medical centers in California with the goal of encouraging breastfeeding by newborn babies who had lost ≥ 5% of their body weight. In this randomized clinical trial, the proportion of mothers who reported exclusive breastfeeding at 3 months to a blinded interviewer was 79% in the ELF group, compared with 42% in the control group ($P = 0.02$) (Flaherman et al. *Pediatrics* 2013;131 [in press]. For each of the following statements, indicate (1) whether it is an internal validity or external validity inference; (2) whether you think it is a valid inference; and (3) any reasons why it might *not* be valid.
   a. For the women in this study, provision of early limited formula increased breastfeeding rates at 3 months.
   b. Provision of early limited formula to infants with ≥ 5% weight loss in the first 36 hours born in a Boston community hospital will likely lead to higher breastfeeding rates at age 6 months.
   c. Based on the results of this study, an international effort to provide formula to most newborns is likely to enhance successful breastfeeding and improve the health of the newborns and their mothers.
2. For each of the following summaries drawn from published studies, *write* a *single sentence* that specifies the design and the research question, including the main predictor and outcome variables and the population sampled.
   a. Investigators in Winston-Salem, North Carolina, surveyed a random sample of 2,228 local high school students about their frequency of watching wrestling on television in the previous 2 weeks, and 6 months later asked the same students about fighting at school and on dates. The adjusted odds of reporting fighting with a date increased by 14% for each episode of wrestling the students reported having watched 6 months before. (DuRant et al. *Pediatrics* 2006;118:e265–272.)
   b. To assess whether the amount of breastfeeding protects women against ovarian cancer, investigators surveyed 493 Chinese women with newly diagnosed ovarian cancer and 472 other hospitalized women, all of whom had breastfed at least one child. They found a dose–response relationship between total months of breastfeeding and reduced risk of ovarian cancer. For example, women who had breastfed for at least 31 months had an odds ratio of 0.09 (95% CI 0.04, 0.19) compared with women who had breastfed less than 10 months (Su et al. *Am J Clin Nutr* 2013;97:354–359).
   c. To see whether an association between dietary saturated fat intake and reduced sperm concentration in infertile men extended to the general population, Danish investigators collected semen samples and food frequency questionnaires from consenting young men at the time of their examination for military service. They found a significant dose–response relation between self-reported dietary saturated fat intake and reduced sperm concentrations (e.g., 41% [95% CI 4%, 64%] lower sperm concentration in the highest

quartile of saturated fat intake compared with the lowest) (Jensen et al. *Am J Clin Nutr* 2013;97:411–418).

d. There is no known effective drug treatment for the ~20% of patients with *Clostridium difficile* diarrhea who relapse after treatment with antibiotics. Investigators in Amsterdam studied patients ≥ 18 years of age who had a relapse of *C. difficile* diarrhea following at least one course of adequate antibiotic therapy. They were randomly assigned (without blinding) to one of three regimens: a 5-day course of vancomycin followed by bowel lavage and infusion of a suspension of volunteer donor feces through a nasoduodenal tube or a standard 14-day course of vancomycin with or without bowel lavage on day 4 or 5. The trial was stopped early after an interim analysis showed the rate of cure without relapse for 10 weeks was 13 of 16 (81%) in the donor feces group, compared with 4 of 13 with vancomycin alone and 3 of 13 with vancomycin plus lavage ($P < 0.001$ for both comparisons)(van Nood et al. *New Engl J Med* 2013;368:407–415).

Chapter 2 Conceiving the Research Question and Developing the Study Plan

1. Consider the research question: "What is the relationship between depression and health?" First, convert this into a more informative description that specifies a study design, predictor, outcome, and population. Then discuss whether this research question and the design you have chosen meet the FINER criteria (Feasible, Interesting, Novel, Ethical, Relevant). Rewrite the question and design to resolve any problems in meeting these criteria.
2. Consider the research question: "Does acetaminophen (paracetamol) cause asthma?" Put yourself back to the year 2000, when this question was just starting to be asked, and provide 1-sentence descriptions of two observational studies and one clinical trial to progressively address this research question. Make sure each sentence specifies the study design, predictor, outcome, and population. Then, for each, consider whether this research question and the design you have chosen meet the FINER criteria (Feasible, Interesting, Novel, Ethical, Relevant).
3. Use the ideas in this chapter and your own interests to conceive a research question and devise a 1-page outline of a study you might carry out. Does it meet the FINER criteria? Discuss different designs, population samples, and variables with a colleague, seeking to optimize your study's FINER nature.

Chapter 3 Choosing the Study Subjects: Specification, Sampling, and Recruitment

1. An investigator is interested in the following research question: "What are the factors that cause people to start smoking?" She decides on a cross-sectional sample of high school students, invites eleventh graders in her suburban high school to participate, and studies those who volunteer.
   a. Discuss the suitability of this sample for the target population of interest.
   b. Suppose that the investigator decides to avoid the bias associated with choosing volunteers by designing a 25% random sample of the entire eleventh grade, and that the actual sample turns out to be 70% female. If it is known that roughly equal numbers of boys and girls are enrolled in the eleventh grade, then the disproportion in the sex distribution represents an error in drawing the sample. Could this have occurred through random error, systematic error, or both? Explain your answer.

2. An investigator is considering designs for surveying rock concert patrons to determine their attitudes toward wearing ear plugs during concerts to protect their hearing. Name the following sampling schemes for selecting individuals to fill out a brief questionnaire, commenting on feasibility and whether the results will be generalizable to all people who attend rock concerts.
   a. As each patron enters the theater, he or she is asked to throw a virtual die (on the investigator's cell phone). All patrons who throw a 6 are invited to fill out the questionnaire.
   b. As each patron enters the theater, he or she is asked to throw a virtual die. Men who throw a 1 and women who throw an even number are invited.
   c. Tickets to the concert are numbered and sold at the box office in serial order, and each patron whose ticket number ends in 1 is invited.
   d. After all the patrons are seated, five rows are chosen at random by drawing from a shuffled set of cards that has one card for each theater row. All patrons in those five rows are invited.
   e. The first 100 patrons who enter the theater are invited.
   f. Some tickets were sold by mail and some were sold at the box office just before the performance. Whenever there were five or more people waiting in line to buy tickets at the box office, the last person in line (who had the most time available) was invited.
   g. When patrons began to leave after the performance, those who seemed willing and able to fill out the questionnaire were invited.
3. Edwards et al. (Edwards et al. *N Engl J Med* 2013;368:633–643) reported on the burden of infection caused by human metapneumovirus (HMPV) among children < 5 years old. The subjects were children in counties surrounding Cincinnati, Nashville, and Rochester, NY, during the months of November to May, from 2003 to 2009, who sought medical attention for acute respiratory illness or fever. Consenting inpatients were enrolled Sunday through Thursday, outpatients 1 or 2 days per week, and emergency department patients 1 to 4 days per week. The authors combined the proportion of children testing positive at each site with nationwide data (from the National Ambulatory Medical Care Survey and the National Hospital Ambulatory Care Survey) on the population frequency of visits for acute respiratory illness or fever to estimate the overall burden of HMPV in the United States. They estimated HMPV was responsible for 55 clinic visits and 13 emergency department visits per 1,000 children annually.
   a. What was the target population for this study?
   b. What was the accessible population, and how suitable was it for generalizing to the target population?
   c. What was the sampling scheme, and how suitable was it for generalizing to the accessible population?
   d. Describe in general terms how the sampling scheme would need to be taken into account when calculating confidence intervals for the HMPV rates they calculate.

Chapter 4 Planning the Measurements: Precision, Accuracy, and Validity

1. Classify the following variables as dichotomous, nominal, ordinal, continuous, or discrete numerical. Could any of them be modified to increase power, and how?
   a. History of heart attack (present/absent)
   b. Age
   c. Education (college degree or more/less than college degree)
   d. Education (highest year of schooling)
   e. Race
   f. Number of alcohol drinks per day

g. Depression (none, mild, moderate, severe)
h. Percent occlusion of coronary arteries
i. Hair color
j. Obese (BMI ≥ 30)/non-obese (BMI<30)

2. An investigator is interested in the research question: "Does intake of fruit juice at age 6 months predict body weight at age 1 year?" She plans a prospective cohort study, measuring body weight using an infant scale. Several problems are noted during pretesting. Are these problems due to lack of accuracy, lack of precision, or both? Is the problem mainly due to observer, subject, or instrument variability, and what can be done about it?
   a. During calibration of the scale, a 10-kilogram (kg) reference weight weighs 10.2 kg.
   b. The scale seems to give variable results, but weighing the 10-kg reference weight 20 times gives a mean of 10.01 ± 0.2 (standard deviation) kg.
   c. Some babies are frightened and when they try to climb off the scale the observer holds them on it to complete the measurement.
   d. Some babies are "squirmy," and the pointer on the scale swings up and down wildly.
   e. Some of the babies arrive for the examination immediately after being fed, whereas others are hungry; some of the babies have wet diapers.
3. The investigator is interested in studying the effect of resident work hour limitations on surgical residents. One area she wishes to address is burnout, and she plans to assess it with two questions (answered on a 7-point scale) from a more extensive questionnaire: (a) "How often do you feel burned out from your work?" and (b) "How often do you feel you've become more callous toward people since you started your residency?"

   The investigator sets out to assess the validity of these questions for measuring burnout. For each of the following descriptions, name the type of validity being assessed:
   a. Residents with higher burnout scores were more likely to drop out of the program in the following year.
   b. These items seem like reasonable questions to ask to address burnout.
   c. Burnout scores increase during the most arduous rotations and decrease during vacations.
   d. A previous study of more than 10,000 medical students, residents, and practicing physicians showed that these two items almost completely captured the emotional exhaustion and depersonalization domains of burnout as measured by the widely accepted (but much longer) Maslach Burnout Inventory (West et al. *J Gen Intern Med* 2009;24:1318–1321).

Chapter 5 Getting Ready to Estimate Sample Size: Hypotheses and Underlying Principles

1. Define the concepts in **red font**.

   An investigator is interested in designing a study with sufficient **sample size** to determine whether body mass index is associated with stomach cancer in women between 50 and 75 years of age. She is planning a case–control study with equal numbers of cases and controls. The **null hypothesis** is that there is no difference in mean body mass index between cases of stomach cancer and controls; she has chosen an **alternative hypothesis** with two sides. She would like to have a **power** of 0.80, at a **level of statistical significance** (α) of 0.05, to be able to detect an **effect size** of a difference in body mass index of 1 kg/m^2 between cases and controls. Review of the literature indicates that the **variability** of body mass index among women is a standard deviation of 2.5 kg/m^2.
2. Which of the following is likely to be an example of a type I error? A type II error? Neither?
   a. A randomized trial finds that subjects treated with a new analgesic medication had greater mean declines in their pain scores during a study than did those treated with placebo ($P = 0.03$).

b. A 10-year study reports that 110 subjects who smoke do not have a greater incidence of lung cancer than 294 non-smokers ($P = 0.31$).
c. An investigator concludes that "Our study is the first to find that use of alcohol reduces the risk of diabetes in men less than 50 years of age ($P < 0.05$)."

Chapter 6 Estimating Sample Size and Power: Applications and Examples

1. Review exercise 1 of Chapter 5. Determine how many cases of stomach cancer would be required for the study. What if the investigator wanted a power of 0.90? Or a level of statistical significance of 0.01?
   Extra credit: Suppose the investigator only had access to 60 cases. What could she do?
2. Muscle strength declines with advancing age. Preliminary evidence suggests that part of this loss of muscle strength might be due to progressive deficiency of dehydroepiandrosterone (DHEA). Investigators plan a randomized trial to administer DHEA or an identical placebo for 6 months to elderly subjects, and then measure muscle strength. Previous studies have reported a mean grip strength in elderly persons of 20 kg with a standard deviation of 8 kg. Assuming α (two-sided) = 0.05 and $\beta = 0.10$, how many subjects would be required to demonstrate a 10% or greater difference between strength in the treated and placebo groups? How many subjects would be needed if $\beta = 0.20$?
3. In exercise 2, sample size calculations indicated more subjects were needed than can be enrolled. A colleague points out that elderly people have great differences in grip strength. This accounts for much of the variability in the strength measured after treatment and might be obscuring the treatment effect. She suggests that you measure strength at baseline and again after treatment, using the change in strength as the outcome variable. A small pilot study shows that the standard deviation of the change in strength during a 6-month period is only 2 kg. How many subjects would be required per group using this design, assuming α (two-sided) = 0.05 and $\beta = 0.10$?
4. An investigator suspects that left-handedness is more common in dyslexic than in non-dyslexic third graders. Previous studies indicated that about 10% of people are left-handed and that dyslexia is uncommon. A case–control study is planned that will select all the dyslexic students in a school district as cases, with an equal number of nondyslexic students randomly selected as controls. What sample size would be required to show that the odds ratio for dyslexia is 2.0 among left-handed students compared with right-handed students? Assume $\alpha = 0.05$ (two-sided) and $\beta = 0.20$.
5. An investigator seeks to determine the mean IQ of medical students in her institution, with a 99% CI of ± 3 points. A small pilot study suggests that IQ scores among medical students range from about 110 to 150. Approximately what sample size is needed?

Chapter 7 Designing Cross-Sectional and Cohort Studies

1. The research question is: "Does vitamin B_{12} deficiency cause hip fractures in the elderly?"
   a. Briefly outline a study plan to address this research question with a prospective cohort study.
   b. An alternative approach would be to take a sample from a geriatric clinic population and compare vitamin B_{12} levels in those who have had a previous hip fracture with levels in those who have not. Compared with this *cross-sectional* approach, list at least one advantage and one disadvantage of your prospective cohort study.
   c. Could the cohort study be designed as a retrospective study? If so, how would this affect these advantages or disadvantages?

2. Sung et al. (Sung et al. *Am J Obstet Gynecol* 2009 May;200(5):557.e1-5) examined the baseline association between the frequency of urinary incontinence and depressive symptoms among 338 overweight or obese women at least 30 years old enrolled in the PRIDE (Program to Reduce Incontinence by Diet and Exercise) clinical trial. They reported that women with depressive symptoms (N = 101) reported a higher mean number of incontinence episodes per week than women without depressive symptoms (28 vs 23; P = 0.005).
   a. What kind of study is this?
   b. One possible explanation for this is that depression increases the frequency of urinary incontinence. What are some other explanations for this association, and how might changes in the study design help you sort them out?

Chapter 8 Designing Case–Control Studies

1. The research question is: "How much does a family history of ovarian cancer increase the risk of ovarian cancer?" The investigator plans a case–control study to answer this question.
   a. How should she pick the cases?
   b. How should she pick the controls?
   c. Comment on potential sources of bias in the sampling of cases and controls.
   d. How would she measure "family history of ovarian cancer" as the predictor variable of interest? Comment on the sources of bias in this measurement.
   e. What measure of association would she use, and what test of statistical significance?
   f. Do you think the case–control method is an appropriate approach to this research question? Discuss the advantages and disadvantages of the case–control design relative to other possibilities for this research question.
2. The investigator wants to investigate the relationship between playing video games involving car racing and the risk of being involved in a real car crash (as the driver).
   a. Assume the exposure of interest is long-term effects of habitual use of these games. How would she select cases and controls and measure the exposure for a case–control study of this question?
   b. Now imagine the exposure of intererst is whether use of such games in the hour immediately preceding driving increases short-term risk. What is a design for studies of short-term effects of intermittent exposures? Lay out how such a study would be carried out for this research question.

Chapter 9 Enhancing Causal Inference in Observational Studies

1. The investigator undertakes a case–control study to address the research question: "Does eating more fruits and vegetables reduce the risk of coronary heart disease (CHD)?" Suppose that her study shows that people in the control group report a higher intake of fruits and vegetables than people with CHD.

   What are the possible explanations for this inverse association between intake of fruits and vegetables and CHD? Give special attention to the possibility that the association between eating fruits and vegetables and CHD may be confounded by exercise (if people who eat more fruits and vegetables also exercise more, and this is the cause of their lower CHD rates). What approaches could you use to cope with exercise as a possible confounder, and what are the advantages and disadvantages of each plan?
2. A study by the PROS (Pediatric Research in Office Settings) network of pediatricians found that among young infants (< 3 months) brought to their pediatricians for fever, uncircumcised

boys had about 10 times the risk of urinary tract infection, compared with circumcised boys (Newman et al.: *Arch Pediatr Adolesc Med* 2002 Jan;156(1):44–54), an association that has been seen in numerous studies. Interestingly, uncircumcised boys in that study appeared to have a *lower* risk of ear infections (risk ratio = 0.77; $P = 0.08$). Explain how including only babies with fever in this study could introduce an association between circumcision and ear infections that is not present in the general population of young infants.

3. In exercise 1 of Chapter 2, we asked you to suggest studies to address the question of whether acetaminophen causes asthma. A proposed mechanism for this association is acetaminophen-induced depletion of glutathione, which protects the lungs from oxidative injury that can lead to inflammation. Describe briefly how you could take advantage of variation in maternal anti-oxidant genotypes to enhance the inference that an association between maternal acetaminophen use and asthma in the offspring is causal.

Chapter 10 Designing a Randomized Blinded Trial

1. An herbal extract, huperzine, has been used in China as a remedy for dementia, and preliminary studies in animals and humans have been promising. The investigator would like to test whether this new treatment might decrease the progression of Alzheimer's disease. Studies have found that the plasma level of Abeta (1–40) is a biomarker for Alzheimer's disease: Elevated levels are associated with a significantly increased risk of developing dementia and the levels of Abeta (1–40) increase with the progression of dementia. In planning a trial to test the efficacy of huperzine for prevention of dementia in elderly patients with mild cognitive impairment, the investigator considers two potential outcome measurements: change in Abeta (1–40) levels or incidence of a clinical diagnosis of dementia.
   a. List one advantage and one disadvantage of using Abeta (1–40) as the primary outcome for your trial.
   b. List one advantage and one disadvantage of using the new diagnosis of dementia as the primary outcome for the trial.
2. A relatively large (>200 person per arm) trial of huperzine is being planned. The primary aim is to test whether this herbal extract decreases the incidence of a clinical diagnosis of dementia among elderly men and women with mild cognitive impairment.
   a. Huperzine is expected to occasionally cause gastrointestinal symptoms, including diarrhea, nausea, and vomiting. Describe a plan for assessing adverse effects of this new treatment on symptoms or diseases besides dementia.
   b. Describe a general plan for baseline data collection: what types of information should be collected?
   c. People who carry an Apoε4 allele have an increased risk of dementia. List one reason in favor and one against using stratified blocked randomization instead of simple randomization to assure a balance of people with the Apoε4 genotype in the treatment and control group.

Chapter 11 Alternative Clinical Trial Designs and Implementation Issues

Topical finasteride is moderately effective in treating male pattern baldness and is approved by the U.S. Food and Drug Administration (FDA) for treating this condition. Statins have been found to increase hair growth in rodents and they act by a different pathway than does finasteride. Imagine that a start-up company wants to obtain FDA approval for marketing a new topical statin (HairStat) for the treatment of male pattern baldness.

1. Describe a phase I trial of HairStat for male pattern baldness. What would be the treatment group(s)? What type of outcomes are expected?

2. The company wants to compare the efficacy of HairStat to finasteride. List at least one advantage and one disadvantage of the following approaches to testing the relative effectiveness of finasteride and the topical statin.
   a. Randomize bald men to either finasteride or topical statin.
   b. In a factorial design, randomly assign men to (1) finasteride and HairStat, (2) finasteride and HairStat-placebo, (3) finasteride-placebo and HairStat, or (4) double placebo.
3. Imagine that the company plans a 1-year placebo-controlled study of HairStat for treatment of baldness. The outcome is change in rating of the amount of hair in photographs of the bald region that is undergoing treatment. Follow-up visits (with photographs) are scheduled every 3 months. Outline a plan—with at least two elements—for encouraging compliance with the study treatment and return for visits to assess the outcome.
4. During the study, 20% of the men in the trial did not return for the 3-month follow-up visit and 40% stopped by 1 year. Some stopped because a rash developed on their scalp. List one disadvantage and one advantage of analyzing the effect of treatment on hair growth by a strict intention-to-treat approach.
5. In the intention-to-treat analysis, HairStat increased hair growth (rated by blinded outcome assessors based on comparison of baseline and 1-year photographs) 20% more than placebo ($P = 0.06$). Subsequent analyses showed that HairStat increased hair growth 45% more than placebo in men younger than age 40 ($P = 0.01$ in that subgroup). What are the problems with the company's conclusion that HairStat is effective for treating baldness in men younger than age 40?

Chapter 12 Designing Studies of Medical Tests

1. You are interested in studying the usefulness of the erythrocyte sedimentation rate (ESR) as a test for pelvic inflammatory disease (PID) in women with abdominal pain.
   a. To do this, you will need to assemble groups of women who do and do not have PID. What would be the best way to sample these women?
   b. How might the results be biased if you used final diagnosis of PID as the gold standard and those assigning that diagnosis were aware of the ESR?
   c. You find that the sensitivity of an ESR of at least 20 mm/hr is 90%, but the specificity is only 50%. On the other hand, the sensitivity of an ESR of at least 50 mm/hr is only 75%, but the specificity is 85%. Which cutoff should you use to define an abnormal ESR?
2. You are interested in studying the diagnostic yield of computed tomography (CT) head scans in children presenting to the emergency department (ED) with head injuries. You use a database in the radiology department to find reports of all CT scans done on patients less than 18 years old and ordered from the ED for head trauma. You then review the ED records of all those who had an abnormal CT scan to determine whether the abnormality could have been predicted from the physical examination.
   a. Out of 200 scans, 10 show intracranial injuries. However, you determine that in 8 of the 10, there had been either a focal neurological examination or altered mental status. Since only two patients had abnormal scans that could not have been predicted from the physical examination, you conclude that the yield of "unexpected" intracranial injuries is only 2 in 200 (1%) in this setting. What is wrong with that conclusion?
   b. What is wrong with using all intracranial injuries identified on CT scan as the outcome variable for this diagnostic yield study?
   c. What would be some advantages of studying the effects of the CT scan on clinical decisions, rather than just studying its diagnostic yield?
3. You now wish to study the sensitivity and specificity of focal neurological findings to predict intracranial injuries. (Because of the small sample size of intracranial injuries, you increase the sample size by extending the study to several other EDs.) One problem you

have when studying focal neurological findings is that children who have them are much more likely to get a CT scan than children who do not. Explain how and why this will affect the sensitivity and specificity of such findings if:

a. Only children who had a CT scan are included in the study.
b. Eligible children with head injuries who did not have a CT scan are included, and assumed not to have had an intracranial injury if they recovered without neurosurgical intervention.

Chapter 13 Research Using Existing Data

1. The research question is: "Do Latinos in the United States have higher rates of gallbladder disease than whites, African Americans, or Asian Americans?" What existing databases might enable you to determine race-, age-, and sex-specific rates of gallbladder disease at low cost in time and money?
2. A research fellow became interested in the question of whether mild or moderate renal dysfunction increases risk for coronary heart disease events and death. Because of the expense and difficulty of conducting a study to generate primary data, he searched for an existing database that contained the variables he needed to answer his research question. He found that the Cardiovascular Health Study (CHS), a large, NIH-funded multicenter cohort study of predictors of cardiovascular disease in older men and women, provided all of the variables required for his planned analysis. His mentor was able to introduce him to one of the key investigators in CHS who helped him prepare and submit a proposal for analyses that was approved by the CHS Steering Committee.
   a. What are the advantages of this approach to study this question?
   b. What are the disadvantages?
3. An investigator is interested in whether the effects of treatment with postmenopausal estrogen or selective estrogen receptor modulators (SERMs) vary depending on endogenous estrogen levels. How might this investigator answer this question using an ancillary study?

Chapter 14 Addressing Ethical Issues

1. The research question is to identify genes that are associated with an increased risk of developing Type II diabetes mellitus. The investigator finds that frozen blood samples and clinical data are available from a large prospective cohort study on risk factors for coronary artery disease that has already been completed. That study collected baseline data on diet, exercise, clinical characteristics, and measurements of cholesterol and hemoglobin A1c. Follow-up data are available on coronary endpoints and the development of diabetes. The proposed study will carry out DNA sequencing on participants; no new blood samples are required.
   a. Can the proposed study be done under the original informed consent that was obtained for the cohort study?
   b. If the original informed consent did not provide permission for the proposed study, how can the proposed study be done?
   c. When designing a study that will collect blood samples, how can investigators plan to allow future studies to use their data and samples?
2. The investigator plans a phase III randomized controlled trial of a new cancer drug that has shown promise in treating colon cancer. To reduce sample size, he would like to carry out a placebo-controlled trial rather than compare it to current therapy.
   a. What are the ethical concerns about a placebo control in this situation?
   b. Is it possible to carry out a placebo-controlled study in an ethically acceptable manner?

3. The investigator plans a study to prepare for a future HIV vaccine trial. The goals of the study are to determine (1) if it is possible to recruit a cohort of participants who have a high HIV seroconversion rate despite state-of-the-art HIV prevention counseling, and (2) if the follow-up rate in the cohort will be sufficiently high to carry out a vaccine trial. Participants will be persons at increased risk for HIV infection, including injection drug users, persons who trade sex for money, and persons with multiple sexual partners. Most participants will have low literacy and poor health literacy. The study will be an observational cohort study, following participants for 2 years to determine seroconversion and follow-up rates.
   a. What do the federal regulations require to be disclosed to participants as part of informed consent?
   b. What steps can be taken to ensure that consent is truly informed in this context?
   c. What is the investigator's responsibility during this observational study to reduce the risk of HIV infection in these high-risk participants?

Chapter 15 Designing Questionnaires, Interviews, and Online Surveys

1. As part of a study of alcohol and muscle strength, an investigator plans to use the following item for a self-response questionnaire to determine current use of alcohol:
   "How many drinks of beer, wine, or liquor do you drink each day?"
   ◯ 0
   ◯ 1–2
   ◯ 3–4
   ◯ 5–6
   ◯ 7–8
   Briefly describe at least two problems with this item.
2. Write a short series of questions for a self-response questionnaire that will better assess current alcohol use.
3. Comment on the advantages and disadvantages of a self-response questionnaire versus a structured interview to assess risky sexual behavior.

Chapter 16 Data Management

1. Refer to the first six items on the sample questionnaire about smoking in Appendix 15. You have responses for three study subjects:

Subject ID	Description of Smoking History
1001	Started smoking at age 17 and has continued to smoke an average of 30 cigarettes/day ever since
1002	Started smoking at age 21 and smoked 20 cigarettes/day until quitting 3 years ago at age 45
1003	Smoked a few cigarettes (<100) in high school

   Create a data table containing the responses of these subjects to the first six questions in Appendix 15. The table should have three rows (one for each subject) and seven columns (one for Subject ID, and one each for the six questions).
2. The PHTSE (Pre-Hospital Treatment of Status Epilepticus) Study (Lowenstein et al. *Control Clin Trials* 2001;22:290–309; Alldredge et al., *N Engl J Med* 2001;345:631–637) was a randomized blinded trial of lorazepam, diazepam, or placebo in the treatment of prehospital status epilepticus. The primary endpoint was termination of convulsions by hospital arrival. To enroll patients, paramedics contacted base hospital physicians by radio. The following are base-hospital physician data collection forms for two enrolled patients:

PHTSE

Base Hospital Physician Data Collection Form

PHTSE Subject ID :

Study Drug Administration: *189*

Study Drug Kit #: *A322*

Date and Time of Administration : *3 / 12 / 94* *17 : 39* **(Use 24 hour clock)**

Transport Evaluation
Seizure Stopped
Time Seizure Stopped: *17 : 44* (Use 24 hour clock)

Final ("End-of-Run") Assessment
Time of Arrival at Receiving Hospital ED: *17 : 48* (Use 24 hour clock)

On arrival at the receiving hospital:
[*X*] 1 Seizure activity (active tonic/clonic convulsions) continued
[ ] 0 Seizure activity (active tonic/clonic convulsions) stopped Verbal GCS
- [ ] 1 No Verbal Response
- [ ] 2 Incomprehensible Speech
- [ ] 3 Inappropriate Speech
- [ ] 4 Confused Speech
- [ ] 5 Oriented

PAGE ON-CALL PHTSE STUDY PERSONNEL !!

PHTSE

Base Hospital Physician Data Collection Form

PHTSE Subject ID :

Study Drug Administration *410*

Study Drug Kit #: *B536*

Date and Time of Administration : *12 / 01 / 98* *01 : 35* **(Use 24 hour clock)**

Transport Evaluation
[*X*]Seizure Stopped
Time Seizure Stopped *01 : 39* (Use 24 hour clock)

Final ("End-of-Run") Assessment
Time of Arrival at Receiving Hospital ED: *01 : 53* (Use 24 hour clock)

On arrival at the receiving hospital:
[ ] 1 Seizure activity (active tonic/clonic convulsions) continued
[*X*] 0 Seizure activity (active tonic/clonic convulsions) stopped
Verbal GCS
[ ] 1 No Verbal Response
[ ] 2 Incomprehensible Speech
[ ] 3 Inappropriate Speech
[*X*] 4 Confused Speech
[ ] 5 Oriented

PAGE ON-CALL PHTSE STUDY PERSONNEL !!

a. Display the data from these two data collection forms in a two-row data table.
b. Create a nine-field data dictionary for the data table in exercise 2a.
c. The paper data collection forms were completed by busy base hospital physicians who were called from the emergency department to a radio room. What are the advantages and disadvantages of using an on-screen computer form instead of a paper form? If you designed the study, which would you use?

3. The data collection forms in exercise 2 include a question about whether seizure activity continued on arrival at the receiving hospital (which was the primary outcome of the study). This data item was given the field name ***HospArrSzAct*** and was coded 1 for yes (seizure activity continued) and 0 for no (seizure activity stopped).

   Interpret the average values for ***HospArrSzAct*** as displayed below:

	HospArrSzAct	
	(1 = Yes, seizure continued; 0 = No, seizure stopped)	
	N	**Average**
Lorazepam	66	0.409
Diazepam	68	0.574
Placebo	71	0.789

Chapter 17 Implementing the Study and Quality Control

1. An investigator carried out a study of the research question: "What are the predictors of death following hospitalization for myocardial infarction?" Research assistants collected detailed data from charts and conducted extensive interviews with 120 hospitalized patients followed over the course of 1 year. About 15% of the patients died during the follow-up period. When data collection was complete, one of the research assistants entered the data into a computer using a spreadsheet. When the investigator began to run analyses of the data, he discovered that 10% to 20% of some predictor variables was missing, and others did not seem to make sense. Only 57% of the sample had been seen at the 1-year follow-up, which was now more than a year overdue for some subjects. You are called in to consult on the project.
   a. What can the investigator do now to improve the quality of his data?
   b. Briefly describe at least three ways that he could reduce missing values and errors in his next study.

Chapter 18 Community and International Studies

1. The investigator decides to study the characteristics and clinical course of patients with abdominal pain of unclear etiology. He plans to enroll patients with abdominal pain in whom no specific cause can be identified after a standard battery of tests. There are two options for recruiting study subjects: (1) the G.I. clinic at his university medical center, or (2) a local network of community clinics. What are the advantages and disadvantages of each approach?
2. The investigator has been assigned to work with the Chinese Ministry of Health in a new program to prevent smoking-related diseases in China. Of the following research questions, to what degree does each require local research as opposed to research done elsewhere?
   a. What is the prevalence and distribution of cigarette smoking?
   b. What diseases are caused by smoking?
   c. What strategies are most effective for encouraging people to quit smoking?

Chapter 19 Writing a Proposal for Funding Research

1. Search the NIH website (http://grants.nih.gov/grants/oer.htm) to find at least three types of investigator-initiated R-series grant awards.
2. Search the Foundation Center website (http://fdncenter.org/) for foundations that might be interested in the area of your research. List at least two.
3. Contact mentors and colleagues to find a research protocol that addresses a question in your area of interest and was funded. Read this protocol carefully.

Answers to Exercises

Chapter 1 Getting Started: The Anatomy and Physiology of Clinical Research

1a. This is an internal validity inference (because it refers to the women in this study) that is probably valid. However, it could be invalid if something other than Early Limited Formula (ELF) caused the difference in breastfeeding rates (e.g., if the control intervention adversely affected breastfeeding), if self-reported breastfeeding does not reflect actual breastfeeding, or if the association is not causal (the $P = 0.02$ does not rule out that it occurred by chance).

1b. This is an external validity inference (because it involves generalizing outside the study) that may be valid. However, in addition to the threats to internal validity above (which also threaten external validity) it is likely that women giving birth in community hospitals and in other parts of the country might respond differently to the intervention, or that other clinicians providing the ELF might carry out the intervention differently from the way it was done in the original study, or that the benefits might not last as long as 6 months.

1c. This is an external validity inference that goes far beyond the population and intervention that were studied and is probably not valid. It involves generalizing not only to other mothers and newborns in other locations, but also includes newborns who have not lost 5% of their body weight; expands the intervention from early, limited formula to providing formula without limitation; and asserts broad, vague health benefits that, while reasonable, were not examined in the ELF study.

2a. This is a cohort study of whether watching wrestling on television predicts subsequent fighting among Winston-Salem high school students.

2b. This is a case–control study of whether the duration of breastfeeding is associated with reduced risk of ovarian cancer among Chinese women who have breastfed at least one baby.

2c. This is a cross-sectional study of the relationship between self-reported saturated fat intake and sperm concentration in Danish men being examined for military service.

2d. This is an open-label randomized trial of whether a short course of vancomycin, bowel lavage, and duodenal infusion of donor feces improves the 10-week cure rate in adults with recurrent *C. difficile* diarrhea, compared with a standard vancomycin regimen with and without bowel lavage.

Each of these four sentences is a concise description that summarizes the entire study by noting the design and the main elements of the research question (key variables and population). For example, in exercise 2a the design is a cohort study, the predictor is watching wrestling on television, the outcome is fighting, and the population is high school students in Winston-Salem.

Chapter 2 Conceiving the Research Question and Developing the Study Plan

1. The process of going from research question to study plan is often an iterative one. One might begin with an answer like: "a cross-sectional study to determine whether depression

is associated with health status among young adults." The possibility that "depression" is related to "health status" seems Interesting and Relevant, but the question as stated is still too vague to judge whether the study is Feasible, Novel, and Ethical. How will depression and health status be measured, and in what population? Also, it will be difficult to establish causality in a cross-sectional design—does depression lead to worse health or vice versa?

A more specific design that could better meet the FINER criteria (depending on how it is fleshed in) might be: "A cohort study to determine whether depression among college juniors, assessed by the CES-D questionnaire, predicts their number of medical illness visits to the student health service in the next year."

2. In the case of the association between acetaminophen and asthma, the observation that acetaminophen use and asthma prevalence have both increased worldwide (and biologic plausibility related to depletion of reduced glutathione by acetaminophen) lead to all studies being Interesting and Relevant; as more studies are done they become less Novel.
Study #1: A case–control study to compare the self-reported frequency of acetaminophen use among adults with asthma symptoms seen in South London general practices (the cases), with the frequency reported by randomly selected adults without such symptoms from the same general practices (the controls). Case–control studies are often a good way to start investigating possible associations (Chapter 8). This study was especially Feasible because it was part of a larger population-based case–control study already investigating the role of dietary antioxidants in asthma. Odds ratios for asthma increased with frequency of acetaminophen use, up to 2.38 (95% CI 1.22 to 4.64) among daily users; *P* for trend = 0.0002). The study was Ethical because it was an observational study that did not put the subjects at risk (Shaheen et al. *Thorax* 2000;55:266–270).
Study #2: A multinational cross-sectional study of parent-reported allergic symptoms (asthma, hay fever, and eczema) among 6- to 7-year-old children that included questions about use of acetaminophen in the previous year and usual use for fevers in the first year after birth. This study (which included 205,487 children ages 6 to 7 years from 73 centers in 31 countries) would not have been feasible if it had not been part of the more general International Study of Asthma and Allergies in Childhood (ISAAC) study. This illustrates the importance of seeking existing data or existing studies when investigating a new research question (Chapter 13). The authors found a strong dose–response relationship between current use of acetaminophen and wheezing, and an odds ratio of 1.46 (95% CI 1.36 to 1.56) for wheezing and a "yes" answer to the question: "In the first 12 months of your child's life, did you usually give paracetamol [acetaminophen] for fever?" (Beasley et al. *Lancet* 2008;372:1039–1048).
Study #3: A randomized double-blind trial of the effect of acetaminophen (12 mg/kg) versus ibuprofen (5 or 10 mg/kg) on hospitalizations and outpatient visits for asthma over 4 weeks among febrile children 6 months to 12 years old who were being treated for asthma at enrollment. A randomized trial is generally the least feasible design, because of the expense and logistics involved. In addition, as evidence of a potential adverse drug effect accumulates, randomized trials to confirm it become less ethical. In this case, investigators did a retrospective analysis of data on the subset of children with asthma in the Boston University Fever Study, a randomized double-blind trial that had completed enrollment in 1993. They found that children randomized to acetaminophen had a 59% higher risk of asthma hospitalization (NS) and a 79% higher risk of an outpatient visit for asthma (RR = 1.79, 95% CI: 1.05, 2.94; *P* = 0.01) (Lesko et al. *Pediatrics* 2002;109:E20).

Chapter 3 Choosing the Study Subjects: Specification, Sampling, and Recruitment

1a. This sample of eleventh graders may not be well suited to the research question if the antecedents of smoking take place at an earlier age. A target population of greater

interest might be students in junior high school. Also, the accessible population (students at this one high school) may not adequately represent the target population—the causes of smoking differ in various cultural settings, and the investigator might do better to draw her sample from several high schools randomly selected from the whole region. Most important, the sampling design (calling for volunteers) is likely to attract students who are not representative of the accessible population in their smoking behavior.

1b. The unrepresentative sample could have resulted from random error, but this would have been unlikely unless it was a very small sample. If the sample numbered 10, a 7:3 disproportion would occur fairly often as a result of chance; in fact, the probability of selecting at least seven girls from a large class that is 50% girls is about 17% (plus another 17% chance of selecting at least seven boys). But if the sample size were 100, the probability of sampling at least 70 girls is less than 0.01%. This illustrates the fact that the investigator can estimate the magnitude of the random component of sampling error once the sample has been acquired and that she can reduce it to any desired level by enlarging the sample size.

The unrepresentative sample could also have resulted from systematic error. The large proportion of females could have been due to different rates of participation among boys and girls. The strategies for preventing non-response bias include the spectrum of techniques for enhancing recruitment discussed in Chapter 3. The large proportion of females could also represent a technical mistake in enumerating or selecting the names to be sampled. The strategies for preventing mistakes include the appropriate use of pretesting and quality control procedures (Chapter 17).

2a. Random sample (probability). The main concern for generalizability will be non-response—it will be important to keep the questionnaire short, and to provide some incentive to fill it out. (Possible non-response bias is an issue for all sampling schemes discussed in this question.)

2b. Stratified random sample (probability), with a threefold over-sampling of women, perhaps because the investigator anticipated that fewer women would attend the concert.

2c. Systematic sample (non-probability). While perhaps convenient, this systematic sampling scheme would lead to underrepresentation of both members of couples. Also, at least theoretically, the vendor in the box office could manipulate which patrons receive tickets ending in 1.

2d. Cluster sample (probability). This may be convenient, but the clustering needs to be taken into account in analyses, because people who sit in the same row may be more similar to one another than randomly selected concert-goers. This could be a particular problem if the music was louder in some rows than others.

2e. Consecutive sample (non-probability). Consecutive samples are usually a good choice, but people who arrive early for concerts may differ from those who arrive later, so several consecutive samples selected at different times would be preferable.

2f. Convenience sample (non-probability). This scheme will miss subjects who bought tickets by mail. In addition, people who come to concerts in groups could be over- or under-represented.

2g. Convenience sample (non-probability). This sampling scheme is not only biased by the whims of the investigator, it may run into non-response by patrons who are unable to hear the invitation.

3a. The target population (to which the authors wished to generalize) was the U.S. population of children under age 5 in the years they studied. We know this because the authors used nationwide survey data to estimate the U.S. human metapneumovirus (HMPV) disease burden. Of course it would be of great interest to generalize to future years as well, and

many readers will do so without a second thought. However, it is important to realize, especially with infectious diseases that can vary from year to year, that generalizing beyond the years of the study is an additional, potentially fragile, inference.

3b. The accessible population (the population from which they drew their subjects) was children < 5 years old living in the counties surrounding the three study sites (Cincinnati, Nashville, and Rochester, NY) and obtaining care from the study sites. Presumably these cities were selected because of their proximity to the investigators. It is not clear how representative they are of other areas in the United States with respect to the frequency of HMPV infection.

3c. The sampling scheme was a convenience sample. The choice of days of the week (which is not specified) could have led to some bias if, for example, parents of children with milder respiratory symptoms over the weekend wait until Monday to bring them to see a doctor and HMPV symptoms are more or less severe than those of other viruses. On the days when the investigators were enrolling subjects, they may have tried to get a consecutive sample (also not specified), which would have helped to control selection bias. The reason for restriction to certain months of the year is not provided, but was presumably because the authors believed almost all HMPV cases would occur during these months.

3d. The observations were clustered by geographic area, so the clustering by city would need to be taken into account statistically. The more different the estimates were between cities, the more this would widen the confidence intervals. Intuitively, this makes sense. Very different rates by city would lead one to wonder how much different the estimate would have been if different cities had been included, and we would expect to see this uncertainty reflected in a wider confidence interval.

A more subtle level of clustering occurs by year. Again, if there is a lot of year-to-year variation in the incidence of HMPV, then if the desire is to generalize to future years (rather than just to estimate what the incidence was in the years studied), clustering by year would need to be accounted for statistically and significant year-to-year variation in incidence would also lead to a wider confidence interval.

Chapter 4 Planning the Measurements: Precision, Accuracy, and Validity

1a. Dichotomous
1b. Continuous
1c. Dichotomous
1d. Discrete numerical
1e. Nominal
1f. Discrete numerical
1g. Ordinal
1h. Continuous
1i. Nominal
1j. Dichotomous

Power is increased by using an outcome variable that contains ordered information. For example, highest year of schooling has more power than college degree or more/less than college degree. Similarly, use of body mass index as a continuous outcome would offer more power (contain more information) for most research questions than presence or absence of obesity. A commonly used intermediate choice is the ordinal variable normal/overweight/obese.

2a. This is a problem with accuracy. It could be due to an observer not visualizing the reading correctly (a second observer could check the result), but more likely the scale needs to be adjusted.

2b. This is a problem with precision. The excessive variability could be an observer error, but more likely the scale needs refurbishing.

2c. This situation can reduce both accuracy and precision. Accuracy will suffer because the observer's hold on the baby will likely alter the observed weight; this might tend to consistently increase the observed weight or to consistently decrease it. This problem with the subjects might be solved by having the mother spend some time calming the baby; an alternative would be to weigh the mother with and without the baby, and take the difference.

2d. This is primarily a problem with precision, because the pointer on the scale will vary around the true weight (if the scale is accurate). The problem is with the subjects and has the same solution as in exercise 2c.

2e. This is mainly a problem with precision, since the babies' weights will vary, depending on whether or not they ate and wet their diapers before the examination. This problem of subject variability could be reduced by giving the mothers instructions not to feed the babies for 3 hours before the examination, and weighing all babies naked.

3a. Predictive validity: The burnout scores predicted an outcome that we might expect to be associated with burnout.

3b. Face validity: Asking people how often they feel burned out seems like a reasonable approach to assessing burnout.

3c. Construct validity: This measure of burnout is responsive to circumstances that we would expect to affect burnout.

3d. Criterion-related validity: These two items agree closely with a well-accepted standard measure.

Chapter 5 Getting Ready to Estimate Sample Size: Hypotheses and Underlying Principles

1. Sample size = the projected number of subjects in a study that are required for the investigator to be able to detect a given effect size (at the specified levels of α and β).

   Null hypothesis = a statement of the research hypothesis that indicates that there is no difference between the groups being compared.

   Alternative hypothesis = a statement of the research hypothesis that indicates that there is a difference between the groups being compared.

   Power = the likelihood of detecting a statistically significant difference between the groups being compared (with a given sample size, at a given level of statistical significance) if the real difference in the population equals the effect size.

   Level of statistical significance = the preset chance of falsely rejecting the null hypothesis.

   Effect size = the minimum size of the difference in the two groups being compared that the investigator wishes to detect.

   Variability = the amount of spread in a measurement, usually calculated as either the standard deviation or the standard error of the mean.

2a. Neither. This is a statistically significant result, and there is nothing to suggest that it represents a type I error.

2b. The sample size was small and very few subjects would have developed lung cancer during the study. These negative results are almost certainly due to a type II error, especially given extensive evidence from other studies that smoking causes lung cancer.

2c. There is no prior epidemiologic or pathophysiologic reason to believe that alcohol use reduces the risk of developing diabetes; this result is likely due to a type I error. The investigator could have been more informative: $P < 0.05$ could be $P = 0.04$ or $P = 0.001$; the latter would reduce (though not rule out) the likelihood of type I error.

Chapter 6 Estimating Sample Size and Power: Applications and Examples

1. H_0: There is no difference in the body mass index of stomach cancer cases and controls.

   H_A (two-sided): There is a difference in the body mass index of stomach cancer cases and controls. Body mass index is a continuous variable and case–control is dichotomous, so a *t* test should be used.

   Effect size = 1 kg/m^2
   Standard deviation = 2.5 kg/m^2
   $E/S = 0.4$

   From Appendix 6A,

   If $\alpha = 0.05$, $\beta = 0.20$, then 100 subjects are needed per group.
   If $\alpha = 0.05$, $\beta = 0.10$, then 133 subjects are needed per group.
   If $\alpha = 0.01$, $\beta = 0.20$, then 148 subjects are needed per group.

   Extra credit: If the investigator only had access to 60 cases, which of the following strategies for increasing power will help the most?

   a. Use a continuous variable—body mass index is already being measured as a continuous variable.
   b. Use a more precise variable—both weight and height are fairly precise variables, so the standard deviation of body mass index is composed mostly of between-individual variation, which cannot be reduced. Careful standardization of height and weight measurements to reduce measurement error would still be a good idea, but this is not the best choice.
   c. Use paired measurements—not applicable; "change" in body mass index is not relevant in this situation.
   d. Use a more common outcome—not applicable.
   e. Use unequal group sizes—the *n* of controls can be increased, as it is easy to find subjects without stomach cancer. For example, if the number of controls can be increased fourfold to 240, one can use the approximation formula on page 69:

   $$n' = ([c + 1] \div 2c) \times n$$

   where n' represents the "new" number of cases, c represents the control-to-case ratio (in this example, 4), and n represents the "old" number of cases (assuming a control per case). In this example,

   $$n' = ([4 + 1] \div 8) \times 100 = (5/8) \times 100 = 63,$$

   which is just about the number of cases that are available. Therefore, a study with 60 cases and 240 controls will have similar power as one with 100 cases and 100 controls.

2. H_0: There is no difference in mean strength between the DHEA-treated and placebo-treated groups.

   H_A: There is a difference in mean strength between the DHEA-treated and placebo-treated groups.

   $\alpha = 0.05$ (two-sided); $\beta = 0.10$
   Test = *t* test
   Effect size = 10% × 20 kg = 2 kg
   Standard deviation = 8 kg

   The standardized effect size (E/S) is 0.25 (2 kg/8 kg). Looking at Appendix 6A, go down the left column to 0.25, then across to the fifth column from the left, where α (two-sided) = 0.05 and $\beta = 0.10$. Approximately 338 subjects per group would be needed. If $\beta = 0.20$, then the sample size is 253 per group.

3. H_0: There is no difference in the mean change in strength between the DHEA-treated and placebo-treated groups.
   H_A: There is a difference in mean change in strength between the DHEA-treated and placebo-treated groups.

   $\alpha = 0.05$ (two-sided); $\beta = 0.10$
   Test = *t* test
   Effect size = 10% × 20 kg = 2 kg
   Standard deviation = 2 kg

   The standardized effect size (*E*/*S*) is 1.0 (2 kg/2 kg). Looking at Appendix 6A, go down the left column to 1.00, then across to the fifth column from the left where α (two-sided) = 0.05 and $\beta = 0.10$. Approximately 23 subjects per group will be needed.
4. H_0: There is no difference in frequency of left-handedness in dyslexic and nondyslexic students.
   H_A: There is a difference in frequency of left-handedness in dyslexic and nondyslexic students.

   $\alpha = 0.05$ (two-sided); $\beta = 0.20$
   Test = chi-squared test (both variables are dichotomous)
   Effect size = odds ratio of 2.0

   Given that the proportion of nondyslexic students who are left-handed (P_2) is about 0.1, the investigator wants to be able to detect a proportion of dyslexic students who are left-handed (P_1) that will yield an odds ratio of 2.0. The sample size estimate will use a chi-squared test, and one needs to use Appendix 6B. However, that appendix is set up for entering the two proportions, not the odds ratio, and all that is known is one of the proportions ($P_2 = 0.1$).

   To calculate the value for P_1 that gives an odds ratio of 2, one can use the formula on page 59:

   $$P_1 = \text{OR} \times P_2 \div ([1 - P_2] + [\text{OR} \times P_2]).$$

   In this example:

   $$P_1 = (2 \times 0.1) \div ([1 - 0.1] + [2 \times 0.1]) = 0.18$$

   So P_1 is 0.18 and P_2 is 0.1. $P_1 - P_2$ is 0.08.

   Table 6B.2 in Appendix 6B reveals a sample size of 318 per group.

   Extra credit: Try this using the formula on page 78; just slog on through, carrying 6 places after the decimal. Then get an instant answer from the calculator on our website, www.epibiostat.ucsf.edu/dcr/.
5. Standard deviation of IQ scores is about one-fourth of the "usual" range (which is 170 – 130 = 40 points), or 10 points.

   Total width of the confidence interval = 6 (3 above and 3 below). Confidence level = 99%.

   Standardized width of the confidence interval = total width/standard deviation:

   $$W/S = 0.6$$

   Using Table 6D, go down the *W*/*S* column to 0.60, then across to the 99% confidence level. About 74 medical students' IQ scores would need to be averaged to obtain a mean score with the specified confidence interval.

Chapter 7 Designing Cross-Sectional and Cohort Studies

1a. Measure serum vitamin B_{12} levels in a cohort of persons more than 70 years of age and without a history of hip fractures, follow them for a period of time (say, 5 years) for the occurrence of hip fractures, and then analyze the association between B_{12} levels and incident hip fractures. (A smaller, albeit less generalizable study could be performed by studying only women, who have a higher rate of hip fractures; an even smaller study would enroll only white women, who have the highest rates of fractures.)

1b. An advantage of the prospective cohort design for studying vitamin B_{12} sufficiency and hip fractures:

- Temporal sequence (i.e., the hip fracture follows the vitamin B_{12} deficiency) helps establish a cause–effect relationship. People who fracture their hips might become vitamin B_{12} deficient after the fracture because they have reduced B_{12} intake, perhaps because of nursing home placement.

A disadvantage of the prospective cohort design:

- A prospective cohort study will require that many subjects be followed for multiple years. The study will therefore be expensive and the findings delayed.

1c. A retrospective cohort study could be done if you find a cohort with stored serum and with reasonably complete follow-up to determine who suffered hip fractures. The main advantage of this design is that it would be less time consuming and expensive. The major drawback is that measurements of vitamin B_{12} might be altered by long-term storage, and that measurements of potential confounders (such as physical activity, cigarette smoking, etc.) may not be available.

2a. Although the PRIDE study is a randomized trial, the report of the baseline examination is an (observational) cross-sectional study. Cross-sectional studies are often the first step in cohort studies or randomized trials.

2b. While it is possible that depression increases urinary incontinence, it seems at least equally plausible that urinary incontinence increases the risk of depression. As we will discuss in Chapter 9, it is also possible that the association is due to bias, e.g., if depressed women were more likely to *report* incontinence episodes even if they did not have more of them, or confounding, if a third factor (e.g., severity of obesity) caused both depression and incontinence.

A longitudinal (cohort) study could help by clarifying the time sequence of the association. For example, depressed and non-depressed women with little or no incontinence at baseline could be followed to see whether the depressed women develop more or worse incontinence over time. Similarly, continent and incontinent women with no history of depression could be followed to determine whether more incontinent women are more likely to become depressed. Finally, and most convincingly, the investigators could study *changes* in depression or incontinence, either naturally occurring or (ideally) as a result of an intervention, and see if changes in one preceded changes in the other. For example, do depressive symptoms improve when incontinence is successfully treated? Does (reported) continence improve when depression lifts?

Chapter 8 Designing Case–Control Studies

1a. The cases might consist of all women between 30 and 75 years of age with ovarian cancer reported to a local tumor registry, and who can be contacted by telephone and agree to participate.

1b. The controls might be a random sample of all women between 30 and 75 years of age from the same counties as in the tumor registry. The random sample might be obtained using random-digit dialing (hence the need to restrict cases to those who have telephones).

1c. Since ovarian cancer requires intensive therapy and can be fatal, some cases may be unwilling to enroll in the study or may have died before they can be interviewed. If a family history of ovarian cancer is related to more aggressive forms of ovarian cancer, then the study might underestimate its relative risk, because those cases with a positive family history would be less likely to survive long enough to be included in the sample of cases. If familial ovarian cancer is more benign than other ovarian cancers, the opposite could occur.
Similarly, it is possible that healthy women who have a family member with ovarian cancer will be more interested in the study and more likely to enroll as a control. In that situation, the prevalence of family history of ovarian cancer in the control group will be artificially high, and the estimate of the risk for ovarian cancer due to family history will be falsely low. This problem might be minimized by not telling the potential control subjects exactly what the research question is or exactly which cancer is being studied, if this can be done in a way that is acceptable to the human subjects committee.

1d. Family history of ovarian cancer is generally measured by asking subjects about how many female relatives they have, and how many of them have had ovarian cancer. Recall bias is a possible problem with this approach. Women with ovarian cancer, who may be concerned about the possibility of a genetic predisposition to their disease, may be more likely to remember or find out about relatives with ovarian cancer than healthy women who have not had reason to think about this possibility. This would cause the estimate of the association between family history and ovarian cancer to be falsely high.

In addition, women may confuse the gynecological cancers (cervical, uterine, and ovarian) and confuse benign gynecological tumors that require surgery with malignant tumors. This may cause misclassification (some women without a family history of ovarian cancer will report having the risk factor and be misclassified). If misclassification occurs equally in the cases and controls, the estimate of the association between family history and ovarian cancer will be falsely low. If this type of misclassification is more common in cases (who may be more likely to misinterpret the type of cancer or the reason for surgery in relatives), then the estimate of the association between family history and ovarian cancer will be falsely high. Misclassification could be decreased by checking pathological records of family members who are reported to have ovarian cancer to verify the diagnosis.

Finally, it would be desirable to take into account the *opportunity* for cases and controls to have a positive family history: women with many older sisters have greater opportunity to have a positive family history than those with only brothers or younger sisters. As discussed in Chapter 9, matching and stratification are two ways of dealing with this possibility.

1e. The simplest approach would be to dichotomize family history of ovarian cancer (e.g., first-degree relatives or not) and use the odds ratio as the measure of association. The odds ratio approximates the relative risk because the outcome (ovarian cancer) is rare. A simple chi-squared test would then be the appropriate test of statistical significance. Alternatively, if family history were quantified (e.g., proportion of first- and second-degree female relatives affected), one could look for a dose–response, computing odds ratios at each level of exposure.

1f. The case–control design is a reasonable way to answer this research question despite the problems of sampling bias, recall bias, and misclassification that were previously noted. The chief alternative would be a large cohort study; however, because ovarian cancer is so rare, a cohort design to answer just this specific question is probably not feasible. A retrospective cohort study, in which data on family history were already systematically collected, would be ideal, if such a cohort could be found.

2a. Cases could be younger drivers (perhaps 16 to 20 years old) who were involved in crashes, and controls could be friends or acquaintances they identify. It would be important to exclude friends with whom they play video games, to avoid overmatching. Random digit

dialing would likely be less successful as a strategy for identifying controls, given the high prevalence of cellular phones (which, unlike land lines, are not geographically localized) in this age group. Cases and controls could also be identified if the investigator had access to the records of an automobile insurance company. An argument could be made that cases and controls should be matched for sex, given that both playing video games and crashing cars are more common in young men. The exposure would be measured using a questionnaire or interview about video game use. It would be important to ask about video games that do not involve driving as well as about those that do, because causal inference would be enhanced if the association were specific, namely, if there was an effect for use of driving/racing games, but not for shooting or other games.

2b. For intermittent exposures hypothesized to have a short-term effect, like use of a video game just before driving, a case–crossover study is an attractive option. As in exercise 2a, cases could be younger drivers who were involved in crashes. In a case–crossover study there are no controls, just control time periods. Thus, case drivers would be asked about use of racing video games just before the trip that included the crash, and also about control time periods when they did not crash. The time period just before the crash is compared in a matched analysis with other time periods to see if racing video game use was more common in the pre-crash period than in other time periods.

Chapter 9 Enhancing Causal Inference in Observational Studies

1. There are five possible explanations for the association between dietary fruit and vegetable intake and CHD:
   a. Chance—The finding that people with CHD eat fewer fruits and vegetables was due to random error. As discussed in Chapter 5, the *P* value allows quantification of the magnitude of the observed difference compared with what might have been expected by chance alone; the 95% confidence interval shows the range of values consistent with the study results. All else being equal, the smaller the *P* value and the further the null value is from the closer end of the confidence interval, the less plausible that chance is as an explanation.
   b. Bias—There was a systematic error (a difference between the research question and the way the study plan was carried out) with regard to the sample, predictor variable, or outcome variable. For example, the sample may be biased if the controls were patients at the same health plan as the cases, but were selected from those attending an annual health maintenance examination, as such patients may be more health conscious (and hence eat more fruits and vegetables) than the entire population at risk for CHD. The measurements of diet could be biased if people who have had a heart attack are more likely to recall poor dietary practices than controls (recall bias), or if unblinded interviewers asked the questions or recorded the answers differently in cases and controls.
   c. Effect–cause—It is possible that having a heart attack changed people's dietary preferences, so that they ate fewer fruits and vegetables than they did before the heart attack. The possibility of effect–cause can often be addressed by designing variables to examine the historical sequence—for example, by asking the cases and controls about their previous diet rather than their current diet.
   d. Confounding—There may be other differences between those who eat more fruits and vegetables and those who eat fewer, and these other differences may be the actual cause of the lower rate of CHD. For example, people who eat more fruits and vegetables may exercise more.

Possible approaches to controlling for confounding by exercise are summarized in the following table:

Method	Plan	Advantages	Disadvantages
Design Phase			
Specification	Enroll only people who report no regular exercise	Simple	Will limit the pool of eligible subjects, making recruitment more difficult. The study may not generalize to people that exercise
Matching	Match each case to a control with similar exercise level	Eliminates the effect of exercise as a predictor of CHD, often with a slight increase in the precision (power) to observe diet as a predictor	Requires extra effort to identify controls to match each case. Will waste cases if there is no control with a similar exercise level. Eliminates the opportunity to study the effect of exercise on CHD
Analysis Phase			
Stratification	For the analysis, group the subjects into three or four exercise strata	Easy, comprehensible, and reversible	Can only reasonably evaluate a few strata and a few confounding variables. Will lose some of the information contained in exercise measured as a continuous variable by switching to a categorical variable, and this may result in incomplete control of confounding
Statistical adjustment (modeling)	Use logistic regression model to control for fitness as well as other potential confounders	Can reversibly control for all the information in fitness as a continuous predictor variable, while simultaneously controlling for other potential confounders such as age, race, and smoking	The statistical model might not fit the data, resulting in incomplete control of confounding and potentially misleading results. For example, the effect of diet or physical fitness may not be the same in smokers and nonsmokers. The important potential confounders must have been measured in advance. Sometimes it is difficult to understand and describe the results of the model, especially when variables are not dichotomous

In addition to these four strategies for controlling confounding in observational studies, there is the ultimate solution: designing a randomized trial.

e. Cause–effect—The fifth possible explanation is that eating fruits and vegetables really does reduce the rate of CHD events. This explanation is made likely partly by a process of exclusion, reaching the judgment that each of the other four explanations is unlikely and partly by seeking other evidence to support the causal hypothesis. Examples of the latter are biologic evidence that there are components of fruits and vegetables (e.g., antioxidants) that protect against atherosclerosis, and ecological studies that find that CHD is much less common in populations that eat more fruits and vegetables.

2. This is an example of conditioning on a shared effect: The study included only infants with fever, which can be caused by both urinary tract infections and ear infections. Because uncircumcised boys were much more likely to have a urinary tract infection, they were more likely to have a cause for their fever other than an ear infection (i.e., they were overrepresented among boys who did not have an ear infection).
3. The association between maternal acetaminophen use and asthma in offspring could be examined in a cohort study, in which mothers were asked about acetaminophen use during

pregnancy and offspring were followed for the development of asthma. Investigators would look for evidence that maternal genotype modifies the effect of maternal acetaminophen exposure on asthma in the children (interaction), with a stronger association between exposure and outcome among those predicted to be most genetically susceptible. In fact, this is what was reported by Shaheen et al (*J Allergy Clin Immunol* 2010;126(6):1141–1148 e7.) in the Avon Longitudinal Study of Parents and Children (ALSPAC).

Chapter 10 Designing a Randomized Blinded Trial

1a. The main advantage of using the biomarker (a continuous variable) as the primary outcome of the trial is a smaller sample size and a shorter duration to determine whether the treatment reduces the level of the marker. The main disadvantage is the uncertainty of whether change in the level of the marker induced by the treatment means that the treatment will reduce the incidence of the clinically far more important outcome, developing dementia.

1b. The clinical diagnosis of dementia is a more meaningful outcome of the trial that could improve clinical practice for prevention of dementia. The disadvantage is that such a trial would be large, long, and expensive.

2a. Participants should be asked at each follow-up visit whether they have experienced diarrhea, nausea, or vomiting. This could be done using a check box format that is easy to code and analyze. To find other unanticipated adverse effects, participants should also be asked at each visit to describe other symptoms, conditions, or medical care (such as hospitalization or new prescription drug) that have occurred since the previous visit. These questions would be asked in an open-ended way, and the responses would subsequently be classified and categorized for data analysis.

2b. Baseline collection of data should include (1) information about how to contact the participant, a close friend, a family member, or a doctor to allow more complete follow-up; (2) characteristics of the enrolled population (such as age, ethnicity/race, and gender) to allow description of the study cohort; (3) risk factors for the outcome (such as hypertension or family history of dementia) that might identify participants with the highest rate of the outcome and could be used to demonstrate that the study groups were comparable at baseline, as well as define subgroups for secondary analyses; and (4) measurement of the outcome (severity of cognitive impairment). Biological specimens should be stored to allow future measurement of factors, such as genotypes of enzymes that metabolize the drug, that could influence the effectiveness of the treatment.

2c. Stratified blocked randomization could guarantee that there would be a very similar number of participants with the Apoε4 genotype in the treatment and the placebo groups. This could be especially important if the effect of the treatment is influenced by the presence of the genotype. On the other hand, this process makes the trial more complicated (assessing Apoε4 genotype before enrollment will delay randomization and raises issues that include how to counsel participants about the results). The risk of a substantial imbalance in a relatively large trial (>200 per arm) is low, so that simple randomization would be a good choice.

Chapter 11 Alternative Clinical Trial Designs and Implementation Issues

1. The main goal of a phase I trial is to determine if the treatment is sufficiently safe and well tolerated to permit additional trials to find the best dose and test its clinical effectiveness. A phase I trial would enroll men with male pattern baldness, and use one or more potential human doses of the treatment (escalating the dose only if the prior dose did not cause side effects) with the main outcome of adverse events, such as the occurrence of rash. There would be no control group.

2a. The value of comparing finasteride to HairStat depends, in large part, on how strong the data are to support using finasteride as the standard-of-care for treatment of male pattern baldness. If these data are not very strong or finasteride is not commonly used in clinical practice, it would be better to compare HairStat to placebo. A placebo-controlled trial will provide clear evidence that HairStat is better than placebo. It may be reasonable to compare HairStat to finasteride if finasteride is considered the standard of care for male pattern baldness and there are good quality randomized trials to document the efficacy of finasteride. In this case, the investigators should first decide if they think HairStat is more effective than finasteride. If so, an active comparison trial would be the best choice to compare HairStat to finasteride. If the investigators think that HairStat is just as good as finasteride, but will be much cheaper, they should consider a non-inferiority trial. In this case, they must take care to use a trial design that is very similar to that used to document the efficacy of finasteride (inclusion criteria, dose, duration of treatment, outcome measures), and must conduct the trial to ensure that there is minimal non-adherence and loss to follow-up. A major drawback of a non-inferiority trial is that the sample size is likely to be much larger than that required for a placebo-controlled trial.

2b. A factorial design that includes a placebo has the advantages of comparing each treatment to a placebo, and (if planned with adequate statistical power) testing whether the combination of treatments is better than either one alone. The disadvantages are the larger size and greater cost and complexity of the trial.

3. Adherence to the visits, protocol, and study medication could be improved by:
   - Employing friendly research staff who are enthusiastic about the study
   - Reminders (by digital messaging, e-mail, telephone, or mail) of upcoming visits and the importance of adherence to treatment
   - Reimbursement for travel, parking, and other expenses related to the study
   - Two screening visits before randomization to identify those participants more likely to miss follow-up visits
   - A run-in period during which participants are asked to use placebo hair gel, and those who are non-adherent are excluded
   - Other potential strategies listed in Table 11.2

4. The main disadvantage of intention-to-treat analysis is that it includes participants who did not comply with the randomized treatment, and who therefore reduce the apparent magnitude of any effect that is observed for the whole randomized group. However, the disadvantages of using as-treated rather than intention-to-treat analysis are even greater. Because participants who do not comply with the intervention usually differ from those who do comply in important but unmeasured ways, as-treated analysis no longer has a true randomized comparison and may incorrectly conclude that the HairStat is effective.

5. The conclusion that HairStat works better in younger men, based on a subgroup analysis, may be wrong because the result may be due to chance. The probability of finding a "significant" effect in a subgroup when there is no significant effect overall increases with the number of subgroups tested; it is not clear how many subgroups were tested to find this "significant" effect. The claim that the treatment is effective in men younger than age 40 implies that the treatment was ineffective—or even had the opposite effect—in older men. This result should also be reported and tested statistically for modification of the effect on hair growth of HairStat due to age. The claim that HairStat is effective in the subgroup of younger men should only be made if the subgroup analyses were specified in advance (ideally based on a biologic basis to suspect that HairState might work better in younger men), there were not a large number of subgroups tested, and the *P* value for effect modification (interaction) between the effect of treatment and age is statistically significant.

Chapter 12 Designing Studies of Medical Tests

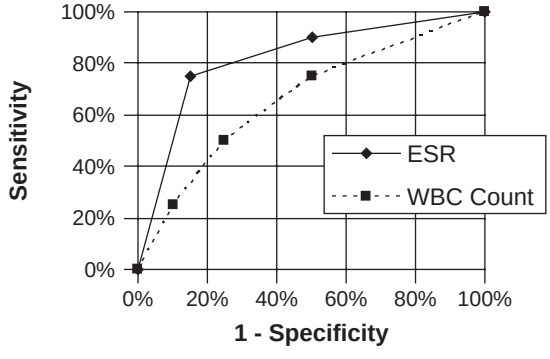
1a. The best way to sample subjects for a diagnostic test is generally to sample patients at risk of a disease, before it is known who has the disease and who does not. In this case, sampling women who present acutely to a clinic or emergency department with abdominal pain consistent with pelvic inflammatory disease (PID) would probably be best. Comparing the erythrocyte sedimentation rates (ESRs) of women hospitalized for PID with those of a healthy control population would be the worst approach, because both the spectrum of disease and especially the spectrum of non-disease are not representative of the groups in whom the test would be used clinically. (Those hospitalized for PID probably have more severe disease than average, and healthy volunteers are much less likely to have high ESRs than women with abdominal pain due to causes other than PID.)

1b. If those assigning the final diagnosis used the ESR to help decide who had PID and who did not, both the sensitivity and specificity might be falsely high. The more those assigning the diagnosis relied on the ESR, the greater the bias (called "**incorporation bias**") in the study.

1c. The best answer is that you should not use any particular cutoff for defining an abnormal result. Rather, you should graphically display the trade-off between sensitivity and specificity using a receiver operating characteristics (ROC) curve and present likelihood ratios for various ESR intervals (e.g., <20, 20 to 49, ≥50 mm/hr) rather than sensitivity and specificity at different cutoffs. This is illustrated by the following table, which can be created from the information in the question:

ESR	PID	No PID	Likelihood Ratio
≥ 50	75%	15%	5.00
20–49	15%	35%	0.43
< 20	10%	50%	0.20
	100%	**100%**	

The ROC curve could also be used to compare the ESR with one or more other tests such as a white blood cell count. This is illustrated in the following hypothetical ROC curve, which suggests that the ESR is superior to the WBC for predicting PID:



2a. This problem illustrates the common error of excluding people from the numerator without excluding them from the denominator. Although it is true that there were only two children with "unexpected" intracranial injuries, the denominator for the yield must be the number of children in whom intracranial injuries would be considered unexpected, i.e., those with normal neurological examinations and mental status. This is probably a

much smaller number than 200. For example, suppose that only 50 of those sent for a CT scan had both a normal mental status and no neurological findings. In that situation, the yield would be 2 of 50, or 4%—nearly 4 times greater.

2b. Unless the finding of an intracranial injury leads to changes in management and there is some way to estimate the effects of these management changes on outcome, it will be very hard to know what yield is sufficient to make the CT scan worth doing. It would be better to use "intracranial injury requiring intervention" as the outcome in this study, although this will require some consensus on what injuries require intervention and some estimate of the effectiveness of these interventions for improving outcome.

2c. The first advantage of studying the effects of the CT scan on clinical decisions is the ability to examine possible benefits of normal results. For example, a normal CT scan might change the management plan from "admit for observation" to "send home." In diagnostic yield studies, normal results are generally assumed to be of little value. Second, as mentioned earlier, abnormal CT scan results might not lead to any changes in management (e.g., if no neurosurgery was required and the patient was going to be admitted anyway). Studying effects of tests on clinical decisions helps to determine how much useful new information they provide, beyond what is already known at the time the test was ordered.

3a. If only children who had a CT scan are included, the study will be susceptible to verification bias (Appendix 12B), in which sensitivity is falsely increased and specificity is falsely decreased, because children without focal neurologic abnormalities (who are either "false negatives" or "true negatives") will be underrepresented in the study.

3b. If children with head injuries who did not have a CT scan are included, and assumed not to have an intracranial injury if they recover without neurosurgery, then the study will be susceptible to differential verification bias ("double gold standard bias"; Appendix 12C), which will tend to increase both sensitivity and specificity if some intracranial injuries resolve without neurosurgery.

Chapter 13 Research Using Existing Data

1. Some possibilities:
   a. Analyze data from the National Health and Nutrition Examination Survey (NHANES). These national population-based studies are conducted periodically and their results are available to any investigator at a nominal cost. They contain data that include variables on self-reported clinical history of gallbladder disease and the results of abdominal sonography.
   b. Analyze Medicare data on frequency of gallbladder surgery in patients more than 65 years of age in the United States, or National Hospital Discharge Survey data on the frequency of such surgery for all ages. Both data sets contain a variable for race. Denominators could come from census data. Like the NHANES, these are very good population-based samples, but have the problem of answering a somewhat different research question (i.e., what are the rates of surgical treatment for gallbladder disease). This may be different from the actual incidence of gallbladder disease due to factors such as access to care.

2a. The main advantages are that using CHS data in a secondary data analysis was quick, easy, and inexpensive—especially compared to the time and expense of planning and conducting a large cohort study. In addition, the research fellow has since developed an ongoing collaboration with the investigators in CHS and has been able to add more sophisticated measures of kidney function to CHS as ancillary studies.

2b. In some cases, the secondary data set does not provide optimal measures of the predictor, outcome, or potential confounding variables. It is important to be sure that the data set will provide reasonable answers to the research question before investing the time and effort required to obtain access to the data. A further drawback is that it can be difficult to obtain data from some studies—the investigator generally needs to write a proposal, find

a collaborator who is a co-investigator on the study, and obtain approval from the study Steering Committee and sponsor.

3. There have been several large randomized controlled trials of the effect of estrogen and selective estrogen receptor modulators on various disease outcomes, including cancer, cardiovascular events, and thromboembolic events. These trials include the Women's Health Initiative randomized trials, the Breast Cancer Prevention trial, the Multiple Outcomes of Raloxifene Evaluation trial, and the Raloxifene Use for the Heart trial. The best place for this investigator to begin would be to determine if estrogen can be measured in stored frozen sera and, if so, determine if any of these large trials have stored sera that could be used for this measurement. The best design for this question is a nested case–control or case–cohort study. The investigator will likely need to write a proposal for this ancillary study, obtain approval from the trial Steering Committee and sponsor, and obtain funding to make the measurements—a relatively inexpensive prospect, since most of the costs of the study have already been covered by the main trial.

Chapter 14 Addressing Ethical Issues

1a. It depends on whether the participants in the original study gave consent for their samples to be used for DNA sequencing, whether they gave consent for the DNA measurements to be used in future studies, and what kinds of future studies were specified. The original consent would not cover the proposed research if the blood samples were collected to be used only to repeat the tests specified in the protocol in case of lost samples or laboratory accidents (such as cholesterol and hemoglobin A1c). Similarly, the original consent would not cover the proposed research if the participants gave consent for the blood specimens to be used for genetic measurement of DNA in future studies of coronary artery disease, but there was no mention of using the specimens in studies of diabetes.

1b. Under federal law, a study can be carried out on existing specimens and data if the new investigator cannot identify the participants, either directly or with the assistance of someone else. Thus, if the new researcher receives samples and data labeled by ID only, and the code that links the samples and the identities of the participants is destroyed or not accessible to the new researcher, additional consent need not be obtained for the secondary study. The ethical justification is that making materials anonymous in this fashion protects participants from breaches of confidentiality, which is the major risk in research with existing materials and data. The presumption is that no one would object to their materials and data being used if there was no risk of breaches of confidentiality. Note, however, that some participants might find it objectionable for someone to sequence their DNA, even if confidentiality is maintained, since the DNA contains information that could lead ultimately to a loss of confidentiality.

1c. When researchers collect new samples in a research project, it is prudent to ask permission to collect and store additional blood to be used in future research studies. Storing samples allows future research to be carried out more efficiently than assembling a new cohort. Tiered consent is recommended: The participant is asked to consent (1) to the specific study (for example, the original cohort study), (2) to other research projects on the same general topic (such as risk of coronary artery disease), or (3) to all other future research that is approved by an IRB and by a scientific review panel. To address the issues raised in exercise 1b, the participant might also be asked to consent specifically to research in which his or her DNA would be sequenced. The participant may agree to one, two, or all options. Of course, it is impossible to describe future research. Hence, consent for future studies is not really informed in the sense that the participant will not know the nature, risks, and benefits of future studies. The participant is being asked to trust that IRBs and scientific review panels will only permit future studies that are scientifically and ethically sound.

2a. Withholding from the control group drugs that are known to be effective would subject them to harm and would therefore be unethical. Even if participants would give informed consent to participate in such a placebo-controlled trial, an IRB would not approve such a study, because it violates the regulatory requirements that the risk/benefit balance be acceptable and that the risks be minimized.

2b. If all participants in the trial were treated with current standard of care chemotherapy, the participants could also be randomized to the new treatment or placebo. Alternatively, the investigators might try to identify a subgroup of patients for whom no therapy has been shown to prolong survival (the most clinically significant endpoint in most cancer treatments). For example, patients whose disease has progressed despite several types of standard chemotherapy and have no options that are proven effective could be asked to participate in a placebo-controlled trial of the experimental intervention. An acceptable control arm could be placebo or best current treatment. This approach assumes that if the drug is active in previously untreated patients it will also be active after other treatments have failed. It is, of course, possible that a drug that does not work in refractory disease may be effective as first-line treatment.

3a. During informed consent, the investigators must discuss: (1) the nature of the study; (2) the number and length of visits; (3) the potential benefits and risks of participation (in this case primarily stigma and discrimination if confidentiality is breached); (4) alternatives to participation in the trial, including HIV prevention measures that are available outside the trial; (5) the voluntary nature of participation and the right to withdraw at any time; (6) protection of confidentiality consistent with state public health reporting requirements.

3b. Investigators need to present information in a manner that participants can understand. Participants with low health literacy will not be able to comprehend a detailed written consent form. It would be useful for the researchers to consult with community and advocacy groups on how to present the information. Suggestions might include videotapes, DVDs, and comic books. Extensive pretesting should be carried out. Furthermore, researchers should determine what misunderstandings about the study are common and revise the consent process to address them.

3c. Even though the study is an observational study, researchers have an ethical obligation to provide information to participants about how to reduce their risk for HIV infection. There are both ethical and scientific reasons for doing so. Researchers have an ethical obligation to prevent harm to participants in their study. They may not withhold feasible public health measures that are known to prevent the potentially fatal illness that is the endpoint of the study. Such measures would include counseling, condoms, and referral to substance abuse treatment and needle exchange programs. Researchers must also invoke these measures to prevent harm to participants in the subsequent vaccine trial, even though the power of the trial will be reduced.

Chapter 15 Designing Questionnaires, Interviews, and Online Surveys

1a. There is no definition of how big a “drink” is.

1b. There is no way to respond if the subject is drinking more than 8 drinks per day.

1c. The question does not specify time—weekdays versus weekend, every day versus less than daily.

1d. It may be better to specify a particular time frame (e.g., in the past 7 days).

2a. Which of the following statements best describes how <u>often</u> you drank alcoholic beverages during the past year? An alcoholic beverage includes wine, liquor, or mixed drinks. Select one of the 8 categories.

◯ Every day
◯ 5–6 days per week
◯ 3–4 days per week
◯ 1–2 days per week
◯ 2–3 times per month
◯ About once a month
◯ Less than 12 times a year
◯ Rarely or not at all

2b. During the past year, how many drinks did you usually have on a typical day when you drank alcohol? A drink is about 12 oz. of beer, 5 oz. of wine, or $1\frac{1}{2}$ oz. of hard liquor. _____ drinks

2c. During the past year, what is the largest number of alcoholic drinks you can recall drinking during one day? _____ drinks

2d. About how old were you when you first started drinking alcoholic beverages?
_____ years old (If you have never consumed alcoholic beverages, write in "never")

2e. Was there ever a period when you drank quite a bit more than you do now?

◯ Yes →
◯ No
↓

If Yes, which of the following statements best describes how often you drank during that period? Select one of the 8 categories

e(i).
◯ Every day
◯ 5–6 days per week
◯ 3–4 days per week
◯ 1–2 days per week
◯ 2–3 times per month
◯ About once a month
◯ Less than 12 times a year
◯ Rarely or not at all

e(ii). During that period, how many drinks did you usually have on a typical day when you drank alcohol? _____ drinks

e(iii). For about how many years did you drink more than you do now? _____years

2f. Have you ever had what might be considered a drinking problem?
◯ Yes
◯ No

3a. Obtaining data through interviews requires more staff training and time than a self-administered questionnaire and is therefore much more expensive.

3b. Some subjects do not like to tell another person the answer to sensitive questions in the area of sexual behavior.

3c. Unless the interviewers are well trained and the interviews are standardized, the information obtained may vary.

3d. However, interviewers can repeat and probe in a way that improves comprehension and produces more accurate and complete responses in some situations than a self-administered questionnaire.

Chapter 16 Data Management

1.

SubjectID	EverSmoked-100Cigs	AgeFirstCig	AvgCigs-PerDay	PastWeek-CigsAny	PastWeek-CigsPerDay	Age-Stopped-Smoking
1001	1	17	30	1	30	
1002	1	21	20	0		45
1003	0			0		

This is how the data might look in a spreadsheet program such as Excel. There are many acceptable possibilities for the field names (column headings). These field names use IntraCaps (capitals in the middle of the word to separate its parts). Database designers are about equally divided between those who like IntraCaps and those who don't.

2a.

SubjectID	KitNumber	AdminDate	AdminTime	SzStopPreHosp	SzStopPreHospTime	HospArrTime	HospArrSzAct	HospArrGCSV
189	A322	3/12/1994	17:39	0		17:48	1	
410	B536	12/1/1998	01:35	1	01:39	0.1:53	0	4

2b.

Field Name	Data Type	Description	Validation Rule
SubjectID	Integer	Unique subject identifier	
KitNumber	Text(5)	Five-character Investigational Pharmacy Code	
AdminDate	Date	Date study drug administered	
AdminTime	Time	Time study drug administered	
SzStopPreHosp	Yes/no	Did seizure stop during prehospital course?	
SzStopPreHospTime	Time	Time seizure stopped during prehospital course (blank if seizure did not stop)	
HospArrTime	Time	Hospital arrival time	
HospArrSzAct	Yes/no	Was there continued seizure activity on hospital arrival?	Check against SzStopPreHosp
HospArrGCSV	Integer	Verbal GCS on hospital arrival (blank if seizure continued)	Between 1 and 5

2c. **Advantages of an on-screen form:**

- No need for transcription from paper forms into the computer data tables
- Immediate feedback on invalid entries
- Programmed skip logic (if seizure stopped during prehospital course, computer form prompts for time seizure stopped; otherwise, this field is disabled and skipped)
- Can be made available via a Web browser at multiple sites simultaneously

Disadvantages of an on-screen form:

- Hardware requirement—a computer workstation
- Some user training required

Advantages of a paper form:

- Ease and speed of use
- Portability
- Ability to enter unanticipated information or unstructured data (notes in the margin, responses that were not otherwise considered, etc.)
- Hardware requirement—a pen
- User training received by all data entry personnel in elementary school

Disadvantages of a paper form:

- Requires subsequent transcription into the computer database
- No interactive feedback or automated skip logic
- Data viewing and entry limited to one person in one place

Although data entry via on-screen data collection forms has many advantages and we recommend it for most research studies, in this study it is impractical. The simplest, fastest, and most user-friendly way to capture data on a nonvolatile medium is still to use a pen and paper.

3. When coded with 0 for *no* or *absent* and 1 for *yes* or *present*, the average value of a dichotomous (yes/no) variable is interpretable as the proportion with the attribute. Of those randomized to lorazepam, 40.9% (27 of 66) were still seizing on hospital arrival; of those randomized to diazepam, 57.4% (39 of 68) were still seizing; and of those randomized to placebo, 78.9% (56 of 71) were still seizing.

Chapter 17 Implementing the Study and Quality Control

1a. Not enough! But here are some steps he can take:

- Identify all missing and out-of-range values and recheck the paper forms to make sure that the data were entered correctly.
- Retrieve missing data from charts.
- Collect missing interview data from surviving participants (but this will not help for those who died or for those whose responses might have changed over time).
- Make a special effort to find subjects who had been lost to follow-up, and at least get a telephone interview with them.
- Obtain vital status, using the National Death Index or a company that helps find people.

1b.

- Collect fewer data.
- Check forms on site immediately after collecting the data to be certain that all items are complete and accurate.
- Use interactive data entry with built-in checks for missing, out-of-range and illogical values.
- Review the database shortly after data entry, so that missing data can be collected before the participant leaves the hospital (or dies).
- Periodically tabulate the distributions of values for all items during the course of the study to identify missing values, out-of-range values, and potential errors.
- Hold periodic team meetings to review progress and emphasize the importance of complete data.

Chapter 18 Community and International Studies

1a. The G.I. clinic

- **Advantages.** This is likely to be a convenient and accessible source of patients. The clinic staff probably has experience participating in research. Implementing a standard battery of diagnostic tests for patients with abdominal pain should not be difficult.
- **Disadvantages.** Patients in this clinic might be a highly selected subset of all patients in the community with abdominal pain, and the clinical course of these patients may differ from others in the community. The results may therefore have limited generalizability.

1b. Community clinics

- **Advantages.** Here you can identify patients at first presentation without the selection and delay caused by the referral process. Community physicians may benefit from the opportunity to participate in research.

- **Disadvantages.** These are mainly logistic. Identifying participating physicians and patients and implementing a standard research protocol will be a major organizational task, and quality control will be a challenge.

2a. This can only be answered with local data. Research elsewhere will not help.

2b. This is well known from the international literature. Repeating such research in China is unlikely to be an efficient use of resources.

2c. For this question, the generalizability of research from elsewhere is likely to be intermediate. Strategies for smoking cessation that have proven successful in other countries may serve as a basis for strategies to be tried in China, but one cannot be sure they will have the same success in China without local research. Previous studies in populations elsewhere with cultural ties to China, such as recent Chinese immigrants to the United States, may be helpful.

Chapter 19 Writing a Proposal for Funding Research

1–3. We hope you came up with some useful ideas for planning your own research agenda, and we encourage you to involve your mentors and peers in discussions of how best to move forward.

Glossary

Accessible population. The group of people to whom the investigator has access and who could be selected for, or approached about participating in, the study. For example, the accessible population for the study consisted of women with breast cancer who were treated within 6 weeks of their original diagnosis at Longview Hospital from January 1, 2013, through June 30, 2014. See also *intended sample* and *target population*.

Accuracy. The degree to which a measurement corresponds to its true value. For example, self-reported bodyweight is a less accurate measurement of actual bodyweight than one made with a calibrated electronic scale.

Adjustment. A general name for various statistical techniques used to account for the effects of one or more variables on an association between two other variables. For example, adjustment for income reduced the magnitude of the association between education and mortality.

Alpha. When designing a study, the preset maximum probability of committing a type I error, that is, rejecting the null hypothesis when it is true. For example, by choosing an alpha of 0.05, the investigator set a maximum probability of 5% that her study would find a statistically significant association between non-white race and the risk of colon cancer by chance alone. Also called the level of statistical significance.

Alternative hypothesis. The proposition, used in estimating sample size, that there is an association between the predictor and outcome variables in the population. For example, the study's alternative hypothesis was that teenagers who smoke cigarettes have a different likelihood of dropping out of school than those who do not smoke. See also *null hypothesis*.

Analytic study. A study that looks for associations between two or more variables. For example, the investigator did an analytic study of whether height was correlated with blood pressure in medical students. See also *descriptive study*.

Association. A quantifiable relationship between two variables. For example, the study found an association between male sex and risk of cognitive impairment among 60- to 69-year-olds, with a risk ratio of 1.6.

Before-after study. A study that compares the attributes of subjects before and then again after an intervention. For example, the study compared mean serum cholesterol levels before and after institution of a low-fat diet.

Beta. When designing a study, the preset maximum probability of committing a type II error, that is, failing to reject the null hypothesis when it is false. This measure is only meaningful in the context of an effect size. For example, if an investigator specifies a beta of 0.20 (and alpha of 0.05), she would need about 25,000 subjects per group followed for 10 years to show that daily aspirin halves the risk of colon cancer. Put another way, if aspirin actually had exactly that effect, her study of 25,000 per group would have a 20% chance of failing to reject the null hypothesis of no difference (at alpha = 0.05). See also *power*.

Between-groups design. A study design that compares the characteristics or outcomes of subjects in two (or more) different groups. For example, the investigator used a between-groups design to compare in-hospital mortality rates among patients treated in intensive care units that had round-the-clock intensivists with those among patients treated in units that used electronic monitoring of patients. See also *within-group design*.

Bias. A systematic error in a measurement, or in an estimated association, due to a shortcoming in a study's design, execution, or analysis. For example, due to a bias in the way subjects remembered their

exposure to toxic chemicals, patients with leukemia were more likely to report use of insecticides than were controls.

Blinding. The process of ensuring that subjects and/or investigators are unaware of the group (e.g., intervention or control) to which subjects are assigned, usually in the context of a randomized trial. Also called masking, especially in ophthalmologic studies. For example, by using identical placebo pills and keeping the list of subject assignments off-site, both the subjects and the investigators (including research assistants) were blinded to which subjects were treated with the active medication.

Blocked randomization. A method of assigning subjects to a particular intervention in blocks (groups) of a pre-specified size (e.g., four or six) to ensure that similar numbers of subjects are assigned to the intervention and control groups. Often used in multi-center studies in which the investigators want the total numbers of intervention and control subjects to be similar at each site. For example, patients within each clinic were randomly assigned to either the treatment or control groups in blocks of six, ensuring that the number of subjects per group would differ by no more than 3. See also *stratified blocked randomization*.

Bonferroni correction. A technique to prevent type I errors by dividing the overall alpha in a study by the number of hypotheses tested. For example, because the investigators were testing four different hypotheses, they used Bonferroni correction to reduce alpha for each hypothesis from 0.05 to 0.0125.

Calibration. The process of ensuring that an instrument gives a consistent reading; usually done by measuring a known standard and then adjusting (calibrating) the instrument accordingly. For example, the scale was calibrated monthly by weighing a 50-kg block of steel.

Case. A subject who has, or who develops, the outcome of interest. For example, cases were defined as those who had unstable angina, myocardial infarction, or sudden death during follow-up. See also *control*.

Case–cohort study. A research design in which subjects who develop a disease (or other outcome) are selected as cases during follow-up of a larger cohort, and then compared with a random sample of the overall cohort. For example, a case–cohort study enrolled a cohort of 2,000 men with early prostate cancer, and compared levels of androgens and vitamin D from samples obtained at baseline among those who died during follow-up with levels in a random sample of the entire cohort.

Case–control study. A research design in which cases who have a disease (or other outcome) are compared with controls who do not. For example, a case–control study compared average weekly consumption of nuts and seeds among cases of diverticulitis seen in an emergency room with nut and seed consumption of controls who had other gastrointestinal diagnoses.

Case–crossover study. A variant of the case–control design, in which each case serves as his own control, and the value of a specific time-dependent exposure in the period before the outcome occurred is compared with its value during one or more control periods of time. This design is susceptible to recall bias and is therefore most useful when an exposure can be ascertained objectively. For example, a case–crossover design was used to determine whether patients who presented to an emergency room with a migraine headache were more likely to have eaten chocolate within the previous 2 hours than during a similar time one day previously.

Categorical variable. A variable that can have only several possible values. For example, the investigator transformed her measurements of reported educational level into a categorical variable with four values: less than high school, high school or some college, college degree, or post-college degree. See also *continuous variable*, *dichotomous variable, nominal variable,* and *ordinal variable.*

Cause–effect. The concept that a predictor is responsible for producing an outcome—or increasing the likelihood of an outcome's occurrence. The purpose of most observational studies is to demonstrate cause–effect, though this is difficult to do unless the cause (e.g., a treatment) is assigned randomly. For example, the investigator performed a case–control study to determine whether there was a cause–effect relation between drinking alcohol (the cause) and pancreatic cancer (the effect). See also *confounding* and *effect–cause*.

Chi-squared test. A statistical technique that compares two (or more) proportions to determine if they are significantly different from one another. For example, a study determined whether the risk of dementia was similar among people who exercised at least twice a week as compared with those who exercised less frequently by comparing those risks statistically with a chi-squared test.

Classification and Regression Trees (CART). See *recursive partitioning*.

Clinical prediction rule. An algorithm that combines several predictors, including the presence or absence of various signs and symptoms and the results of medical tests, to estimate the probability of a particular disease or outcome. For example, the investigators developed a clinical prediction rule for the diagnosis of wrist fracture among postmenopausal women based on information about prior fractures, the characteristics of the fall (if any), physical examination of the forearm, and current medications.

Clinical trial. A research design in which subjects receive one of (at least) two different interventions. Usually, the interventions are is assigned randomly; thus, the term *randomized clinical trial*. Clinical trials are sometimes called experiments. For example, the investigator performed a clinical trial to determine whether prophylactic treatment with penicillin reduced the risk of bacterial endocarditis among patients with abnormal heart valves who were undergoing dental procedures.

Clinic-based control. In the context of a case–control study, the selection of control patients from the same clinics (or practices) from which the cases were chosen. For example, the investigator used clinic-based controls in her study of whether running on pavement for at least two miles per week was associated with radiographic osteoarthritis of the knee.

Cluster randomization. A technique in which groups of participants, known as clusters, are randomly assigned to different treatments, rather than having each participant assigned randomly as an individual. For example, in a study of the effects of noise reduction on recovery from cardiac surgery, the investigator used cluster randomization to assign intensive care units in 40 different hospitals to either a "post-operative quiet" intervention or a "usual care" control.

Cluster sampling. A sampling technique in which subjects are selected in groups (clusters) rather than as individuals. Most often used for convenience when sampling large populations. For example, an investigator interested in determining the prevalence of drug use used cluster sampling to enroll 300 patients. First, she identified potential subjects by choosing 10 three-digit prefixes (e.g., 285-, 336-, etc.) within an area code; then she used random-digit dialing to find 30 willing subjects within each three-digit cluster.

Coefficient of variation (CV). A measure of the precision of a measurement, obtained by dividing the standard deviation of a series of measurements performed on a single sample by the mean of those measurements. Sometimes, the CV is obtained for values at the middle and the extremes of the measurement. For example, the lab determined that its coefficient of variation for serum estradiol levels was 10% in a sample from a peri-menopausal woman (in whom the estradiol level was very low), but only 2% in a younger woman.

Cohort study. A prospective cohort study involves enrolling a group of subjects (the cohort), performing some baseline measurements, and then following them forward in time to observe outcomes; a retrospective cohort study involves identifying a group of subjects (the cohort) in whom the measurements have already been made, and in whom some or all of the follow-up has already occurred. For example, an investigator did a retrospective cohort study of whether the results of an emotional intelligence test administered when soldiers enlisted in the U.S. Army was associated with the subsequent likelihood of developing post-traumatic stress disorder (PTSD).

Co-intervention. In a clinical trial, an intervention that occurs after randomization, other than the intervention being studied, that affects the likelihood of an outcome. Co-interventions that occur at different rates in the study groups can bias the outcome and make it difficult to ascribe causality to the intervention being studied. For example, a study of the effect of a breastfeeding promotion intervention on subsequent allergic disease in infants was hard to interpret because the women in the intervention group not only breastfed longer, but were also more likely than the control group to delay the introduction of solid foods and to purchase hypo-allergenic formula, both of which were potential co-interventions.

Complex hypothesis. A research hypothesis that has more than one predictor or outcome variable. Complex hypotheses should be avoided, because they are difficult to test statistically. For example, the investigators reformulated their complex hypothesis ("That a new program in case management would affect both length of stay and the likelihood of readmission") into two simple hypotheses ("That a new program in case management would affect length of stay" and also "That a new program in case management would affect the likelihood of readmission"). See also *simple hypothesis*.

Concordance. A measure of agreement between two (or more) observers about the occurrence of a phenomenon. For example, the concordance between radiologists A and B was 96% for the presence of a lobar pulmonary infiltrate, but only 76% for cardiomegaly. See also *kappa*.

Conditioning. The process of examining the associations between two or more other variables at fixed levels of a "conditioned-on" variable. Specification, matching, stratification, and multivariate adjustment are the most common ways of conditioning on a variable. For example, the investigator found no association between cocaine use and the risk of syphilis after conditioning on the number of sexual partners.

Conditioning on a shared effect. A source of bias in epidemiologic studies in which an association is introduced between two different causes of the same effect by conditioning on that effect. For example, due to conditioning on a shared effect (total screen time), there is an inverse association between television and gaming among children with at least 6 hours per day of screen time: Those who spend more time watching television spend less time playing videogames.

Confidence interval. A term that is often misunderstood, a confidence interval is best thought of as a measure of precision: the narrower a confidence interval, the more precise the estimate. Confidence intervals are closely related to statistical significance: A $(1 - \alpha)$% confidence interval (approximately) includes the range of values that were not statistically significantly different (at significance level α) from what was observed. Confidence intervals are often erroneously interpreted as direct statements about *posterior probability* (e.g., that there is a 95% probability that the true value is contained within the 95% confidence interval). This is incorrect because posterior probability depends on other information besides what was found in the study. For example, a relative risk of 1.6, with a 95% confidence interval from 0.9 to 2.8, would not be statistically significant at an alpha of 0.05, because the interval includes "no effect" (a relative risk of 1.0). See also *alpha* and *P value*.

Confounder. See *confounding*.

Confounding. An epidemiologic phenomenon in which an association between a predictor and an outcome is due to a third variable (called the confounder or the confounding variable), rather than being a cause–effect relation between the predictor and the outcome. For example, the apparent association between cigarette smoking and cervical cancer was confounded by human papilloma virus (HPV) infection, because women who smoked were also more likely to have (multiple sexual partners and) HPV infection. See also *effect modification*.

Confounding by indication. A specific form of confounding in which one of the indications for a treatment is the confounder; usually occurs in observational studies of the association between a treatment and an outcome. For example, the reviewers of an observational study were concerned that the reported association between a new treatment for bipolar disorder and increased suicide risk might have occurred because patients with more severe underlying disease had been selectively treated with the new medication.

Confounding variable. See *confounding*.

Consecutive sample. A study sample in which the subjects are chosen one after another until the sample size is achieved. Usually used to refer to the intended sample; it may also refer to the actual sample when performing medical records reviews, since informed consent may not be required. For example, the investigators performed consecutive sampling to review the charts of the first 100 patients with rheumatoid arthritis seen in the rheumatology clinic, beginning January 15, 2013.

Construct validity. A term that describes how well a measurement corresponds to the theoretical definitions of the trait (the "construct") that is being measured. For example, a measurement of social anxiety was thought to have construct validity, because there were substantial differences in its values among people whose friends described them as "fun-loving" and "extroverted" as compared with those who were described as "shy" and "unlikely to go to parties." See also *content validity* and *criterion-related validity*.

Contamination. The undesirable process by which some or most of the effects of an intervention also affect subjects in the control group. For example, a study of the effects of whether teaching children to count backwards improved their overall arithmetic skills was plagued by contamination, because the children in the intervention group couldn't resist teaching that skill to their friends in the control group.

Content validity. A term that describes how well a measurement represents several aspects of the phenomenon being studied. For example, a measurement of insomnia was thought to have content validity, because it measured total amount of sleep, episodes of nighttime awakening, early morning awakening, energy on arising for the day, and daytime sleepiness. See also *construct validity* and *criterion-related validity*.

Continuous variable. A measurement that, in theory, can have an infinite number of possible values. In practice, the term is often used for measurements that have "many" (some say 10 or more, others say 20 or more) possible values. For example, systolic blood pressure was measured as a continuous variable in mm Hg using a mercury sphygmomanometer. See also *categorical variable, dichotomous variable,* and *discrete variable*.

Control. A term that has two distinct meanings. First, control refers to a subject who does not have the outcome of interest, and is therefore a member of a comparison group to which those with the outcome (the "cases") are compared. For example, for a study of risk factors for peptic ulcer disease, controls were selected from patients hospitalized during the study period with a non-gastrointestinal diagnosis. Second, control refers to the inactive "treatment" (i.e., a placebo or "usual care") received by participants in a clinical trial who did not receive the study intervention; in that context, control is also used to refer to a participant who received the inactive treatment. For example, the controls were given placebo tablets that looked identical to the active drug. See also *case* and *intervention*.

Convenience sample. A group of subjects who were selected for a study simply because they were relatively easy to access. For example, the investigator used a convenience sample of patients from her clinic to serve as controls for her case–control study of risk factors for meningioma.

Correlation coefficient. A statistical term that indicates the degree to which two continuous measurements are related linearly, such that a change in one measure is associated with a proportional change in the other. Often abbreviated as r. For example, height and weight were correlated in a sample of middle-aged women with $r = 0.7$.

Cox model. Also called *Cox proportional hazards model*. A multivariable statistical technique that measures the individual effects of one or more predictor variables on the rate (hazard) at which an outcome occurs in a sample, accounting for differing lengths of follow-up among subjects. For example, using a Cox proportional hazards model, men were about twice as likely as women, and blacks about three times as likely as whites, to develop strokes, adjusting for age, blood pressure, and diabetes, as well as length of follow-up. See also *logistic regression model*.

Criterion-related validity. A term that describes how well a measurement correlated with other ways of measuring the same phenomenon. For example, a measurement of depression in adolescents was thought to have criteria-related validity, because it had a high correlation with scores on the Beck depression inventory. See also *construct validity* and *content validity*.

Crossover. A term used to describe a subject, usually in a clinical trial, who starts out in one group (say, usual care) and switches to the other group (say, the active treatment) during the study. Most commonly occurs when the active treatment involves a procedure. For example, 15 subjects with prostate cancer who were initially assigned to watchful waiting crossed-over to receive radiation therapy or surgery during the trial.

Crossover study. A research design in which all the subjects from one treatment (or control) group are switched to the other group, usually at the midway point of the study. Sometimes, there is a washout period between the two phases. This design, which enables all subjects to receive the active treatment, is only useful for conditions that return to baseline after treatment. For example, patients with migraine headaches were involved in a crossover study comparing a new drug with a placebo for the prevention of migraines.

Cross-sectional study. A research design in which subjects are selected and measurements made within a limited period of time, usually to estimate the prevalence of an exposure or a disease. For example, the prevalence of myopia was estimated in a cross-sectional study of 1,200 college students in Berkeley, California.

Cumulative incidence. See *incidence*.

Data. A plural noun used to describe measurements, usually in numeric format. (The singular of data is datum.) For example, data concerning the prevalence of various diseases are useful when making decisions about allocation of health care resources.

Data dictionary. A table or spreadsheet that includes information about each of the variables in a study, including its name and type (e.g., numeric or string), the definition of each value, and the allowed range of values. For example, the investigator consulted the data dictionary because she had forgotten that a "5" in the field named "race" was used to indicate American Indian/Alaska Native.

Data table. A table of study data in which each row corresponds to a unique record and each column corresponds to a field or attribute. All studies will have a table of study subjects in which each row corresponds to an individual participant and the columns correspond to participant-specific information, such as sex and date of birth. Most studies will use additional tables in which the rows correspond to study visits, laboratory results, telephone contacts, etc.

Dependent variable. See *outcome variable*.

Descriptive study. A study that does not look for associations, test hypotheses, or make comparisons. For example, the investigator performed a descriptive study of the prevalence of obesity among preschool children. See also *analytic study*.

Diagnostic test study. A study that looks at whether the results of a medical procedure are useful in assessing the likelihood of a particular diagnosis in a patient. For example, a diagnostic test study was designed to determine whether serum bicarbonate levels were useful in the diagnosis of sepsis among patients with fever.

Dichotomous variable. A variable that can have only one of two values, such as yes/no or male/female. For example, the examiner dichotomized systolic blood pressure into hypertensive ($\geq$ 140 mm Hg) or not. See also *categorical variable* and *continuous variable*.

Differential bias. A general term for the situation in which a measurement varies systematically by the status of the subject, usually by whether or not the subject is a case or a control; it most commonly occurs with recalled exposures. For example, because cases of adult celiac disease were more likely to recall childhood exposures to wheat-containing products as children than their siblings who had grown up in the same household, the investigators suspected that there was differential recall bias. See also *non-differential bias*.

Differential verification bias. A bias that occurs in studies of diagnostic tests when different gold standards are applied to different subjects, depending at least in part on the result of the test being studied. For example, in a study of prostate specific antigen (PSA) screening for prostate cancer in men, those with high PSA levels received prostate biopsies, while those with normal PSA levels were followed clinically; this raised the concern that differential verification bias falsely increased the sensitivity, and decreased the specificity, of PSA screening in men with indolent prostate cancer.

Discrete variable. A type of variable that takes on only integer values. For practical purposes, continuous variables are sometimes treated as discrete variables. For example, age is usually expressed as age in years at last birthday, and current smoking as average number of cigarettes smoked per day. See also *continuous variable*.

Dose–response. The phenomenon by which the greater the exposure (dose), the greater the magnitude or likelihood of the outcome (response). (If an exposure is protective, then the greater the exposure, the lower the likelihood of the outcome.) For example, one study reported a dose–response relation between sun exposure and numbers of melanocytic nevi; another reported a dose–response relation between numbers of nevi and risk of melanoma.

Double-cohort study. A study design in which subjects are enrolled into one of two distinct cohorts, often by occupation. For example, a double-cohort study was used to compare the risks of contact dermatitis of the hands, as well as fungal infections of the feet, among potters versus dancers.

Double gold standard bias. See *differential verification bias*.

Dropout. A study subject in whom outcome status cannot be ascertained, often because she refused follow-up. Sometimes this includes subjects who drop out because they died during the study. For example, there were 17 dropouts in a study: 8 due to refusal, 6 due to death, and 3 because of the development of dementia.

Effect–cause. The situation in which an outcome causes the predictor, rather than vice-versa. For example, although a case–control study observed that exposure to inhaled bronchodilators was associated with an increased risk of interstitial lung disease, the most likely explanation was effect–cause, namely that patients with interstitial lung disease were more likely to have been treated (erroneously) with inhalers. See also *cause–effect*.

Effectiveness. Although there is no standard definition of this term, we use it to refer to a measure of how well an intervention works in actual practice, as opposed to how well it worked in a randomized trial. For example, because clinical trials have found that tissue plasminogen activator (tPA) reduces morbidity and mortality from stroke in several trials performed in urban settings, the investigators studied its effectiveness in 25 rural emergency rooms. See also *efficacy*.

Effect modification. The condition in which the strength of the association between a predictor and an outcome is affected by a third variable (often called the effect modifier, though it can be difficult to determine which is the predictor and which the effect modifier). For example, the investigators found that the effects of income on stroke risk differed in whites and blacks, such that poverty had stronger association with stroke in blacks than in whites. See also *confounding*.

Effect size. In the context of sample size planning, a measure of how big a difference the investigator wishes to detect between the groups that will be compared, or of the size of the association. More generally, the actual size of that difference or association after the study is completed. For example, the investigators based their sample size estimates on an effect size of a 20 mg/dL difference in mean blood glucose levels in the two groups.

Efficacy. Although there is no standard definition of this term, we use it to refer to a measure of how well an intervention worked in a clinical trial, as opposed to how well it would work in actual practice. For example, a clinical trial reported that tissue plasminogen activator (tPA) had an efficacy of 25% in reducing morbidity and mortality among patients with acute stroke. See also *effectiveness*.

Entry criteria. A list of the attributes that subjects must have to be eligible to participate in a study. The entry criteria may vary if subjects are enrolled in different groups, such as in case–control or double–cohort studies. For example, the entry criteria for a study of a new treatment for gout included age between 20 and 75 years, at least one episode of physician-diagnosed gout in the previous 12 months, and a serum uric acid level of at least 6 mg/dL. See also *exclusion criteria* and *inclusion criteria*.

Epidemiologist. A physician, broken down by age and sex. For example, one of the authors (but we are not saying which one!).

Epidemiology. The science of determining the frequency and determinants of diseases or other health outcomes in populations. For example, a study investigated the epidemiology of handgun violence in inner cities.

Equipoise. The situation in which it is not known which of two possibilities (drug X is better than placebo; drug X is worse than placebo) is more likely to be true. Thus, it is ethical to compare drug X and placebo in a randomized trial. For example, the investigators believed that there was clinical equipoise in a trial, since it was not known whether a proposed new treatment for esophageal cancer would result in better outcomes than the current standard of care.

Equivalence study. A study whose purpose is to show that two (or more) treatments have similar outcomes; usually, one of the treatments is new, and the other is known to be effective. For example, an equivalence study design was used to compare two antibiotics (new drug A with old drug B) for the treatment of pneumonia.

Exclusion criteria. A list of attributes that prevent a potential subject from being eligible for a study. For example, the exclusion criteria for the study were prior treatment with an antidepressant medication in the previous two years, current use of alpha-blockers or beta-blockers, and an inability to read English at the sixth-grade level. See also *inclusion criteria*.

Experiment. In clinical research, a study in which subjects are assigned randomly to one (or more) treatment or comparison groups. It is also called a randomized trial. For example, the investigators performed an experiment to determine whether drug X was better than placebo in the treatment of fibromyalgia.

Exposure. A term used to indicate that a study subject has a particular risk factor. For example, exposure to aspirin was defined as taking an average of one or more aspirin tablets (of any size) a week during the previous 6-month period.

Face validity. A term that describes how well a measurement appears to measure a particular phenomenon, based on whether it seems reasonable; it is generally not a very reliable method for assessing validity. For example, a measurement of popularity in adolescents was regarded as having face validity, because the investigators thought that it differentiated the popular students in their high schools from those who were not. See also *construct validity, content validity,* and *criterion-related validity*.

Factorial trial. A clinical trial of two or more treatments (e.g., A and B), sometimes with two unrelated outcomes, in which subjects are assigned randomly to receive active treatment A and placebo B, active treatment B and placebo A, both active treatments A and B, or both placebos A and B. For example, the investigator performed a factorial trial to determine whether long-term use of beta carotene and aspirin affected the risk of gastrointestinal cancer.

False-negative result. A term that can be used in two different ways. In the context of a medical test, it refers to a test result that is falsely negative in a patient with the condition being tested for. For example, though the patient had biopsy-proven breast cancer, her mammogram had given a false-negative result. In the context of a research study, it refers to a study result that fails to detect an effect in the sample (i.e., the study result was not statistically significant) that is present in the population. For example, though subsequent studies showed that cigarette smoking increases the risk of stroke, an early case–control study had a false-negative result ($P = 0.23$).

False-positive result. A term that can be used in two different ways. In the context of a medical test, it refers to a test result that is falsely positive in a patient without the condition being tested for. For example, though the patient did not have breast cancer or develop it during 6 years of follow-up, her mammogram had a false-positive result. In the context of a research study, it refers to a study result that detects an effect in the sample (i.e., the study result was statistically significant) that is not present in the population. For example, though subsequent studies showed that cigarette smoking does not increase the risk of Parkinson's disease, an early case–control study had a false-positive result ($P = 0.03$).

Field. A column in a relational database table that includes data on a specific attribute of the record. For example, two of the fields in the Encounters table were the SubjectId (to link back to subject-specific information) and WghtKg (weight in kg).

Fixed-effects model. A general term used in multi-level statistical analysis; discussed in this book only with respect to meta-analysis, where it describes a statistical model in which the study weights and the variance of the summary effect estimate are based only on the within-study variances of the included studies. For example, in a meta-analysis of clinical trials of the effect of practicing yoga on depression, the results of the trials were variable; the summary effect based on the fixed-effects model was dominated by one large study, and the confidence interval was narrower than would have been estimated with a random-effects model. See also *random-effects model*.

Generalizability. The degree to which the results in a study sample are thought to apply to other populations. For example, the reviewer questioned the generalizability of the reported 90% success rate of intraluminal radioablation of lower esophageal webs, because the procedures were all performed by the gastroenterologist who had invented and then perfected the technique in 350 patients, whereas most practicing gastroenterologists would see only a handful of patients with the same problem in their careers.

Gold standard. An unambiguous method of determining whether or not a patient has a particular disease or outcome. For example, the gold standard for the diagnosis of hip fracture required radiologic confirmation by a board-certified radiologist.

Hazard rate. An epidemiologic term that measures the instantaneous rate at which an outcome occurs in a population. For practical purposes, it is almost always estimated as the rate of an outcome. For example, the hazard rate for developing coronary artery disease among women ages 50 to 59 years old was estimated as 0.008 per year.

Hazard ratio. The ratio of the hazard rate in those exposed to a risk factor divided by the hazard rate in those who are not unexposed; it is almost always estimated from a proportional hazards model

(Cox model). For example, the hazard ratio for developing coronary artery disease was 2.0 comparing men ages 50 to 59 years with women of the same ages.

Heterogeneity. A situation in which the association between a predictor and an outcome is not uniform, either among different studies or among different subgroups of subjects. For example, there is substantial heterogeneity among studies that have looked at the effects of postmenopausal estrogen on mood and cognition, with some studies showing positive effects, some adverse effects, and some no effect.

Homogeneity. A situation in which the association between a predictor and outcome is uniform in different studies. For example, there is homogeneity among reasonably sized studies that have looked at the effects of smoking on lung cancer: All have found a substantially increased risk among smokers.

Hospital-based controls. In the context of a case–control study, the selection of control patients from the same hospital(s) from which the cases were chosen. For example, in her study of whether eating processed meats was associated with upper gastrointestinal cancer, the investigator used hospital-based controls selected from patients who had non-malignant gastrointestinal diseases treated at the same hospital as the cases.

Hypothesis. A general term for a statement of belief about what the study will find. For example, the study hypothesis was that chronic use of anti-epileptic medication was associated with an increased risk of oral cancer. See also *null hypothesis* and *research hypothesis*.

Incidence. The proportion of subjects who develop an outcome during the follow-up period; sometimes called incidence proportion or cumulative incidence. For example, the investigators found that pregnant vegetarians had a lower incidence of preterm delivery than pregnant women who ate meat.

Incidence-density sampling. Within a nested case–control study, a technique to sample controls when an important exposure changes with time; thus, the exposure needs to be measured at a similar time in both cases and controls. For example, a nested case–control study to determine whether use of antihistamine medications, which varies seasonally, increases the short-term risk of hip fractures (presumably due to an increased risk of falling) used incidence-density sampling of controls, such that a control's use of an antihistamine was measured during the same month that a hip fracture occurred in a case.

Incidence rate. The rate at which a particular disease or outcome occurs in a group of subjects previously free of that disorder. Usually calculated as the number of new cases of the outcome divided by the person-time at risk. For example, the incidence rate of myocardial infarction was 35.3 per 1,000 person-years in middle-aged men, about twice the rate (17.4 per 1,000 person-years) in middle-aged women. See also *person-time*.

Inclusion criteria. A list of attributes required of the potential subjects for a study. For example, the inclusion criteria for a study were people ages 18 to 65 years who lived in San Francisco and had no prior history of depression. See also *exclusion criteria*.

Independent. This term can be used in at least two ways. First, it is the condition in which two variables do not influence each other. For example, the investigators determined that dietary consumption of nuts and serum glucose levels were independent: there was no evidence in their study that nut consumption affected glucose levels, or vice versa. Second, independent is used to refer to an effect that one variable has on another variable that does not depend upon (i.e., "is independent of") a third variable. For example, because she was concerned that maternal education and breastfeeding were associated with one another, the investigator adjusted for maternal education to estimate the independent effect of breastfeeding on language skills at age 2 years.

Independent variable. See *predictor variable*.

Inference. The process of drawing conclusions about a population based on observations in a sample. For example, because twice as many cases of bladder cancer as controls reported drinking well water ($P = 0.02$), the investigators made the inference that consumption of well water increases the risk of bladder cancer in the population.

Instrumental variable. A variable that is associated with the predictor variable, but not otherwise associated with the outcome variable; it therefore can be used to indirectly estimate the effect of the predictor on the outcome. For example, investigators found marked regional differences in the use of a new

influenza vaccine, so they were able to use region of residence as an instrumental variable to study the effect of the influenza vaccine on total mortality in older adults.

Intended sample. The group of subjects the investigator intended to include in a study, as described in the study protocol. For example, the intended sample for the study consisted of women with breast cancer who were seen initially for treatment on a Monday or Thursday at Longview Hospital (the days that the investigator or her research staff were available) and who were within 6 weeks of their original diagnosis, during the period from January 1, 2013, through June 30, 2014. See also *accessible population* and *sample*.

Intention-to-treat analysis. In a randomized trial, the process of comparing subjects based on the group to which they were randomly assigned, even if this is not the same as the treatment they actually received. This is the most rigorous form of analysis. For example, the investigators performed an intention-to-treat analysis to determine whether random assignment to receive 6 months of psychotherapy improved symptoms of anxiety as compared with random assignment to a control group that received a pamphlet about stress reduction. See also *per-protocol analysis*.

Interaction. Another name for *effect-modification*.

Intervention. In a randomized trial, the active treatment that subjects receive. Often used as an adjective (intervention group). For example, in a randomized trial of psychotherapy for the treatment of anxiety, the intervention consisted of 6 months of weekly one-hour sessions with a licensed psychologist that emphasized cognitive-behavioral approaches. See also *control* (second definition).

Kappa. A statistical term that measures the degree to which two (or more) observers agree whether or not a phenomenon occurred, beyond that expected by chance. Varies from –1 (perfect disagreement) to 1 (perfect agreement). For example, the kappa comparing how well two pathologists agreed about the presence of cirrhosis in a sample of liver biopsy specimens was 0.85.

Level of statistical significance. See *alpha*.

Likelihood ratio. A term used to describe the quantitative effects of a medical test result on the likelihood that a patient has the disease being tested for. It is defined as the likelihood of that test result in a patient *with* the disease divided by the likelihood of that result in a patient *without* the disease (the mnemonic is WOWO: *w*ith *o*ver *w*ithout). For example, the likelihood ratio for the characteristic symptoms of typical angina (exertional substernal pressure) is about 50 for the diagnosis of coronary artery disease.

Likert scale. A set of answers (usually 5) to a question that provides similarly spaced range of choices. For example, the potential answers to the question "How likely are you to return to this emergency room for care?" were as follows: Very likely, Somewhat likely, Neither likely nor unlikely, Somewhat unlikely, Very unlikely.

Logistic regression model. A statistical technique used to estimate the effects of one or more predictor variables on a dichotomous outcome variable, adjusting for the effects of other predictor and confounding variables. For example, in a logistic regression model, men were about twice as likely as women, and blacks about three times as likely as whites, to develop strokes, adjusting for age, blood pressure, and diabetes.

Marginals. The row and column totals of a contingency table. For example, looking at the marginals in the 2 x 2 table showed that there were similar numbers of men and women in the study.

Masking. See *blinding*.

Matching. In a case–control study, the process of selecting controls to be similar in certain attributes to cases, to reduce confounding by those attributes. For example, in a case–control study of the risk factors for brucellosis, controls were matched to cases by age (within 3 years), sex, and county of residence. See also *overmatching*.

Mean. The average value of a continuous variable in a sample or population; calculated as the sum of all the values of that variable divided by the number of subjects. For example, the mean serum cholesterol level in a sample of 287 middle-aged women was 223 mg/dL. See also *median* and *standard deviation*.

Measurement error. The situation in which the precision or accuracy (or both) of a measurement is less than perfect; thus, there is at least some measurement error for most variables (with the possible

exception of death). For example, to reduce measurement error, the investigator used a 2 kg stainless steel weight to calibrate the baby scale weekly.

Median. The value of a variable that divides a sample or population into two halves of (approximately) equal size; equivalent to the 50th percentile. Often used when a continuous variable has a few very high (or very low) values that would overly influence the mean value. For example, the median annual income in the sample of 54 physicians was $225,000. See also *mean* and *standard deviation*.

Mediator. A variable that is caused by the predictor of interest, and also causes the outcome; it accounts at least in part for *how* the predictor causes the outcome. For example, in studying the effect of obesity on the risk of stroke, the investigators did not control for diabetes, because they believed one mechanism by which obesity might lead to stroke was as a mediator that caused diabetes.

Medical test studies. A general term used for studies that measure how well a test (or a series of tests) identifies patients with a particular diagnosis or outcome. For example, the investigator performed a medical test study to determine the likelihood ratios for the presence and absence of typical angina (defined as exertional substernal chest pain or pressure) for the diagnosis of coronary artery disease.

Mendelian randomization. A technique for enhancing causal inference by taking advantage of the random inheritance of genes that affect susceptibility to a risk factor or treatment. For example, the likelihood of a causal relationship between maternal use of acetaminophen and asthma in children was enhanced by the observation that the association was significantly greater in mothers with the T1 genotype of glutathione S-transferase, an enzyme involved in the detoxification of an acetaminophen metabolite.

Meta-analysis. A process for combining the results of several studies with similar predictor and outcome variables into a single summary result. For example, a meta-analysis of 12 published studies found that use of nonsteroidal anti-inflammatory drugs was associated with a 28% greater risk of developing asthma.

Misclassification. A measurement error for a categorical variable in which subjects with one value of the variable are counted (misclassified) as having another value. For example, investigators were worried that because medical records were incomplete, some subjects who really had fallen during their hospitalization were misclassified as not having had a fall. See also *differential misclassification* and *nondifferential misclassification*.

Missing data. Data that were not collected during a study, whether at baseline or during follow-up. For example, the investigator was concerned that the relatively large proportion (34%) of subjects who had missing data on alcohol use may have biased her study of the risk factors for falls.

Multiple–cohort study. A cohort study that enrolls two or more distinct groups of subjects (the cohorts), and then compares their outcomes. Often used in studies of occupational exposures, in which the cohorts being compared are either exposed to a potential risk factor or not. For example, the investigators performed a multiple–cohort study of whether exposure to cosmic rays during airplane flights is associated with an increased risk of hematologic malignancies; the investigators studied four cohorts: pilots and flight attendants (who would be exposed to cosmic rays) and ticket agents and gate attendants (who would not). See also *double-cohort study*.

Multiple hypothesis testing. The situation in which an investigator studies more than one—and usually many more than one—hypothesis in a study, thereby increasing the risk of making a type I error unless the level of statistical significance is adjusted. For example, although the investigator reported a statistically significant ($P = 0.03$) association between consumption of vitamin D and cognitive decline, her results were criticized because she did not account for the effect of multiple hypothesis testing, since the study had looked at more than 30 nutritional supplements. See also *Bonferroni correction*.

Multivariate adjustment. A general term for the statistical techniques used to adjust for the effects of one or more potential confounding variables on the association between a predictor and outcome. For example, using multivariate adjustment, the study found that ingestion of supplemental vitamin D was associated with an increased risk of cognitive decline, adjusting for age, sex, education, baseline cognitive function, and smoking.

Negative predictive value. See *predictive value, negative*.

Nested case–control study. A case–control study in which the cases and controls are selected from a (larger) defined cohort or from among previously enrolled subjects in a cohort study. This design is

usually used when it is too expensive to make certain measurements in all of the subjects in the cohort; instead, they are made in samples that were stored at baseline in those subjects. For example, the investigators performed a nested case–control study to determine whether cytokine levels on newborn screening blood spots were associated with the development of cerebral palsy in the 2009 birth cohort of the state of Ohio.

Nominal variable. A categorical variable for which there is no logical order. For example, religious affiliation (Christian, Buddhist, Hindu, Moslem, Jewish, other, none) was coded as a nominal variable.

Non-differential bias. A type of bias that is not affected by whether a subject was a case or control (or occasionally, by whether a subject was exposed, or not exposed, to a third variable). Non-differential bias tends to make associations harder to find because it reduces apparent differences between groups. For example, although recall of past exposure to antibiotics was imperfect in both cases and controls, the bias appeared to be non-differential, in that a review of medical records indicated that both groups had similar inaccuracies. See also *differential bias*.

Non-inferiority trial. A clinical trial comparing a new treatment that has some advantages over an established treatment (e.g., the new treatment is safer, less expensive, or easier to use), with the goal of demonstrating that the efficacy of the new treatment is not inferior to the established treatment. For example, a trial of a new pain medication that does not cause drowsiness demonstrated that the new medication was not inferior to oxycodone for relief of post-operative pain.

Non-response bias. A type of bias in which failure to respond (e.g., to a questionnaire) affects the results of a study. For example, the investigators were concerned about non-response bias in their study of the effects of illicit drug use on the risk of developing renal failure.

Normalization. In a relational database, the process of eliminating redundancy and improving reliability by making sure that each data item is stored in no more rows or tables than necessary. For example, after the database consultant normalized the database, he could update a subject's telephone number by altering just one row in a single table.

Null hypothesis. The form of the research hypothesis that specifies there is no difference in the groups being compared. For example, the null hypothesis stated that the risk of developing claudication would be the same in subjects with normal lipid levels who were treated with a statin as in those treated with placebo.

Number needed to treat. The absolute number of people who need to receive a treatment in order to prevent the occurrence of one outcome. Calculated as the reciprocal of the risk difference. For example, when evaluating the benefits of treating mild-to-moderate hypertension, the number needed to treat was about 800 patients per year to prevent one stroke.

Observational study. A general term for a research design in which the investigators simply observe the subjects without making any interventions. Thus, this term includes cross-sectional, case–control, and cohort studies, but not randomized trials or before-after studies. For example, the examiners performed an observational study to determine the risk factors for melanoma.

Observer bias. The situation in which an investigator (or research assistant) makes a non-objective assessment that is affected by her knowledge of one or more of the subject's attributes, such as whether the subject is a case or control, or was exposed or not exposed to a particular risk factor. For example, observer bias was apparently responsible for the finding that, based on an interview, Hispanic teenagers were more likely to be characterized as having issues with anger management than Asians, since a self-administered survey and a review of school records found no differences between the two groups.

Odds. The risk of a disease (or other outcome) divided by 1 – risk. For example, if the lifetime risk of breast cancer among women is 15%, then the lifetime odds of developing breast cancer are 0.18 (0.15/0.85). Risk and odds are similar for rare diseases (those that develop in less than about 10% of persons).

Odds ratio. The ratio of the odds of a disease (or other outcome) in those exposed to a risk factor compared with the odds of that disease in those not exposed. The risk ratio and the odds ratio are similar when a disease is rare in both the exposed and the unexposed, because the odds and the risks of the disease are similar. For example, the odds ratio for renal failure among those with hypertension is 2.0, meaning that hypertensive patients are about twice as likely to develop renal failure as those who are not hypertensive.

One-sample *t* test. A statistical test used to compare the mean value of a variable in a sample to a fixed constant (a particular number). The most common type of one-sample *t* test is a *paired* t *test*, in which the sample mean for the difference between paired measurements (e.g., on the same subject at different points in time) is compared with zero. For example, the investigators found that men gained a mean (± SD) of 4 ± 3 kg in weight during their residencies ($P = 0.03$, by one-sample *t* test). See also *two-sample* t *test*.

One-sided hypothesis. An alternative hypothesis in which the investigator is interested in evaluating the possibility of committing a type I error in only one of the two possible directions (e.g., greater or lesser risk, but not both). For example, the investigator tested the one-sided hypothesis that smoking was associated with an increased risk of dementia. See also *two-sided hypothesis*.

Ordinal variable. A categorical variable whose values have a logical order. For example, current alcohol use was treated as an ordinal variable: The values were no alcohol consumption, 1 or 2 drinks per week, > 2 but < 7 drinks per week, 1 to 2 drinks per day, and ≥ 3 drinks per day. See also *nominal variable*.

Outcome. A general term for the endpoint(s) of a study, such as death or the occurrence of a disease. For example, in a study of whether radiosurgery was beneficial for patients with solitary brain metastasis, subjects were followed for the outcomes of death or placement in a skilled nursing facility.

Outcome variable. The formal definition of the outcome for each subject. For example, in a study of the effects of different types of exercise on body weight and body composition, the outcome variables were defined as the change in weight in kg from baseline to the final measurement after 1 year, and the change in waist circumference in cm during that same time period.

Overfitting. A problem that arises when investigators select variables or cutoff points for a multivariate model based partly on chance variation in the sample, leading to poor generalizability of results. For example, reviewers suspected overfitting when the authors reported the best model to predict recurrent cataracts included birth in the months of March or August for women between ages 65 and 74 years.

Overmatching. The situation in which matching beyond that necessary to control for confounding reduces the investigator's ability to determine whether a risk factor is associated with an outcome because the controls have become too similar to the cases. For example, because the controls were matched to cases by age (± 3 years), sex, race, and socioeconomic status, overmatching made it impossible to determine whether education was associated with the risk of stroke among subjects ages 65 years and older, since the matching variables are major determinants of education in that age group.

P value. Based on statistical tests, the probability of finding an effect (more precisely, a value of the test statistic) as large or larger than that found in the study by chance alone if the null hypothesis is really true. For example, if the null hypothesis is that drinking coffee is not associated with the risk of myocardial infarction, and the study found that the relative risk of myocardial infarction among coffee drinkers compared with nondrinkers was 2.0 with a *P* value of 0.10, there was a 10% probability of finding a relative risk of 2.0 or larger in this study if there was no association between coffee drinking and myocardial infarction in the population.

Paired measurements. Measurements closely linked with one another in some way, such as those done on different sides of the same person, different members of a twin pair, or (most commonly) the same participant at two different points in time, such as before and after an intervention. For example, in a study of the effect of an exercise program on glycohemoglobin levels in patients with Type II diabetes, paired measurements of glycohemoglobin included measurements made at baseline and again after 3 months of exercise.

Participant. Someone who participates in a research study. The term *participant* is often preferred over *subject* because it emphasizes that the person enrolled in the study is an active participant in advancing science, not merely a subject being experimented upon. For example, in a study of a new drug for treatment of insomnia, the participants are the people who are eligible for and enroll in the study.

Peer review. Review of a protocol, proposal, or manuscript by peers of the investigator who prepared these documents. For example, proposals submitted for funding to the NIH undergo a peer review process in which scientists in the same field score the protocol using well-defined criteria. Similarly, manuscripts submitted to medical journals are peer reviewed by scientists who help the journal editors decide whether the manuscript should be published.

Per-protocol analysis. In a clinical trial, an analysis approach in which data from participants are only included if the participants adhered to the study protocol, which is typically defined as taking or using the study intervention as instructed. For example, in a randomized trial of surgery compared with physical therapy for treatment of severe osteoarthritis of the knee, a per-protocol analysis would include data only from participants in the surgery group who actually underwent surgery and from participants in the physical therapy group who were adherent to the physical therapy regimen. See also *intention-to-treat analysis*.

Person-time. The sum of the amounts of time each of the subjects in a study or population is at risk, used as the denominator for calculation of incidence rates. It can be calculated as the number of subjects who are at risk of an outcome multiplied by their average time at risk. For example, the total person-time of follow-up among the 1,000 subjects who had an average of 2.5 years at risk was a total of 2,500 person-years, although 5% of the subjects were followed for 1 month or less. See also *incidence rate*.

Phase I trial. An early phase, generally unblinded, uncontrolled trial of escalating doses of a new treatment in a small number of human volunteers to test its safety. For example, a phase 1 trial of a new drug for treatment of menopausal hot flashes would generally include a small number of volunteers (with or without hot flashes) who receive escalating doses of the drug to determine its effects on blood counts, liver and renal function, physical findings, symptoms, and other unexpected adverse events.

Phase II trial. A small randomized (and preferably blinded) trial to test the effect of a range of doses of a new treatment on side effects as well as on surrogate or clinical outcomes. For example, a phase II trial of a new drug for treatment of hot flashes that has been shown to be safe in a phase I trial might enroll a small number of postmenopausal women with hot flashes, randomly assign them to two or three different doses of the new medication or placebo, and then follow them to determine the frequency of hot flashes, as well as side effects.

Phase III (pivotal) trial. A randomized (and preferably blinded) trial that is large enough to test the efficacy and safety of a new treatment. For example, if the optimal dose of a new treatment for hot flashes has been established in a phase II trial and the new treatment was acceptably safe, the next step would be a large phase III trial in which postmenopausal women with hot flashes are randomly assigned to the new treatment or placebo and followed for the occurrence of hot flashes and adverse effects.

Phase IV trial. A large study, which may or may not be a randomized trial, conducted after a drug is approved by a regulatory agency such as the U.S. Food and Drug Administration (FDA), often to determine the drug's safety over a longer term than is possible in a phase III trial. For example, after a new drug for the treatment of menopausal hot flashes has been approved by the FDA, a phase IV trial might include women with less severe hot flashes than those included in the phase III trial.

Pilot study. A small study conducted to determine whether a full-scale study is feasible, as well as to optimize the logistics and maximize the efficiency of the full-scale study. For example, a pilot trial of restorative yoga for prevention of diabetes in patients with insulin resistance might aim to demonstrate the feasibility of measuring with insulin resistance; refine and standardize the yoga intervention; and show that it is possible to recruit and randomize participants to yoga and control groups.

Placebo control. An inactive control that is indistinguishable from the active drug or intervention used in a randomized trial. For example, in a randomized, placebo-controlled trial of a new treatment for incontinence, the placebo should look, smell, taste, and feel the same as the new medication that is being tested.

Plagiarism. A type of scientific misconduct in which an investigator appropriates another person's ideas, results, or words without giving appropriate credit. For example, using another investigator's description of a new measurement method without appropriate attribution constitutes plagiarism.

Polychotomous categorical variables. Categorical variables with three or more categories. For example, blood group, which includes A, B, and O, is a polychotomous categorical variable.

Population. A complete set of people with specified characteristics. For example, the adult population of the United States with Type II diabetes could be defined as all U.S. adults who are taking a glucose-lowering medication or who have a fasting blood sugar level above 125 mg/dL.

Population-based sample. A sample of people who represent an entire population. For example, the National Health and Nutrition Examination Survey (NHANES), which provides data on a random sample of the entire population of the United States, is a population-based sample.

Positive predictive value. See *predictive value, positive.*

***Post hoc* hypotheses.** Hypotheses that are formulated after data have been analyzed. For example, in a study of the association between insomnia and the risk of stroke, the hypothesis that insomnia increases the risk of diverticulitis is a *post hoc* hypothesis.

Power. The probability of correctly rejecting the null hypothesis in a sample if the actual effect in the population is equal to or greater than a specified effect size. For example, suppose that exercise leads to an average reduction of 20 mg/dL in fasting glucose among diabetic women in the entire population. If an investigator set power at 90% and drew a sample from the population on numerous occasions, each time carrying out the same study with the same measurements, then in 9 of every 10 studies the investigator would correctly reject the null hypothesis and conclude that exercise reduces fasting glucose levels. See also *beta.*

Practice-based research networks. Networks in which physicians from community settings work together to study research questions of interest. For example, a study from a practice-based research network of treatments for carpal tunnel syndrome in primary care practice showed that most patients improved with conservative therapy. This contrasted with previous literature from academic medical centers that indicated that the majority of patients with carpal tunnel syndrome required surgery.

Precision. The degree to which measurement of a variable is reproducible, with nearly the same value each time it is measured. For example, a beam scale can measure body weight with great precision, whereas an interview to measure severity of depression is more likely to produce values that vary from one observer to the next.

Preclinical trial. Studies that occur before an intervention is tested in humans. Such trials might include cells, tissues, or animals. For example, the U.S. Food and Drug Administration requires preclinical trials in two different animal species to document safety before new drugs can be tested in humans.

Predictive validity. A term that describes how well a measurement represents the underlying phenomenon it is intended to measure, based upon its ability to predict related outcomes. For example, the predictive validity of a measurement of depression would be strengthened if it was associated with the subsequent risk of suicide.

Predictive value, negative. The probability that a person with a negative test result does not have the disease being tested for. For example, in a population of men with a prevalence of prostate cancer of 10%, the negative predictive value of a prostate specific antigen (PSA) ≤ 4.0 ng/mL is about 91%. See *prevalence, prior probability, sensitivity* and *specificity*.

Predictive value, positive. The probability that a person with a positive test result has the disease being tested for. For example, in a population of men with a prevalence of prostate cancer of 10%, the positive predictive value of a prostate specific antigen (PSA) > 4.0 ng/mL is about 30%. See *prevalence, prior probability, sensitivity* and *specificity*.

Predictor variable. In considering the association between two variables, the one that occurs first or is more likely on biologic grounds to cause the other. For example, in a study to determine if obesity is associated with an increased risk of sleep apnea, obesity would be the predictor variable. In a randomized trial analyzed by intention to treat, the predictor variable is group assignment.

Pretest. An evaluation of specific questionnaires, measures, or procedures that can be carried out by study staff before a study starts. Its purpose is to assess the measure's functionality, appropriateness, or feasibility. For example, pretesting the data entry and database management system could be done by having study staff complete forms with missing, out of range, and illogical data to ensure that the data editing system identifies these errors.

Prevalence. The proportion of persons who have a disease or condition at one point in time. Prevalence is affected by both the incidence of a disease and duration of the disease. For example, the prevalence of systemic lupus erythematosis is the proportion of people who have this condition at a specific point in time; it might increase if the disease becomes more common or if treatment improves such that persons with the disease live longer.

Primary key. In a relational database, the field or combination of fields that uniquely identify each row in a particular table. For example, the investigator created a unique VisitNumber to serve as the primary key for a table of outpatient visits.

Principal investigator. The person who has ultimate responsibility for the design and conduct of a study, and the analysis and presentation of the study findings. For example, the institutional review board asked to speak with the study's principal investigator because some members had questions about the protocol.

Probability sampling. A random process, usually using a table of random numbers or a computer algorithm, to guarantee that each member of a population has a specified chance of being included in the sample, thereby providing a rigorous basis for making inferences from the sample to the population. For example, an observation from a probability sample of 5% of persons with chronic obstructive pulmonary disease (COPD) based on hospital discharge diagnoses from all hospitals in California should provide reliable findings about risk factors for rehospitalization and death.

Propensity score. The estimated probability that a study participant will have a specified value of a predictor variable, most often the probability of receiving a particular treatment. Controlling for the propensity score (e.g., by matching, stratification, or multivariable analysis) is one method for dealing with confounding by indication: Instead of adjusting for all factors that might be associated with the outcome, the investigator creates a multivariate model to predict receipt of the treatment. Each subject is then assigned a predicted probability of treatment (the propensity score), which can then be used as the only confounder when estimating the association between the treatment and the outcome. For example, the investigators used a propensity score to adjust for the factors associated with the use of aspirin to determine the association between aspirin use and colon cancer.

Proposal. A document that includes a study protocol, budget, and other administrative and supporting information that is written for the purpose of obtaining funding from a granting agency. For example, the National Institutes of Health (NIH) requires proposals for funding of multiple types of research.

Prospective cohort study. A study design in which a defined group (the cohort) of study participants has baseline values of predictor variables measured and then is followed over time for specific outcomes. For example, the Nurses Health Study is a prospective cohort study of risk factors for common diseases in women. The cohort is a sample of registered nurses in the United States and the outcomes have included cardiovascular diseases, cancer, and mortality.

Protected health information. Individually identifiable health information. Federal health privacy regulations (called HIPAA regulations after the Health Insurance Portability and Accountability Act) require researchers to maintain the confidentiality of protected health information in research. For example, protected health information should not be stored on flash drives or sent via regular e-mail.

Protocol. The detailed written plan for a study. For example, the study protocol for a study specified that only subjects who could understand English at the eighth grade level were eligible for participation.

Publication bias. A distortion of the published literature that occurs when published studies are not representative of all studies that have been done, usually because positive results are submitted and published more often than negative results. For example, publication bias was suspected by authors of a meta-analysis that found that six small positive studies, but only one large negative study, had been published.

Quality control. The processes to ensure that the conduct of a study, including enrollment, measurements, laboratory procedures, and data management and analysis, are of the highest quality. For example, the investigators controlled the quality of data collection by preparing explicit written procedures for all study measurements in an operations manual and intermittently observing study staff to make sure they followed them.

Query. A command or instruction to a relational database to select or manipulate the data. For example, the study coordinator wrote a query to select names and contact information for all study participants who were due for a follow-up visit in the next 2 months that had not yet been scheduled.

Questionnaire. A measurement instrument consisting of a series of questions to obtain information from study participants. Questionnaires can be either self-administered or administered by study staff. For example, the Block Food Frequency Questionnaire asks about usual intake of 110 food items to assess intake of multiple nutrients and food groups.

Random-effects model. A general term used in multi-level statistical analysis; discussed in this book only with respect to meta-analysis, where it describes a statistical model in which the study weights and

variance of the summary effect estimate incorporate a term for the variability between the results of the individual included studies. For example, in a meta-analysis of clinical trials of the effect of practicing yoga on depression, the results of the trials were variable; thus, smaller studies contributed more to the summary effect based on the random-effects model, and the confidence interval was wider than with the fixed-effects model. See also *fixed-effects model*.

Random error. A departure of a measurement or estimate from the true value due to chance variation. Random error can be reduced by repeating measurements and by increasing the sample size. For example, if the true prevalence of the use of fish oil by persons with coronary disease in the population is 20%, a study that enrolls 100 participants might find that exactly 20% use fish oil, but just by random error, the proportion is likely to be a bit higher or lower than that.

Randomization. The process of randomly assigning eligible participants to one of the study groups in a randomized trial. The number of treatment groups and the probability of being assigned to any group are determined before randomization begins. Although eligible participants are usually assigned to two study groups with equal (50%) probability, random assignment can be made to any number of study groups with any predetermined probability. For example, in a study comparing two treatments with a placebo control, randomization could be to three groups, with 30% assigned to each of the two active treatment groups and 40% assigned to placebo.

Randomized blinded trial. A study design in which eligible participants are randomly assigned to the study groups with predetermined probability and study group assignment is not known to investigators, participants, or other staff involved in the study. For example, a randomized blinded trial of a new pill for treatment of diarrhea would require that eligible participants be randomly assigned to the new pill or an identical placebo pill (usually with 50% chance of being assigned to each group) and that the investigators, participants, and study staff not know if a participant is taking the active medication or placebo.

Random sample. A sample drawn by enumerating the units of the population and selecting a subset at random. For example, a random sample of persons with cataracts at an investigator's clinic would require that the investigator list all of the patients with cataracts and use a table of random numbers or computer-generated random numbers to select the sample. See also *probability sampling*.

Rate. A measure of risk, defined as the number of subjects who develop an outcome divided by the person-time at risk. For example, the rate of stroke in the study was 23 per 1,000 person-years. See also *hazard rate*.

Recall bias. A specific type of bias in which whether and how a subject remembers his exposure to a risk factor is influenced by another factor, especially by whether the subject is a case or control. For example, recall bias was thought to be the reason why cases of amyotrophic lateral sclerosis were more likely to recall exposure to insecticides than controls.

Receiver operating characteristic (ROC) curve. A graphical technique to quantify the accuracy of a diagnostic test and illustrate the trade-off between sensitivity and specificity at different thresholds for considering the test positive. The curve displays the rates of true positives (sensitivity) on the Y-axis and the corresponding rates of false positives (1 – specificity) on the X-axis at several cutpoints for considering the test positive. The area under the ROC curve, which ranges from 0.5 for a useless test to 1.0 for a perfect test, is a useful summary of the overall accuracy of the test. For example, the area under the ROC curve for the use of CT scans (which could be interpreted as Clearly positive, Likely positive, Not helpful, Likely normal, or Clearly normal) to diagnose appendicitis was 0.95, substantially better than the value of 0.77 for ultrasound (which had similar categories of interpretation).

Record. A row in a relational database table (best identified by a *primary key*) which includes information about that person, transaction, result, or event. For example, a Subjects table might have one record for each subject in the study, with StudyId as its primary key, as well as other information such as date of birth and sex as fields.

Recruitment. The process of identifying and enrolling eligible participants in a study. Recruitment methods vary depending on the nature of the study. For example, recruitment for the study included identifying eligible patients in specialty clinics, advertising in fliers and newspapers, and using the Internet and social media sites.

Recursive partitioning. A multivariate technique for classifying people according to their risk of an outcome; unlike techniques that require a model, such as logistic regression, it does not require any assumptions about the form of the relationship between predictor variables and outcome. It creates a classification tree that includes a series of yes/no questions, called a Classification and Regression Tree (CART). For example, using recursive partitioning, investigators determined that emergency department patients ages 20 to 65 years who had abdominal pain but who did not have loss of appetite, fever, or rebound tenderness were at low risk of acute appendicitis. See *clinical prediction rule* and *overfitting*.

Registry. A database of persons with a certain disease or who underwent a particular procedure. Studies can be conducted using registries by collecting outcome data as part of the registry, or by linking registry data to other sources, such as cancer registries or the National Death Index. For example, the San Francisco Mammography Registry obtains data on all women who undergo mammography at the three largest mammography centers in San Francisco; investigators have linked it with local cancer registries to estimate mammography accuracy.

Regression to the mean. The tendency for outlying (very high or very low) values to be closer to the population mean when repeated. For example, in a group of children selected for a study based on having systolic blood pressures above the 95th percentile, the majority of children were observed to have lower blood pressures at the first follow-up visit, even though they had not yet received any treatment.

Relational database. Software that allows storage of related information in a series of tables. The tables can be linked with one another through common fields. For example, a relational database for a study could include each subject's StudyID and BirthDate in a Subjects table, and StudyID and VisitDate in an Encounters table, which could have many encounters per subject. A participant's age on the day of an encounter can be calculated easily by using the StudyID to link each VisitDate to that participant's birth date.

Relative risk. See *risk ratio*.

Representative sample. A sample of people enrolled in a study that represents the target population. For example, in the Framingham Heart Study, the target population was all adults. The accessible population (to investigators located in Boston) was the adult population of the town of Framingham, Massachusetts. Investigators enumerated adults in Framingham and asked every second resident to enroll in the study. This approach could result in a representative sample, but some people refused to enroll and were replaced by volunteers. Since volunteers often have more healthy habits than non-volunteers, the sample may have over-represented healthy persons. In addition, the population of Framingham (which was mostly white) does not represent all U.S. adults, and certainly does not represent adults in other countries.

Reproducibility study. A study in which the reproducibility of a measurement is the main research question, typically performed by comparing the results of a measurement done multiple times by the same person or machine (intra-observer reproducibility) or the results of the same measurement done by different persons or machines (inter-observer reproducibility). For example, the investigators performed a reproducibility study to determine whether a new electronic stethoscope could improve the ability to detect diastolic heart murmurs.

Research hypothesis. A statement by the investigator that summarizes the main elements of the study, including the population of interest, the predictor and outcome variables, and an anticipated result. For statistical purposes, the research hypothesis is stated in a form that establishes the basis for tests of statistical significance, generally including a null and alternative hypothesis. For example, the research hypothesis was that migraine headaches would be associated with at least a 20% increase in risk of stroke.

Research misconduct. Illegal or unethical conduct of research, including plagiarism and fabrication or falsification of research data. For example, a research coordinator at the VA Medical Center in Albany, New York, was found to have repeatedly submitted false documentation to allow persons who did not qualify for a study to be enrolled. All data from the Albany site were subsequently excluded, such that the participants' time and effort were wasted. See also *scientific misconduct*.

Research proposal. A document written for the purpose of obtaining research funding that describes the proposed study design, participants, measurements, statistical analyses, and ethical issues. For example, the National Institutes of Health receives thousands of research proposals annually from investigators who seek funding for their studies.

Research question. The question a research project is intended to answer. A good research question should include the predictor and outcome of interest and the population that will be studied. Research questions generally take the form of "Is A associated with B in population C?" or (for a clinical trial) "Does A cause B in population C?" For example, "Does regular use of dental floss reduce the risk of coronary events in persons with diabetes?"

Response rate. The proportion of eligible participants who respond to a questionnaire or to a particular item on it. A low response rate can decrease the internal validity of the study and bias the outcome. For example, in a survey of high school students, a response rate of 20% to a question on marijuana use would suggest the result is not likely to be a valid estimate of the real rate of marijuana use among students. See also *missing data*.

Retrospective cohort study. A cohort study in which assembly of the cohort, baseline measurements, and follow-up happened in the past. For example, to describe the natural history of thoracic aortic aneurysms, an investigator conducting a retrospective cohort study in 2012 could obtain data from hospital discharge records of patients who had a diagnosis of aortic aneurysm in 2007 and use hospital discharge records and the National Death Index to determine which patients subsequently had a ruptured aortic aneurysm or died before 2012.

Risk difference. The risk for an outcome in one group minus the risk in a comparison group. For example, if the risk for venous thromboembolic events among women who are current users of estrogen is 5/1000 (0.5%) and the risk among those who never used estrogen is 2/1000 (0.2%), the risk difference among women using estrogen compared to nonusers is 3/1000 (0.3%). See also *number needed to treat*.

Risk ratio (relative risk). The risk for an outcome in one group divided by the risk in a comparison group. For example, if the risk for venous thromboembolic events among women who are current users of estrogen is 5/1000 (0.5%) and the risk among those who never use estrogen is 2/1000 (0.2%), the relative risk among women using estrogen compared with nonusers is 2.5. See also *hazard ratio* and *odds ratio*.

Run-in period. In a clinical trial, a brief period during which eligible participants take either the placebo or the active intervention; only those who achieve a certain level of adherence, tolerate the intervention, or have a beneficial effect on an intermediate outcome are eligible for the main trial. For example, in the Cardiac Arrhythmia Suppression Trial, only those who had a satisfactory reduction in premature ventricular contractions on active medication during the run-in period were randomized to continue medication or switch to placebo.

Sample. The subset of the population that participates in a study. For example, in a study of a new treatment for asthma, where the target population is all children with asthma and the accessible population is children with asthma in the investigator's town this year, the study sample is the children in the investigator's town this year who actually enroll in the study.

Sample size. This term has two meanings. It can either be the number of participants enrolled in a study, or the estimated number of participants needed for a study to be successful. For example, the investigator estimated that she needed to have a sample size of 54 subjects to have 90% power to detect a doubling in the risk of aggressive behavior among third-grade boys exposed to violent video games.

Sampling. The process of selecting participants to enroll in a study when the number of eligible participants is larger than the estimated sample size. For example, the investigator used a "1 in 3" sampling scheme to select, on average, every third eligible subject. See also *cluster sampling, consecutive sample, convenience sample, probability sampling, stratified random sampling,* and *systematic sample*.

Sampling bias. A systematic error that causes the sample of persons included in a study not to represent the target population. For example, if participants in a study of risk factors for osteoporosis were recruited from among patients hospitalized for hip fracture, falling may falsely appear to be a risk factor for osteoporosis due to sampling bias.

Scale. A common approach to measuring abstract concepts by asking multiple questions that are scored and combined into a scale. For example, the SF36 scale for measuring quality of life asks 36 questions that yield 8 scales related to functional health and well-being. (SF stands for "short form.") See also *Likert scale*.

Scientific misconduct. A general term for intentionally defrauding the scientific community, including research misconduct (fabrication and falsification of data and plagiarism), as well as guest and ghost authorship, and conflict of interest that is not disclosed or managed. For example, the investigator's

institution judged that she was guilty of scientific misconduct because she failed to disclose an equity interest in the company that made the medical device she was studying.

Secondary data analysis. Use of existing data to investigate research questions other than the ones for which the data were originally collected. Secondary data sets may include previous research studies, medical records, health care billing data, and death certificates. For example, hospital discharge data and the National Death Index could be used in a secondary data analysis to determine 1-year mortality among patients with a discharge diagnosis of acute pancreatitis.

Secondary research question. Questions other than the primary research question, often including additional predictors or outcomes. For example, if the primary research question is to determine the association of alcohol consumption among pregnant women and low birth weight infants, a secondary question might be to determine the association of alcohol consumption and anemia during pregnancy.

Selection criteria. Rules that define who is eligible to enroll in a study, including the inclusion and exclusion criteria. For example, in a clinical trial of transdermal testosterone to enhance libido in postmenopausal women, the selection criteria might be women aged 45 to 60 years with low libido who are free of coronary disease and have not had more than three menstrual periods in the prior year.

Sensitivity. The proportion of subjects with disease in whom a test is positive ("positive in disease," or PID). For example, compared with pathology results on biopsy, the sensitivity of a serum PSA test result > 4.0 ng/mL is about 20% for the detection of prostate cancer; in other words, 20% of men with prostate cancer will have a PSA > 4.0 ng/mL. See also *specificity*.

Sensitivity analysis. Using different methods (e.g., alternate definitions of predictor or outcome variables, different statistical tests) to determine if the results of the primary analysis are robust. For example, in a meta-analysis of clinical trials of the effect of selective serotonin reuptake inhibitors on depression, in a sensitivity analysis, the investigator might include only the blinded trials to demonstrate that the results are robust when the analysis is restricted to high-quality trials.

Simple hypothesis. A hypothesis with only one predictor variable and one outcome variable. For example, the investigator rephrased his complex hypothesis into the simple hypothesis that people who eat fruit at least five times a week are less likely to develop colon cancer. See also *complex hypothesis*.

Specific aims. In a research proposal, brief statements of the goals of the research. For example, one specific aim of a randomized trial of the effect of testosterone on bone mineral density in men might be: "To test the hypothesis that, compared with men assigned to receive a placebo patch, those assigned to receive the testosterone patch will have less bone loss during 3 years of treatment."

Specification. A design phase strategy to cope with a confounder by specifying a value of that confounder as an inclusion criterion for the study. For example, in a study of the effect of pacifier use on the risk of sudden infant death syndrome, the investigator might use specification to include only formula-fed infants in the study. If a decreased risk of sudden death was found in pacifier users, it could not be because they were more likely to be breastfed.

Specificity. The proportion of subjects without the disease being tested for in whom a test is negative ("negative in health," or NIH). For example, compared with pathology results on biopsy, the specificity of a serum PSA test result of > 4.0 ng/mL is about 95% for the detection of prostate cancer; in other words, 95% of men without prostate cancer will have a PSA ≤ 4.0 ng/mL. See also *sensitivity*.

Spectrum bias. The situation in which the accuracy of a test is different in the sample than it would have been in the population because the spectrum of disease (which affects sensitivity) or non-disease (which affects specificity) in the sample differs from that in the population in which the test will be used. For example, because of spectrum bias, a new serum test designed to diagnose esophageal cancer was found to be relatively accurate in a study of patients with advanced esophageal cancer compared to healthy medical students, but performed poorly when used in elderly patients with undiagnosed difficulty swallowing.

Spurious association. An association between a predictor variable and an outcome variable that is seen in a study but that is not true in the population, either due to chance or bias. For example, observational studies found a decreased risk of cardiovascular disease among persons who took beta carotene supplements. However, a randomized trial of beta carotene supplements found no effect on risk of cardiovascular disease, suggesting that the association observed in the observational studies was spurious.

Standard deviation. A measure of the variance (spread) in a continuous variable. For example, the investigator reported that the mean age in the cohort of 450 men was 59 years, with a standard deviation of 10 years.

Standard error of the mean. An estimate of the precision of the mean of a continuous variable in a sample; depends on both the standard deviation and the (square root of the) size of the sample. For example, the investigator reported that the mean age in the cohort of 450 men was 59 years, with a standard error of 0.48 years.

Standardization. Specific, detailed instructions for how to perform a measurement designed to maximize reproducibility and precision of the measurement. For example, in a study that measures blood pressure, standardization of the measurement could include instructions on preparing the participant, what size cuff to use, where to place the cuff, how high to inflate and deflate the cuff, and which heart sounds indicate systolic and diastolic blood pressure.

Steering committee. In a multi-center study, a committee that provides overall governance for the study. It is generally composed of the principal investigators of each study site, the coordinating center, and representatives of the sponsor. For example, the study's steering committee decided whether proposed ancillary studies should be conducted.

Stratification. An analysis phase strategy for controlling confounding by segregating the study participants into strata according to the levels of a potential confounder and analyzing the association between the predictor and outcome separately in each stratum. For example, in a study of the association between exercise and the risk of stroke, not exercising regularly might be associated with increased risk of stroke because many people who don't exercise are obese, and obesity increases stroke risk. To minimize the potential confounding effect of obesity, participants were stratified by their body mass index, and the analyses were carried out separately in those who were normal weight, overweight, or obese at baseline.

Stratified blocked randomization. A randomization procedure designed to ensure that equal numbers of participants with a certain characteristic (usually a confounder) are randomly assigned to each of the study groups. Randomization is stratified by the characteristic of interest; within each stratum, participants are randomly assigned in blocks of predetermined size. For example, in a trial of a drug to prevent fractures, a history of vertebral fracture is such a strong predictor of the outcome and of response to many treatments that it would be best to ensure an equal number of participants with and without vertebral fracture in each of the study groups. Therefore, the investigators used stratified blocked randomization to divide participants into two strata (those with vertebral fractures and those without such fractures); within each stratum, randomization was carried out in blocks of six to ten subjects.

Stratified random sampling. A sampling technique in which potential participants are stratified into groups based on characteristics, such as age, race, or sex, and a random sample is taken from each stratum. The strata can be weighted in various ways. For example, the investigators used stratified random sampling in a study of the prevalence of pancreatic cancer in California to oversample racial and ethnic minorities.

Subgroup analysis. Comparisons between randomized groups in a subset of the trial participants. For example, in a randomized trial of the effect of a selective estrogen receptor modulator (SERM) on recurrence of breast cancer, the investigators performed a subgroup analysis of the effect of treatment by stage of cancer, comparing the effect of the SERM to placebo among women with stage I, stage II, and stages III and IV disease.

Subject. See *participant*.

Subject bias. See *recall bias*.

Summary effect. In a meta-analysis, the weighted average effect seen in the included studies; the formula for the weights depends on the model. For example, in a meta-analysis of randomized trials of the effect of an angiotensin-converting enzyme (ACE) inhibitor on mortality in patients with coronary disease, the summary effect with the fixed-effects model was the weighted mean relative risk, weighted by the inverse of the variance of the relative risk in each included study. See also *fixed-effects model* and *random-effects model*.

Suppression. A type of confounding in which the confounder diminishes the apparent association between the predictor variable and the outcome variable because it is associated with the predictor but affects the outcome in the *opposite* direction. For example, an association between smoking and skin wrinkles could be missed ("suppressed") in a study if smokers were younger and confounding by age was not controlled.

Surrogate marker. A measurement thought to be associated with meaningful clinical outcomes. A good surrogate marker usually measures changes in an intermediate factor in the main pathway that determines the clinical outcome. For example, an increased CD4 lymphocyte count in patients with human immunodeficiency virus (HIV) infection is a good surrogate marker for the effectiveness of antiretroviral drugs because it predicts a lower risk of opportunistic infections.

Survey. A cross-sectional study in a specific population, usually involving a questionnaire. For example, the National Epidemiologic Survey on Alcohol and Related Conditions enrolled a representative sample of adults in the United States and asked questions about present and past alcohol consumption, alcohol use disorders, and utilization of alcohol treatment services.

Survival analysis. A statistical technique used to compare times to an outcome (not necessarily survival) among groups in a study. For example, in a randomized trial of the effect of coronary artery bypass surgery compared with percutaneous coronary angioplasty for the prevention of myocardial infarction and death, survival analysis could be used to compare time from starting treatment to either of those outcomes in the two groups.

Systematic error. See *bias*.

Systematic review. A review of the medical literature that uses a systematic approach to identify all studies of a given research question, clear criteria to include a study in the review, and standardized methods to extract data from the included studies. A systematic review may also include a meta-analysis of the study results. For example, the investigator did a systematic review of all studies that tested whether zinc supplements reduced the risk of developing colds.

Systematic sample. A sample that is drawn by enumerating the units of the eligible population and selecting a subset of the population using a pre-specified process. For example, in the Framingham Heart Study, investigators constructed a list of all adult residents of the town of Framingham, Massachusetts, and then selected every other resident to be included in the study as part of a systematic sample.

***t* test (or *Student's* t *test*).** A statistical test used to determine whether the mean value of a continuous variable in one group differs significantly from that in another group. For example, among study participants who were treated with two different antidepressants, a *t* test could be used to compare the mean depression scores after treatment in the two groups (an unpaired two-sample *t* test) or the mean change from baseline to after treatment in the two groups (a paired two-sample *t* test). See also *one-sample* t *test* and *two-sample* t *test*.

Target population. A large set of people defined by clinical and demographic characteristics, to which the study investigator wishes to generalize the results of a study. For example, the target population for a study of a new treatment for asthma in children at the investigator's hospital might be children with asthma throughout the world.

Time series design. A within-group study design in which measurements are made before and after each participant (or a whole community) receives an intervention. This design eliminates confounding because each participant serves as his own control. However, within-group designs are susceptible to learning effects, regression to the mean, and secular trends. For example, using a time series design, fasting blood glucose levels were measured among a group of patients with diabetes before starting an exercise program and again after the program was completed to determine if exercise lowered fasting glucose levels. See also *within-group design*.

Translational research. Research that aims to translate scientific findings to improve health. Translational research may aim to test basic science findings from the laboratory in clinical studies in patients (often called "bench-to-bedside" or "T1 research") or to apply the findings of clinical studies to improve health in populations (often called "bedside to population" or "T2 research"). For example, a study to

determine whether a genetic defect that causes congenital deafness in mice has a similar effect in humans would be a T1 research study; a study to determine whether a statewide effort to screen newborns with a test that measures cortical response to sound to detect hearing loss improves school performance would be a T2 research study.

Two-sample *t* test. A statistical test used to compare the mean value of a variable in a sample with its mean value in another sample. For example, the investigators found that participants treated with olive oil supplements gained a mean of 10 mg/dL in high-density lipoprotein cholesterol levels during the study as compared with an increase of 2 mg/dL among those treated with placebo ($P = 0.14$, by two-sample *t* test). See also *one-sample* t *test*.

Two-sided hypothesis. An alternative hypothesis in which the investigator is interested in evaluating the possibility of committing a type I error in both of the two possible directions (e.g., greater risk or lesser risk). For example, the investigator tested the two-sided hypothesis that salsa dancing was associated with an increased or decreased risk of dementia. See also *one-sided hypothesis*.

Type I error. An error in which a null hypothesis that is actually true in the population is rejected because of a statistically significant result in a study. For example, a type I error occurs if a study of the effects of dietary carotene on the risk of developing colon cancer (with alpha set at 0.05) concludes that carotene reduces the risk of colon cancer ($P < 0.05$) when there is actually no association. See also *false-positive result*.

Type II error. An error in which a null hypothesis that is actually false in the population is not rejected by a study (i.e., $P >$ alpha). For example, a type II error occurs if a study fails to reject the null hypothesis that carotene has no effect on the risk of colon cancer ($P > 0.05$) when carotene actually does reduce the risk for colon cancer. See also *false-negative result*.

Validity. The degree to which a measurement represents the phenomenon of interest. For example, the score on a quality of life questionnaire is valid to the extent that it really measures quality of life.

Variability. The amount of spread in a measurement, usually calculated as the standard deviation. For example, if change in body weight produced by a diet ranges from substantial weight gain to substantial weight loss, the change is highly variable. See also *standard deviation* and *standard error of the mean*.

Variable. A measurement that can have different values. For example, sex is a variable because it can take two different values–male or female. See also *categorical variable, confounding variable, continuous variable, dichotomous variable, discrete variable, nominal variable, ordinal variable, outcome variable,* and *predictor variable*.

Verification bias. (Also called work-up bias or referral bias). A bias in the assessment of the accuracy of a test that occurs when participants selectively undergo disease verification by gold standard testing based partly on the results of the study test itself. For example, if a study of the accuracy of chest percussion to diagnose pneumonia included only patients who had a chest x-ray, and if those with dullness to percussion were more likely to get an x-ray, the sensitivity of percussion would be falsely increased, and specificity falsely decreased due to verification bias.

Visual analog scale. A scale (usually a line) that represents a continuous spectrum of answers, from one extreme to the other. Typically, the line is 10 cm long and the score is measured as the distance in centimeters from the lowest extreme. For example, a visual analog scale for the severity of pain might present a straight line with "no pain" on one end and "unbearable pain" on the other end; the study participant marks an "X" at the spot that best describes the severity of his pain.

Vulnerable persons. Potential study participants who are at greater risk for being used in ethically inappropriate ways in research. For example, because people with cognitive impairments or communication problems may be unable to give fully informed consent to participate in research, they are considered vulnerable persons. Other examples include children, prisoners, fetuses, and persons of low socioeconomic status.

Washout period. In a crossover study, the period of time between the first and second treatment designed to allow the effects of the intervention to wear off and the outcome measure to return to baseline. For example, in a crossover trial comparing a diuretic medication to placebo for treatment of high blood

pressure, the investigator might allow a one-month washout period with no treatment between the two treatment periods to allow blood pressure to return to baseline.

Within-group design. A study design in which measurements are compared in a single group of participants, most often at two different time periods. This design eliminates confounding because each participant serves as his own control. However, within-group designs are susceptible to learning effects, regression to the mean, and secular trends. For example, using a within-group design, fasting blood glucose levels were measured among a group of patients with diabetes before starting an exercise program and after the program was completed to determine if exercise lowered fasting glucose levels. See also *between-groups design, one-sample* t *test,* and *time series design.*

***Z* test.** A statistical test used to compare proportions to determine if they are statistically significantly different from one another. Unlike the chi-squared test, which is always two-sided, the *Z* test can be used for one-sided hypotheses. For example, a one-sided *Z* test can be used to determine if the proportion of prisoners who have diabetes is significantly greater than the proportion of free-living persons who have diabetes. Similarly, a two-sided *Z* test (or chi-squared test) could be used to determine if the proportion of prisoners who have diabetes is significantly different (i.e., smaller or larger) than the proportion not in prison who have diabetes.

Index

Note: Page numbers followed by *f* indicate figures; those followed by *t* indicate tables.