Bryan **Kestenbaum**

EPIDEMIOLOGY AND BIOSTATISTICS

AN INTRODUCTION TO CLINICAL RESEARCH

Springer

Epidemiology and Biostatistics

Bryan Kestenbaum

Epidemiology and Biostatistics

An Introduction to Clinical Research

Editors

Kathryn L. Adeney, MD, MPH
Department of Epidemiology,
University of Washington, Seattle

Noel S. Weiss, MD, Dr. PH
Department of Epidemiology,
University of Washington, Seattle

Contributing Author

Abigail B. Shoben, MS
Department of Biostatistics,
University of Washington, Seattle

Springer

Bryan Kestenbaum
University of Washington
Seattle, WA
USA
brk@u.washington.edu

ISBN 978-0-387-88432-5 e-ISBN 978-0-387-88433-2
DOI 10.1007/978-0-387-88433-2
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009927087

Preface

This textbook was born from a disparate collection of written materials that were created to teach Epidemiology and Biostatistics to second year medical students at the University of Washington. These materials included handouts, practice problems, guides to reading research articles, quizzes, notes from student help sessions, and student emails. The primary goal of these written materials, and now this book, is to recreate the perspective of learning Epidemiology and Biostatistics for the first time. With critical editing assistance from Epidemiology faculty, graduate students in Epidemiology and Biostatistics, and the students themselves, I have tried to preserve the innate logic and connectedness of clinical research methods and to demonstrate their application.

The textbook is designed to provide students with the tools necessary to form their own informed conclusions from the clinical research literature. More than ever, a clear understanding of the fundamental aspects of Epidemiology and Biostatistics is needed to successfully navigate the increasingly complex methods utilized by modern clinical research studies.

This book could not have been created without the dedicated help of the editors, the teaching assistants, and the students, who asked the important questions. I would especially like to thank my family who patiently allowed me so much time to write.

Seattle, WA Bryan Kestenbaum, MD MS

Contents

Epidemiology

Chapter 1
Measures of Disease Frequency

Learning Objectives

1. Measures of disease frequency:
   a. Clarify the significance of a particular health problem
   b. Help guide resource allocation
   c. Provide basic insight into the pathogenesis of disease
2. *Point prevalence* describes the amount of disease at a particular point in time.
3. *Incidence* describes the number of new cases of disease that develop over time.
4. Incidence may be expressed as *incidence proportion* or *incidence rate.*
5. Incidence rate accounts for follow-up time
6. Measures of disease frequency can be stratified, or broken up, by person, place, or time characteristics to gain insight into a disease process.

In December 1998, a 55-year-old woman presented to her local emergency department complaining of profound weakness and difficulty walking. She first noticed pain and weakness in her shoulders about 7 days earlier. The weakness progressed to involve her thigh muscles; she then developed nausea and noticed that her urine appeared dark. Over the next 48 h, her weakness further intensified and she became unable to stand on her own power.

Her previous medical conditions included high blood pressure, asthma, and high serum cholesterol levels. Her father had died at an early age from heart disease. She did not smoke cigarettes and rarely drank alcohol. Her regular medications included aspirin, diltiazem, and cerivastatin. She first started taking cerivastatin 2 weeks earlier to treat high cholesterol levels.

She appeared ill. There was no fever, the blood pressure was 140/95 mmHg, and the pulse was 48 beats/min. She was unable to raise her hips or her shoulders against gravity and her quadriceps muscles were diffusely tender. The rest of her physical examination, including neurological function, was normal.

The urine was dark amber in color. Laboratory testing revealed a serum creatinine level of 8.9 mg/dl, indicating severe kidney failure, and a serum potassium level of 7.6 mEq/liter (normal level is 2.5–4.5 mEq/liter). She was admitted to the hospital for emergent dialysis.

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_1,

Further diagnostic testing revealed a serum level of creatine kinase, an enzyme that normally exists inside of muscle tissue, of 178,000 units/liter (normal level is <200 units/liter). The patient was diagnosed with acute rhabdomyolysis, a condition characterized by severe, systemic muscle breakdown with release of muscle contents into the blood. One of these muscle components, myoglobin, is toxic to the kidney and causes kidney failure.

Cerivastatin (Baycol), a synthetic inhibitor of 3-hydroxy-3-methylglutaryl-coenzyme-A reductase, belongs to a class of cholesterol-lowering medications called "statins." The drug was approved on the basis of lowering serum cholesterol levels in 1997. At the time of this patient's presentation, no rhabdomyolysis cases associated with cerivastatin use had been reported in the literature. Could cerivastatin be causing this patient's rare and potentially fatal condition?

Epidemiology is concerned with investigating the *cause of disease*. In this example, there are some reasons to suspect that cerivastatin might be causing rhabdomyolysis. The disease developed soon after initiation of cerivastatin. Similar cholesterol-lowering medications can also damage muscle tissue, though the severe rhabdomyolysis seen in this case would be rare.

During the first 100 days following approval, the cerivastatin manufacturer received seven case reports of rhabdomyolysis among people using cerivastatin. Should these seven cases be cause for concern? It is difficult to answer this question based on the case report data alone. The next step is to estimate the *frequency* of rhabdomyolysis in cerivastatin users. According to company sales data, there were 3,100 cerivastatin prescriptions dispensed during the first 100 days of drug approval. On the basis of these data, the estimated frequency of rhabdomyolysis in cerivastatin users is 7/3,100, or 0.2%.

This disease frequency seems relatively small, but rhabdomyolysis is a rare and potentially fatal condition. The next step is to compare the frequency of rhabdomyolysis among cerivastatin users to that of an appropriate control population. One possible control population might be people who were using similar cholesterol-lowering medications. In previous clinical trials, a total of 33,683 people had been assigned to a statin medication other than cerivastatin; 8 of these people developed rhabdomyolysis. On the basis of these trial data, the estimated frequency of rhabdomyolysis among people using other statin drugs is 8/33,683, or 0.02%.

While these findings may be partially distorted, due to differences in the compared populations, the observed tenfold greater frequency of rhabdomyolysis among cerivastatin users is concerning.

Two years after cerivastatin was approved, the manufacturer conducted an internal investigation of rhabdomyolysis rates. They found cerivastatin use to be associated with a 20-fold greater risk of rhabdomyolysis compared with other approved statin medications. Their findings were not reported or published. Case reports of rhabdomyolysis associated with cerivastatin use began to surface in the medical literature in 2000–2001.[1] By August 2001, there were 31 fatal cases of rhabdomyolysis attributed to cerivastatin. At this time, the company voluntarily removed the drug from the market.[2]

1.1 Importance of Measures of Disease Frequency

The cerivastatin example demonstrates that *measures of disease frequency* represent key initial information needed to investigate the cause of disease. While it may be tempting to dive in and conduct novel discovery studies or high-profile clinical trials, an important first question about a disease process is, "how frequently does the disease occur?" Measures of disease frequency can help answer several important questions:

1. Measures of disease frequency can provide big picture information about a disease, framing public health questions and guiding resource allocation. For example, years after the invention of chronic dialysis for kidney failure, researchers observed that rates of cardiovascular death among dialysis patients were approximately 30-fold greater than those of the general population.[3] These disease frequency data lead to a dramatic increase in funding for research of links between kidney and cardiovascular diseases.
2. Measures of disease frequency describe the *absolute* risk of a disease. For example, many studies have reported that smoking causes a more than tenfold increase in the *relative* risk of lung cancer. Rate data reveals that the *cumulative lifetime risk* of lung cancer for a person who smokes is approximately 18%. The rate data are important for counseling patients and understanding the impact of the disease on the population.
3. Measures of disease frequency can be categorized, or *stratified*, by person, place, and/or time characteristics to gain insight into the pathogenesis (mechanism) of disease. For example, rates of multiple sclerosis, an autoimmune disease that affects the central nervous system, vary considerably by geographic region within the USA. Areas with the lowest sunlight exposure have the highest incidence of multiple sclerosis. These rate data lead some researchers to investigate whether vitamin D deficiency might participate in the pathogenesis of multiple sclerosis.[4] Vitamin D is obtained from sunlight exposure and can suppress inflammation and T-cell function.

The two most commonly used measures of disease frequency are *prevalence* and *incidence*.

1.2 Prevalence

Point prevalence measures the amount of a disease at one particular point in time. Prevalence is defined as the *proportion* of people who have the disease:

$$\text{Prevalence}\,(\%) = \frac{\text{number of people with disease}}{\text{number of people in the population}} \times 100\%$$

Because prevalence is always a ratio of some number of people and some number of people, prevalence estimates are often multiplied by 100% and expressed as %.

Example 1.1. What is the prevalence of anxiety disorder among second year medical students?

Solution: Administer a standardized test for anxiety disorder to 200 second year medical students; find that 12 meet the definition of anxiety disorder. *Prevalence*= $12/200 \times 100\% = 6\%$.

In the medical literature, the term "*prevalent"* is also used to indicate a *previous history* of a chronic disease. For example, "*prevalent diabetes"* and "*prevalent coronary disease"* may be used in clinical research studies to indicate previous or current diagnoses of these conditions, because they are rarely cured and considered to be present indefinitely after diagnosis. In contrast, a previous history of a short-lived disease, such as influenza, would not be considered to represent prevalent disease, unless that condition was found to exist at the time of measurement.

Prevalence measures help to describe the *current burden* of a disease in a population in order to facilitate planning and resource allocation. For example, if the prevalence of anxiety disorder was truly 6% among second year medical students, the medical school might consider implementing specific counseling programs for students with this disorder. Analogously, if the prevalence of diabetes is found to be 40% among patients in a particular chronic kidney disease clinic, then that clinic might implement routine blood glucose monitoring.

1.3 Incidence

Incidence is a measure of the number of *new* cases of disease that develop over time. There are two definitions of incidence, differing only by the choice of the denominator:

$$\text{Incidence proportion} = \frac{\text{number of new cases of disease}}{\text{population without disease at baseline}}$$

$$\text{Incidence rate} = \frac{\text{number of new cases of disease}}{\text{person - time at risk}}$$

Another term for incidence rate is incidence density.

Example 1.2. What is the incidence of influenza infection among UW medical students during a 3-month period from January through March 2002?

Solution: Suppose that there are 500 UW medical students beginning in January 2000, and 5 new cases of influenza develop from January through March (3 months of follow-up).

$$\text{Incidence proportion} = \frac{5\,\text{cases}}{500\,\text{people}} \times 100\% = 1\%, \text{or } 1 \text{ per } 100 \text{ people}$$

$$\begin{aligned} \textit{Incidence rate} &= \textit{5 cases} \;/\; \textit{(500 people x 3 months)} \\ &= \textit{5 cases} \;/\; \textit{1500 person-months} \\ &= \textit{.003 cases} \;/\; \textit{person-month} \\ &= \textit{3 cases} \;/\; \textit{1000 person-months} \end{aligned}$$

Incidence rates are typically reported as the number of cases of disease per some rounded measurement of time at risk, such as 1,000 or 100,000 person-years. The inclusion of time at-risk in the denominator of incidence rate provides a more precise description of incidence than incidence proportion, particularly if study subjects contribute different amounts of time at risk to a study. For the influenza example, suppose that some of the medical students in the study are assigned to a distant clinical rotation for part of the study period, and cannot report influenza to a research study center during that time. Because the study could not detect the development of influenza for these away months, time at-risk should be adjusted to consider only months in which the disease could be captured. Table 1.1 presents data for the first six students in the study.

Table 1.1 Calculation of individual risk time

	Unadjusted time at risk	Months away	Actual time at risk
Student 1	3 months	0 month	3 months
Student 2	3 months	0 month	3 months
Student 3	3 months	0 month	3 months
Student 4	3 months	1 month	2 months
Student 5	3 months	2 month	1 month
Student 6	3 months	3 month	0 month
Total			12 months

The total time at risk contributed by these six students is 12 months. If two cases of influenza developed in these six students, then the *incidence rate* of influenza for these 6 students would be 2 cases/12 person-months = 16.7 cases per 100 person-months.

The calculation of incidence rate from person-time data may be best appreciated using a diagram representing time at risk and disease status for each individual in a study, as shown in Fig. 1.1.

In this diagram, students one and three are followed for the full 3-month study period and do not develop influenza. Student two is also followed for the full 3-month period, but develops influenza at the end of the study. Student four develops influenza after only 2 months of follow-up. Since a study subject is longer at risk for developing incident (new) disease once the disease has occurred, we would consider the total time at risk for student four to be 2 months. Students five and six do not develop the disease and contribute approximately 1.25 and 0.25 months of time at-risk, respectively.

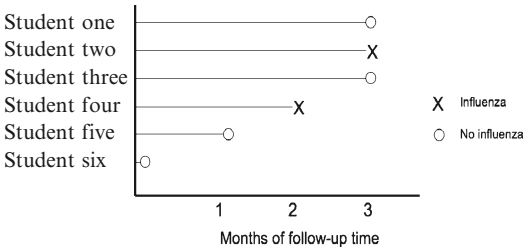


Fig. 1.1 Diagram of individual risk time and disease status

The importance of counting time at risk is highlighted by examples in which follow-up time differs between comparison groups. For example, a study compares rates of cellulitis, a common skin infection, between children seen in primary care clinics at a county and a university hospital. Investigators study 500 children from each site and follow them for up to 5 years. Results using incidence proportion data indicate that cellulitis is more common at the university hospital:

	Number of children	Cellulitis cases	Incidence proportion
County hospital	500	10	2%
University hospital	500	15	3%

However, these raw rate data do not include time at risk. It is possible that children from the county hospital are lost to follow-up or dropout of the study more frequently than those from the university hospital. If this were the case, then the incidence proportion data would be misleading. Examining the same data using incidence *rates* reveal a different result:

	Number of children	Cellulitis cases	Time at risk (person-years)	Incidence rate (per 1,000 person-years)
County hospital	500	10	1,200	8.3
University hospital	500	15	2,000	7.5

The incidence rate of cellulitis is actually higher at the county hospital after accounting for person time at risk.

Incidence measures help to provide clues as to the *cause* or *development* of a disease. For the anxiety disorder example, suppose that the *incidence proportion* of anxiety disorder was 5% per year among medical students. This incidence measure would consider only new cases of anxiety disorder that developed during medical school; students with prevalent anxiety disorder at the beginning of medical school would not be counted. These incidence data suggest that certain aspects of medical school might contribute to anxiety disorder, prompting a more thorough search for possible causal factors. In contrast, the prevalence data for anxiety

disorder alerted to a high burden of disease in the student population, motivating implementation of treatment programs.

1.4 Relationship Between Prevalence and Incidence

The prevalence of a disease is a function of how often new cases develop and how long the disease state lasts. For example, the incidence of influenza may be relatively high during influenza season; however, the prevalence of influenza at any point in time is likely to be low because illness is short-lived; people either recover quickly, or in rare cases, die from the disease. In contrast, the prevalence of diabetes is likely to be high because there is a steady incidence of new cases, and the disease, though treatable, is rarely cured.

The mathematical relationship between prevalence and incidence is $P = I \times D$, where P is the prevalence, I the incidence, and D the duration of disease. Figure 1.2 presents a graphical depiction of the relationship between incidence and prevalence.

Individuals in a population will acquire a disease at some rate (incidence). They will remain with the disease until they either get well, die, or leave the population (and cannot be counted).

1.5 Stratification of Disease Frequency by Person, Place, and Time

Once we calculate measures of disease frequency, we can examine whether these measures vary by personal characteristics, geography, and/or time periods. *Stratification* refers to the process of separating analysis by subgroups. For example, the prevalence of diabetes among all US adults is approximately 9.0%; the prevalence of diabetes *stratified by race* is 8.2% among whites and 14.9% among Native Americans.

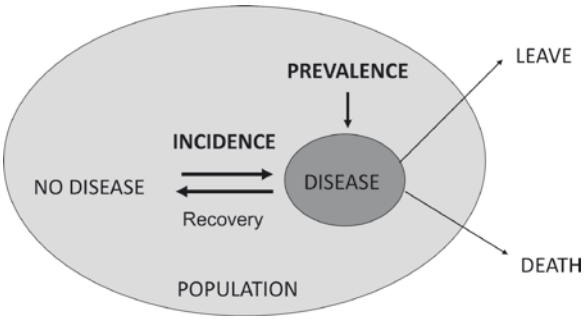


Fig. 1.2 Relationship between incidence and prevalence of disease

Example 1.3

A study was conducted to describe the epidemiology of latex allergy in healthcare workers. At the beginning of the study (*baseline*), 500 females and 540 males underwent skin-prick testing; 40 females and 22 males tested positive for latex sensitivity. Follow-up skin tests were performed 2 years later, and 29 new cases of latex sensitivity were detected.

The overall *prevalence* of latex sensitivity among healthcare workers *at baseline* is 62/1,040 × 100% = 6%.

The prevalence stratified by sex is 40/500 × 100% = 8% in females, and 22/540 × 100% = 4% in males. Therefore, latex sensitivity appears to be twice as common in women compared to men at baseline. To calculate incidence, the number of *new* cases of latex sensitivity that develop over time, we will exclude the 62 people who already had prevalent latex sensitivity at baseline.

$$\text{Incidence proportion of latex allergy} = \frac{29}{1,040-62} \times 100\% = 3\%$$

$$\text{Incidence rate of latex allergy} = \frac{29}{(1,040-62)\times 2 \text{ years}}$$
$$= 15 \text{ cases per } 1,000 \text{ person-years}$$

Information about sex is not provided for the new cases of latex sensitivity, so incidence data cannot be stratified by sex in this example.

1.5.1 Disease Frequency Measurements Stratified by Characteristics of Person

Examples of personal characteristics include age, race/ethnicity, and sex. For example, polycythemia vera is a myeloproliferative disorder characterized by an abnormal increase in red blood cell mass. The estimated *prevalence* of polycythemia vera among individuals aged 35–44 is 9 cases per 100,000, whereas the estimated prevalence in people aged 75–84 is 163 cases per 100,000. Polycythemia rates are also greater in men and in people of Jewish/Eastern European ancestry. These stratified disease frequency data begin to define *risk factors* for the disease.

1.5.2 Disease Frequency Measurements Stratified by Characteristics of Place

The incidence of multiple sclerosis varies considerably by geographic region within the USA. Areas with the lowest sunlight exposure, such as Seattle, have the

highest incidence of multiple sclerosis. Vitamin D is ascertained from sunlight exposure and may play an important role in suppressing autoimmunity. Circulating vitamin D levels are particularly low in regions with reduced sunlight exposure. These disease frequency data, stratified by place, suggest the hypothesis that vitamin D deficiency may play a role in the pathogenesis of multiple sclerosis.

1.5.3 Disease Frequency Measurements Stratified by Characteristics of Time

In 1970, approximately 5% of all births in the USA were by Cesarean section delivery. By the year 2000, nearly 25% of US babies were born by Cesarean section. These strong temporal changes in rates generate a number of hypotheses.[5] One possibility is that maternal age has also increased during this time period, leading to more complicated pregnancies that may require Cesarean section. A second possibility is that improved fetal monitoring technology that can detect small changes in fetal status may prompt more surgical intervention. A third possibility is that the routine use of repeat Cesarean section has become standard practice in the USA because of data demonstrating an increased risk of uterine rupture in women who have a vaginal birth after a first Cesarean section.[6] Disease frequency measurements stratified by time are often hypothesis generating, motivating further studies to uncover the true causes of a disease process.

1.5.4 Disease Frequency Measurements To Complement Experimental Data

Stratified measures of disease frequency can also be used to corroborate experimental data. For example, animal models have suggested that estrogen can slow the progression of chronic kidney disease by reducing expression of proinflammatory cytokines and decreasing the extent of fibrosis within the kidney.[7] Can measures of disease frequency *in humans* be used to corroborate these provocative experimental data?

One possibility is to obtain estimates of the incidence of chronic kidney disease, stratified by sex and menopausal status, as depicted by the hypothetical data presented in Table 1.2.

Table 1.2 Rates of chronic kidney disease according to sex and premenopausal status

	Incident rate of chronic kidney disease (cases per 1,000 person years)
Premenopausal women	4.0
Postmenopausal women	6.3
Men	6.5

These data reveal a lower chronic kidney disease incidence rate among premenopausal women, compared with postmenopausal women and men. These disease frequency data support the hypothesis that estrogen protects against chronic kidney disease, and represent a first step toward investigation of this process in humans.

Chapter 2
General Considerations in Clinical Research Design

Learning Objectives

1. The study population refers to all people who enter a study.
2. Common exclusion criteria in clinical studies include:
   a. Exclusion of people who have prevalent disease to focus on incident outcomes.
   b. Exclusion of people who have major disease risk factors to focus on the exposure of interest.
   c. Exclusion of people whose disease development may be missed in the study.
3. The choice of study population influences the generalizability of study findings.
4. The exposure is a factor that may explain or predict the presence of an outcome.
5. The outcome is a factor that is being explained or predicted in the study.
6. Observational studies observe the exposure; interventional studies assign the exposure.
7. Several factors favor causal inference in epidemiology research:
   a. Randomized evidence
   b. Strong associations
   c. Temporal relationship
   d. Exposure-varying response
   e. Biological plausibility

This chapter presents fundamental elements of a clinical/epidemiological research study: the study population, exposure, outcome, and the general study design. Specific study designs, along with their inherent strengths and weaknesses, are discussed in subsequent chapters. The chapter concludes with a discussion of factors that favor causal inference.

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_2,

2.1 Study Population

2.1.1 Definition of the Study Population

The term *study population*, or *patient population* in a research study, refers to *all of the people who enter a study*, regardless of whether they are treated, exposed, develop the disease, or drop out after the study has begun. Typically, a study population originates from some larger *source population*, which is then narrowed using *exclusion criteria*.

Consider a study to address the hypothesis that estrogen use increases the risk of developing venous thromboembolism (VTE). Biological data suggest a link between estrogen use and VTE because estrogen interferes with a circulating factor that normally inhibits blood clot formation. One approach to studying this question in humans would be to identify a group of estrogen users and a group of nonusers, and then to follow them prospectively for the development of VTE. What exclusion criteria should be applied to best address the research question?

First, investigators may exclude women who have a previous history of VTE. Typically, studies of disease development focus on the incidence of disease, and therefore *exclude people who have prevalent disease at the beginning of the study*. Prevalent disease may be defined by any history of a chronic disease if that disease cannot be fully eradicated. For example, a previous history of coronary heart disease or diabetes is typically considered to represent "prevalent disease" in a clinical research study because these conditions are rarely cured, though frequently treated.

Second, investigators may exclude women with known major VTE risk factors, such as cancer or recent major surgery. Excluding women who have known causes of VTE will increase confidence that any new VTE cases that develop during the study can be attributed to the use versus nonuse of estrogen, rather than to some other factor. However, there are limits to using exclusion as a means to focus on a specific cause of disease. There are many risk factors for VTE, including genetic mutations, smoking, and kidney disease. Excluding women with *any* VTE risk factor would significantly diminish the size of the available study population, and would markedly restrict interpretation of study findings. In clinical practice, physicians do not test for a battery of rare mutations before prescribing estrogen. Results of a study that excluded women who had any predisposing mutation to VTE would have diminished generalizability to clinical practice.

Third, investigators may decide to exclude women whose VTE might be missed during the study. For example, subjects who plan on moving from the area may develop VTE in another geographic location and might not be counted. Subjects who have a history of frequently missing clinic appointments might be difficult to contact and less likely to complete surveillance procedures, potentially developing VTE that could be missed. When selecting a suitable study population, it is important for investigators to consider how they will capture the disease in question, and consider limiting the study population to subjects whose disease would be counted if it were to develop during the study.

2.1.2 Choice of Study Population and Generalizability of Study Findings

The choice of study population directly influences the *generalizability* or *applicability* of study findings. The following examples illustrate how different types of study populations influence the generalizability of the results.

Example 2.1. Clinic-based descriptive study of resistant soft tissue infections in children.

Study objective: Describe prevalence of resistant organisms in kids with soft tissue infection.

Study findings: Among 30 children with soft tissue infection demonstrated by wound culture, 7 (23%) had organisms that were resistant to first line antibiotics.

Study population: We studied 30 consecutive children with soft tissue infection from our outpatient pediatric clinic in greater Minneapolis. Children were included if they were 2–16 years old, had a soft tissue infection that required incision and drainage, and demonstrated organisms by culture examination.

Clinic-based studies such as these are generally the easiest and least expensive to conduct, because potential study subjects are often readily accessible to the investigators and important data may already be available as part of clinical practice. However, findings from these types of studies tend to be poorly generalizable to other populations. In this case, antibiotic resistance patterns may be specific to the geographic region where the study was conducted, and could be influenced by antibiotic prescription practices of this particular pediatric clinic. Results from this study will apply only to children who live in Minneapolis, and have the same age, socioeconomic background and pediatrician practice patterns as the children who attended this particular pediatric clinic. Moreover, clinic-based study populations tend to be relatively small and therefore highly subject to sampling variation or random fluctuation in results. Imagine that there were 500 total soft-tissue infections in this pediatric clinic and that 50 (10%) were caused by resistant organisms. Selecting random samples of 30 cases from the total group of 500 would yield a wide range of estimates of the proportion of resistant organisms.

Example 2.2. Health network-based study of hip fracture in chronic kidney disease.[8]

Study objective: Estimate hip fracture rates in people with and without kidney disease.

Study findings: Late-stage kidney disease associated with fourfold greater risk of hip fracture.

Study population: The source population consisted of all male veterans with at least one outpatient primary care or internal medicine subspecialty clinic visit within the Northwest Veterans Integrated Service Network, a collection of eight Veterans

Affairs facilities located in Washington State, Idaho, Oregon, and Alaska. Exclusion criteria were prior history of hip fracture, diagnosis of cancer, chronic dialysis, or renal transplantation.

Health network-based studies such as these offer an improvement in generalizability compared to clinic-based studies. In this case, the observed association of late-stage kidney disease with hip fracture is more broadly applicable to men in multiple geographic locations, and is not limited to the practice patterns of a single clinic or clinic. Further, the closed nature of the Veterans Affairs medical system increases the likelihood that hip fractures will be captured in this study population. This more general study population helps to support the hypothesis that late-stage kidney disease might play a causal role in the development of hip fracture. It is important however to note that this study population consisted predominantly of older men. Results may not apply to women, who have a considerably higher underlying prevalence of osteoporosis.

Example 2.3. Community-based study of lipoprotein a [Lp(a)] levels and stroke.[9]

Study objective: Examine association of Lp(a) with the risk of incident stroke in older adults.

Study findings: Higher Lp(a) associated with greater stroke risk in men, but not in women.

Study population: The Cardiovascular Health Study (CHS) is a community-based study of heart disease and stroke in 5,888 ambulatory adults aged 65 years and older. Participants were recruited from four communities by randomly sampling from age-stratified Medicare eligibility lists in each area. Subjects were excluded if they were institutionalized, were required a proxy to give consent, were required a wheelchair, or were receiving treatment for cancer.

Community-based studies such as these are typically the most costly and complex to conduct because they involve leaving the health care system in favor of the community for subject recruitment. The use of a community-based study population yields the most generalizable study findings, because many people in a community never see a doctor, let alone the inside of a hospital. Study findings for Lp(a) are expected to broadly apply to the general population of ambulatory older adults, not just those who are receiving health care.

2.1.3 Where to Find Information About the Study Population in a Clinical Research Article

Description of the study population is usually, but not always, described in the *first one or two paragraphs of the methods section* of a research article. This paragraph should explicitly define the source population from which study subjects were

selected and detail and justify the specific inclusion and exclusion criteria. This information may also be presented in flow chart form.

2.2 Exposure and Outcome

2.2.1 Definition

One broad segment of clinical/epidemiological research focuses on the relationship, or association, between an *exposure* and an *outcome* of interest. The term "exposure" is carried over from infectious disease epidemiology; however, it is used to describe *any factor or characteristic that may explain or predict the presence of an outcome*. Examples of exposures include the use of a particular medication, smoking, and blood pressure.

The term *outcome* refers to the factor that is being predicted. The outcome is often a disease, but can be any clinical characteristic, such as cholesterol level, vaccination status, or medication use.

The distinction between exposure and outcome is highlighted in the following examples.

Example 2.4. A study examines whether vaccination against pneumococcal pneumonia is effective at preventing the disease. Investigators review medical charts from 250 patients enrolled in a primary care clinic to determine whether they received the pneumococcal vaccine.

One possible study question might be:

$$\text{Pneumococcal vaccine} \xrightarrow{\text{association?}} \text{Pneumococcal pneumonia}$$

In this example, the *exposure* of interest is pneumococcal vaccination and the *outcome* of interest is pneumococcal pneumonia. The study would estimate the association of pneumococcal vaccination status with the risk of developing pneumococcal pneumonia.

Example 2.5. A study explores whether education level plays a role in a person's decision to use herbal medications. A group of 500 people from a local shopping mall are asked to complete a questionnaire querying their education level and their frequency of herbal medication use.

A specific study question of interest might be:

$$\text{Education level} \xrightarrow{\text{association?}} \text{Herbal medication use}$$

In this example, the *exposure* of interest is education level and the *outcome* of interest is herbal medication use. Note that the outcome that is being predicted in this example is medication use. Other studies may examine medication use as the exposure, or predictor, of a disease.

Example 2.6. A study is conducted to examine whether heart failure influences survival after a first myocardial infarction. A total of 1,000 people who survive a first myocardial infarction undergo a history, physical examination, and echocardiography testing to determine the presence or absence of systolic heart failure. Subjects are followed until they die or drop out of the study.

A specific study question might be:

$$\text{Systolic heart failure} \xrightarrow{\text{association?}} \text{Survival}$$

In this example, the exposure of interest is systolic heart failure and the outcome of interest is survival. Note that the exposure in this case happens to be a disease. Other studies might focus on risk factors for systolic heart failure, and thus examine heart failure as the outcome of interest.

2.2.2 Specifying and Measuring the Exposure and Outcome

Effective clinical research requires *highly focused* definitions of exposure and outcome variables. For the pneumococcal vaccine study example, possible choices of *specific exposures* might be: (1) *any* previous pneumococcal vaccination (yes vs no), (2) *recent* pneumococcal vaccination within the last year (yes vs no), or (3) the number of years since last pneumococcal vaccination. Similarly, possible choices for the outcome variable, pneumococcal pneumonia, might be: (1) pneumonia, defined by chest x-ray findings or (2) pneumonia, defined by cough, fever, rales, and evidence of streptococcal DNA in sputum. For the heart failure example, the exposure variable, systolic heart failure, might be defined by a history of heart failure symptoms, such as shortness of breath and lower extremity swelling, plus evidence of a low cardiac ejection fraction measured by echocardiography.

Once exposures and outcomes are specifically defined, they should be measured as carefully as possible. For example, it is possible that medical chart records that document pneumococcal vaccination are less accurate than computerized pharmacy records that indicate actual disbursement of the vaccine. Errors in measuring exposure and outcome data are common; the consequences of measurement error are discussed in chapter 8.

2.2.3 Where to Find Exposure and Outcome Data in a Clinical Research Article

Information regarding ascertainment of exposure and outcome data is usually described beginning after the description of the study population in the methods section. This section should specifically define the exposure and outcome variables

and spell out exactly how they were collected and measured, so that the validity of the study findings can be judged, and so that the study could be repeated under similar conditions. If possible, studies should also describe the accuracy of the data collection methods. For the pneumococcal vaccine example, the authors might state in the methods section, "We defined pneumococcal pneumonia by a hospitalization code for pneumonia plus culture evidence of streptococcus in the sputum. This definition correctly classified 88% of pneumococcal pneumonia cases compared to gold-standard PCR testing for pneumococcus in a subsample of cases."

2.3 Interventional Versus Observational Study Designs

Epidemiologic research studies can be broadly categorized as *interventional* or *observational*. The distinction arises in the method by which study subjects are exposed. An interventional study *assigns* exposure to study subjects, usually at random, whereas an observational study *observes* the exposure, which occurs "naturally".

For the previous example of estrogen use and venous thromboembolism, consider first the interventional approach. At considerable effort and expense, investigators could recruit a large group of postmenopausal women who were not already using estrogen. Each recruited subject would then meet with a study pharmacist who would flip a coin – if the coin comes up heads, the pharmacist would assign estrogen therapy and if the coin comes up tails, the pharmacist would assign an identical appearing pill that did not contain estrogen (placebo). The coin flip would be conducted in secret, such that neither participant nor study investigators were aware of the results. Following completion of this random exposure assignment process, the baseline characteristics of exposed (estrogen users) and unexposed (placebo) study subjects can be compared in a *table of baseline characteristics*, which is usually the first table of a clinical research article. Baseline characteristics of subjects assigned to estrogen versus placebo are presented in Table 2.1.

Table 2.1 Baseline characteristics from a randomized trial comparing estrogen to placebo

	Estrogen use ($N = 1814$)	Placebo ($N = 1814$)
Age (years)	61.9 (15.7)	61.9 (15.8)
Race		
Caucasian	1241 (68.4)	1247 (68.7)
African–American	450 (24.8)	467 (25.7)
Other	123 (6.8)	100 (5.5)
Smoker	715 (39.4)	702 (38.7)
History of cardiovascular disease	795 (49.4)	833 (48.3)
Mean serum albumin (mg/dl)	3.92 (0.58)	3.90 (0.58)
Mean serum cholesterol (mg/dl)	188.6 (54.7)	188.4 (56.7)

All values in the table are either mean (standard deviation) or number of patients (percent).

Notice how the distributions of baseline characteristics are nearly identical between women assigned to estrogen versus those assigned to placebo, *due to the random assignment of the exposure* between groups. In contrast, one characteristic that is dramatically different between the groups is the use of estrogen at the beginning of the study; this should be 100% in the estrogen group and 0% in the placebo group. Assuming reasonable compliance with the assigned therapy during the study, potential differences in VTE outcomes between the two groups can be ascribed only to differences in estrogen use *and not to other characteristics of estrogen users because in every other respect women assigned to estrogen are similar to those assigned to the placebo*.

Now consider the observational approach. A population of postmenopausal women would be identified, but this time researchers would *observe* whether each study subject was using or not using estrogen. Investigators could ascertain estrogen status by querying computerized pharmacy records or by asking study participants to bring in their medication bottles to the study examination. Notice that in the observational study design, no participants receive a placebo. Baseline characteristics of estrogen users and nonusers are compared in Table 2.2.

Under the observational approach, exposure groups are more unbalanced with regard to some of their baseline characteristics. There are also a lower number of estrogen users; only 611 women in the observational study population were using estrogen. Some baseline characteristics appear to be unrelated to estrogen use, such as the serum albumin level. Others, such as previous cardiovascular disease, appear to be strongly linked with estrogen use, possibly because clinicians falsely believed that estrogen use reduced cardiovascular disease and may have prescribed estrogen for that purpose.

Like the interventional design, baseline estrogen use differs dramatically between exposure groups in the observational design: 100% versus 0%. However, if the two groups are found to have different VTE rates during follow-up, there will be residual uncertainty as to whether observed differences are due to estrogen use or due to differences in other characteristics of the women who used estrogen. This phenomenon is known as confounding and will be covered in detail in Chaps. 9 and 10. *Freedom from confounding is the primary advantage of interventional trials over observational study designs.*

Table 2.2 Characteristics from observational study comparing estrogen use versus no use

	Estrogen users ($N = 611$)	Nonusers ($N = 1850$)
Age (years)	65.9 (15.7)	59.7 (15.8)
Race		
Caucasian	528 (86.4)	1382 (74.7)
African–American	66 (10.8)	266 (14.4)
Other	17 (2.8)	202 (10.9)
Smoker	177 (29.0)	747 (40.4)
History of cardiovascular disease	395 (64.6)	831 (44.9)
Mean serum albumin (mg/dl)	3.9 (0.58)	3.9 (0.51)
Mean serum cholesterol (mg/dl)	192.1 (54.7)	178.4 (56.7)

An important, but difficult concept is the distribution of unmeasured factors that are not presented in the baseline characteristics table, for example, exercise and dietary factors. In the interventional study design, if the sample size is reasonably large then random assignment will balance not only measured participant characteristics but also unmeasured characteristics that do not appear in the baseline table. So, it is expected that women assigned to estrogen in the randomized trial will have similar distributions of exercise patterns and dietary characteristics as those assigned to placebo. In contrast, there is no easy way to predict whether unmeasured characteristics will be balanced in an observational study and increasing the sample size will have no effect on this uncertainty.

2.4 Inferring Causation from Association Studies

2.4.1 Importance of Distinguishing Causation from Association

Epidemiologic research studies report *associations* between an exposure and an outcome, because association and not causation is actually observed. For example, studies detecting an association of LDL cholesterol levels with the occurrence of coronary heart disease did not observe LDL cholesterol entering arterial plaques and subsequently occluding blood vessels. This presumed sequence of events was based on multiple lines of evidence, from clinical association studies to basic experimental work. *There is usually no way of directly observing an exposure causing a disease in humans.*

Many associations are not causal. For example, receiving last rites in the intensive care unit is strongly *associated* with impending death; however, it is unlikely that receiving last rites *causes* a higher risk of death. Estrogen use has been *associated* with lower risks of heart disease in observational studies, but not in intervention trials, possibly because estrogen use is a marker of other healthy characteristics that are also linked with lower risks of heart disease, such as a healthier diet or compliance with other medical therapies. Separating association from causation in clinical research studies is critically important because uncovering causative factors in a disease process can lead to prevention and treatment. For example, an inference that the prone sleeping position caused sudden infant death syndrome (SIDS) led to successful prevention strategies that substantially lowered its risk.[10] An inference that LDL cholesterol levels caused coronary heart disease led to the creation of specific drugs that lower LDL cholesterol, and to subsequent clinical trials proving their efficacy.

Inferring causality from epidemiological studies may not be easy. Causal inference is hampered in clinical research by the fact that (1) multiple exposures often influence a single disease outcome, for example, many people with *low* LDL cholesterol levels still develop heart disease, because they have other risk factors that play an important causal role and (2) many exposures take a long time to influence the outcome; for example, the effect of diet on the risk of cancer.

2.4.2 Factors Favoring an Inference of Causation

Although we can never be sure that a particular exposure causes a particular outcome, a number of factors can be used to help decide whether an exposure of interest is likely to be a cause of a disease, rather than just being associated with it.

2.4.2.1 Evidence Arising from Randomized Studies

Studies that randomly assign subjects to one treatment group versus another are generally the most powerful way to show that an exposure is a *cause* of an outcome. Large randomized trials are usually free from confounding, that is, characteristics of subjects assigned to one particular treatment are usually very similar, on average, to those of subjects assigned to another treatment. If outcomes differ between treatment groups, it is reasonable to conclude that the treatment is the cause of the difference. Unfortunately, randomized studies are limited to exposures that can be easily assigned to people, such as medications or devices.

2.4.2.2 Strength of Association

For both interventional and observational studies, a strong association between exposure and outcome increases the likelihood that the exposure is a *cause* of that outcome. Note that strength of association is *not* the same as statistical significance. For example, a case-control study observed that infants in the prone sleeping position had a fivefold greater risk of SIDS (odds ratio 5.0, p-value = 0.001). Although the p-value is important to rule out chance as a possible explanation for these findings, the *strength of this association* – that infants in the prone sleeping position were *five times more likely* to develop SIDS – was important for establishing the prone sleeping position as a cause of SIDS. One reason that strong associations often indicate causation is that there can only be so much bias and error in a well-conducted observational study. For the SIDS example, some errors in classifying SIDS versus other causes of death may have occurred, and there may have been other aspects of infants who sleep in the prone versus supine position that could have also influenced the risk of SIDS; however, it is unlikely that such potential errors would account for the entirety of such a strong observed association. In general, relative risks greater than 2.0 or less than 0.5 are considered to indicate strong associations.

2.4.2.3 Temporal Relationship Between Exposure and Outcome

For an exposure to be considered as a cause of a disease, there should be evidence that the exposure was present *before* the disease developed. Consider a study that examines the association between cyclophosphamide chemotherapy and the risk of

secondary bladder cancer.[11] The study population includes patients who were free from bladder cancer at the beginning of the study, the exposure is the use of cyclophosphamide chemotherapy, and the outcome is newly diagnosed or *incident* bladder cancer that occurs at least 2 years after the initiation of chemotherapy. The investigators find that cyclophosphamide use is associated with a 4.5-fold greater risk of developing future bladder cancer. In this example, ensuring that the exposure (cyclophosphamide chemotherapy) clearly preceded the outcome (bladder cancer) in time strengthens the case for cyclophosphamide as a potential *cause* of secondary bladder cancer.

For second example, consider a hypothetical study that discovers high levels of a novel neurotransmitter, "DP-1" in the blood of people who have established major depressive disorder. These data alone do not clarify whether higher circulating DP-1 levels were present before the development of depression, or whether depression was present before DP-1 levels increased. The alternative possibility that DP-1 levels might rise *in response* to depression diminishes the case for causal inference.

2.4.2.4 Exposure-Varying Association

If a primary association between exposure and outcome is observed, the case for causal inference may be strengthened by additional evidence that the association differs predictably across different levels of the exposure. For the cyclophosphamide and bladder cancer example, the overall relative risk of secondary bladder cancer associated with cyclophosphamide use was 4.5. The investigators next examined the risk of bladder cancer associated with different cumulative doses of cyclophosphamide. Their findings are presented in Table 2.3.

Table 2.3 Cyclophosphamide dosage and the relative risk of secondary cancer

Cumulative cyclophosphamide dosage (g)	Relative risk of secondary cancer
None	*Reference group*
1–20	2.4
20–50	6.0
>50	14.5

This stepwise increased risk of bladder cancer associated with each higher cyclophosphamide dosage strengthens evidence for a causal relationship between cyclophosphamide and bladder cancer. The "dose–response relationship" need not be limited to a medication, and can apply to different levels of any exposure. For example, childhood streptococcal infections have been associated with the development of neuropsychiatric syndromes, such as Tourette's disorder. A well-conducted observational study observed that children who had a streptococcal infection were 2.2 times more likely to develop a future neuropsychiatric syndrome.[12] The investigators strengthened the case for a causal relationship by further showing that the

risk of neuropsychiatric syndromes increased steadily with the number of previous streptococcal infections.

2.4.2.5 Biological Plausibility

Causal inference relies on translational and basic science knowledge to make sense of observed epidemiologic associations. Associations that have proven biological plausibility based on experimental data are more likely to be causal than those not supported by scientific evidence. Note that biological plausibility derives from scientific evidence obtained from other studies. For the example of LDL cholesterol levels and heart disease, multiple parallel studies took place: basic science studies demonstrated LDL cholesterol deposition in the arterial wall, translational studies showed enlargement of atherosclerotic plaque size by angiography among patients with higher LDL cholesterol levels, observational studies indicated associations of higher LDL cholesterol levels with a greater risk of developing clinical heart disease, and interventional trials demonstrated a reduced risk of death and cardiovascular disease in patients treated with drugs specifically designed to lower LDL cholesterol levels. This example highlights the importance of interdisciplinary collaboration for producing quality science and for moving the medical research field forward.

Chapter 3
Case Reports and Case Series

Learning Objectives

1. Case reports and case series describe the experience of one or more people with a disease.
2. Case reports and case series are often the first data alerting to a new disease or condition.
3. Case reports and case series have specific limitations:
    a. Lack of a denominator to calculate rates of disease
    b. Lack of a comparison group
    c. Selecting study populations
    d. Sampling variation

Case reports and case series represent the most basic type of study design, in which researchers describe the experience of a single person (*case report*) or a group of people (*case series*). Typically, case reports and case series describe individuals who develop a particular new disease or condition. Case reports and case series can provide compelling reading because they present a detailed account of the clinical experience of individual study subjects. In contrast, studies that evaluate large numbers of individuals typically summarize the data using statistical measures, such as means and proportions.

Example 3.1. A case series describes 15 young women who develop breast cancer; 9 of these women report at least once weekly ingestion of foods packaged with the estrogenic chemical bisphenol A (BPA). Urine testing confirms the presence of BPA among all nine case women.

It is tempting to surmise from these data that BPA might be causally related to breast cancer. However, case reports/case series have important limitations that preclude inference of a causal relationship.

First, case reports/case series *lack denominator data* that are necessary to calculate the *rate of disease*. The denominator refers to the population from which the diseased subjects arose. For example, to calculate the incidence proportion or incidence rate of breast cancer among women exposed to BPA, the total number of women who were exposed to BPA or the total number of person-years at risk is needed.

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*, 25
DOI 10.1007/978-0-387-88433-2_3, © Springer Science+Business Media, LLC 2009

$$\text{Incidence proportion of breast cancer in BPA exposed} = \frac{\text{BPA exposed with breast cancer}}{\text{Total BPA exposed women}}$$

$$\text{Incidence rate of breast cancer in BPA exposed} = \frac{\text{BPA exposed with breast cancer}}{\text{BPA exposed person-years at risk}}$$

Disease rates are needed for comparison with historically reported disease rates, or with rates from a selected comparison group. Unfortunately, obtaining the necessary denominator data may not be easy. In this example, additional data sources are needed to determine the total number of BPA-exposed women from whom the breast cancer cases arose. The case series data alone cannot be used to calculate the rate of breast cancer because they do not include the total number of women who were exposed to BPA.

A second problem with case report/case series report data is the *lack of a comparison group.* The 60% prevalence of BPA exposure among women with breast cancer seems unusually high, but what is prevalence of BPA exposure among women *without breast cancer*? This comparison is critical for addressing the hypothesis that BPA might be a cause of breast cancer.

A third limitation of case reports/case series is that these studies often describe *highly select individuals* who may not represent the general population. For example, it is possible that the 15 breast cancer cases originated from a single hospital in a community with high levels of air pollution or other potential carcinogens. Under these conditions, a fair estimate of breast cancer incidence among non-BPA-exposed women from the same community would be required to make an inference that BPA causes breast cancer.

A fourth limitation of case reports/case series is *sampling variation.* This concept will be explored in detail later in this book. The basic idea is that there is tremendous natural variation in disease development in humans. The fact that 9 of 15 women with breast cancer reported BPA exposure is interesting; however, this number may be very different in the next case series of 15 women with breast cancer simply due to chance. A precise estimate of the rate of a disease, independent from chance, can be obtained only by increasing the number of diseased subjects.

Recall the list of factors that are used to judge whether a factor may be a cause of disease:

1. Randomized evidence
2. Strength of association
3. Temporal relationship between exposure and outcome
4. Dose-response association

5. Biological plausibility

In general, *case reports/case series rely almost exclusively on biological plausibility* to make their case for causation. For the BPA and breast cancer case series, there is no randomized evidence, no measure of the strength of association between BPA and breast cancer, no reported dose–response association, and no evidence that BPA exposure preceded the development of breast cancer. The inference for causation derives completely from previous biological knowledge regarding the estrogenic effects of BPA.

Despite limitations of case series data, they may be highly suggestive of an important new association, disease process, or unintended side effect of a medication or treatment.

Example 3.2. In 2007, a case series described three cases of male prepubertal gynecomastia.[13] The report included detailed information on each subjects' age, body size, serum levels of endogenous steroids, and known exposures to exogenous hormones. It was discovered that all three otherwise healthy boys had been exposed to some product containing lavender oil (lotion, shampoo, soap), and that in each case, the gynecomastia resolved upon discontinuation of the product. Subsequent in vitro studies demonstrated endocrine-disrupting activity of lavender oil. This novel case series data may lead to further investigations to determine whether lavender oil, a common ingredient in commercially available products, may be a *cause* of gynecomastia.

Example 3.3. A vaccine designed to prevent rotavirus infection was found to cause weakening of the intestinal muscle layers in animals. Following release of the vaccine, a number of cases of intussusception (when one portion of the bowel slides into the next) were reported in children who received the vaccine, with some fatal cases.[14] The strong biological plausibility underlying this initial association, and knowledge that intussusception is otherwise rare in infants, was highly suggestive of a causal relationship and the vaccine was removed from the market.

Chapter 4
Cross-Sectional Studies

Learning Objectives

1. Cross-sectional studies measure the exposure and the outcome at the same time.
2. Cross-sectional studies estimate the *prevalence* of a disease or condition.
3. Cross-sectional studies cannot establish a temporal relationship between the exposure and the outcome.

A cross-sectional study refers to a study design in which ascertainment of the exposure and the outcome occurs simultaneously. Measuring the exposure and outcome at the same time implies that there is *no follow-up time in a cross-sectional study*.

Example 4.1. Homocysteine, an amino acid formed during the conversion of methionine to cysteine, possesses proinflammatory and prothrombotic properties that might contribute to atherosclerosis. Researchers investigated whether higher serum homocysteine levels were associated with peripheral arterial disease among 6,600 men and women. Findings are presented in Table 4.1.

How can we interpret these data? Among the 6,600 study participants, 1,000 had homocysteine levels that were classified as high and 5,600 had levels that were classified as normal. We can calculate the *prevalence of peripheral arterial disease* among study participants who had high homocysteine levels:

$$\text{Prevalence} = \frac{100 \text{ people with disease}}{1{,}000 \text{ people with high homocysteine}} * 100\% = 10\%.$$

Similarly, we can calculate the *prevalence of peripheral arterial disease* among study participants who had normal homocysteine levels:

$$\text{Prevalence} = \frac{175 \text{ people with disease}}{5{,}600 \text{ people with normal homocysteine}} * 100\% = 3\%.$$

The cross-sectional study design results in knowing the amount of peripheral arterial disease that is present at the time of homocysteine measurement. As defined in Chap. 1, prevalence is used to describe the amount of disease in a population

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_4, © Springer Science+Business Media, LLC 2009

Table 4.1 Association of homocysteine level and peripheral arterial disease

	Peripheral arterial disease		
Homocysteine level	Yes	No	Total
High	100	900	1,000
Normal	175	5,425	5,600
Total	275	6,325	6,600

at any given point in time. This contrasts with incidence that is used to describe the number of new cases of disease that develop over time. The incidence of peripheral arterial disease cannot be determined in this cross-sectional study.

An important disadvantage of cross-sectional studies is the *inability to discern a temporal relationship between the exposure and the outcome*. Two alternate explanations for the homocysteine data are possible. On the one hand, it is possible that higher homocysteine levels precede and participate in the development of peripheral arterial disease. On the other hand, it is also possible that peripheral arterial disease leads to metabolic changes that include a subsequent rise in homocysteine levels. The cross-sectional study design cannot distinguish between these two possibilities because homocysteine levels and peripheral arterial disease were measured simultaneously. As a result, we are left uncertain as to whether higher homocysteine levels might cause peripheral arterial disease, or whether peripheral arterial disease might cause high homocysteine levels.

In some instances, the inability to discern temporality in a cross-sectional study is not of concern because only one direction of causality is biologically plausible. For example, consider a second study that simultaneously measures peripheral arterial disease status and polymorphisms (different sequences of nucleic acids) within the methylenetetrahydrofolate reductase (MTHFR) gene, which codes for an enzyme involved in homocysteine metabolism. The genetic study finds that a particular polymorphism within the MTHFR gene is associated with a twofold greater risk of peripheral arterial disease. In this example, one direction of causality that peripheral arterial disease causes the MTHFR gene polymorphisms is biologically impossible. We are left to conclude that differences in the MTHFR gene precede, and may cause peripheral arterial disease.

Cross-sectional studies are frequently performed while researchers are waiting for follow-up data to become available. For example, after conflicting interpretations of animal studies concerning the safety of the estrogenic chemical BPA in food packaging materials, the first large-scale epidemiological study of BPA in humans was published.[15] The investigators used data from the National Health and Nutrition Examination Study, a cross-sectional study that randomly sampled adults from the US population for interviews and laboratory measures. Participants provided a urine sample which was analyzed for BPA concentration, and at the same time reported their history of a number of physician-diagnosed diseases. Investigators found that participants with a self-reported history of cardiovascular disease or diabetes had higher mean urinary BPA levels compared to participants without these conditions.

These cross-sectional data indicate a greater prevalence of cardiovascular disease and diabetes among people with greater BPA exposure, as indicated by their urinary levels of this compound. It is tempting to conclude from these data that BPA might be a cause of these diseases. However, we should remain alert to the alternate possibility that cardiovascular disease and diabetes might influence dietary habits, which subsequently alter the amount of BPA exposure. This alternative explanation weakens the case for BPA exposure as a novel cause of disease. Ideally, this provocative cross-sectional study will be followed by more formal studies that measure BPA levels in healthy people without diabetes or cardiovascular disease, and then follow them for the development of incident diabetes and cardiovascular disease during long-term follow-up. However, it may take years or decades to collect these follow-up data.

Chapter 5
Cohort Studies

Learning Objectives

1. The fundamentals of a cohort study design are:
   a. Identify people who are free of disease at the beginning of the study
   b. Assemble cohorts of exposed and unexposed individuals
   c. Follow cohorts for the development of incident outcomes
   d. Compare the risks of incident outcomes in each cohort
2. Cohort studies have certain advantages:
   a. Can discern temporal relationships between the exposure and the outcome
   b. Can be used to evaluate multiple outcomes
3. Cohort studies have certain disadvantages:
   a. Observational design: other factors may be responsible for observed association
   b. May be inefficient for studying rare diseases or those with long latency periods
4. Cohort studies can be used to evaluate the risks and benefits of medication use.
5. Cohort studies can be used to calculate relative and attributable risks of disease.

5.1 Overview of Cohort Study Design

Cohort studies are a particular type of observational study design that improve on case series/case reports and cross-sectional studies. Cohort studies provide an estimate of the *incidence of disease or outcome*, typically include a formal comparison group, and temporally dissociate the exposure from the outcome as a means to strengthen the evidence for causation.

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_5, © Springer Science+Business Media, LLC 2009

Cohort studies are conducted in three fundamental steps:

1. Assemble or identify cohorts of exposed and unexposed individuals who are free of the disease/outcome of interest at the beginning of the study
2. Observe each cohort over time for the development of the outcome(s) of interest
3. Compare the risks of outcomes between the cohorts

Recall from Chap. 2 that the term "exposure" is used to describe any factor or characteristic that may explain or predict the presence of an outcome. Examples of exposures include serum levels of cholesterol, the use of a diuretic medication, smoking, and kidney function.

The first step in a cohort study is to identify cohorts of exposed and unexposed individuals. A *cohort* is a group of people, derived from the study population, who share a common experience or condition and whose outcome is unknown at the beginning of the study. In cohort studies, the investigators *observe* exposure status, which occurs "naturally." In contrast, in randomized trials, investigators *assign* the exposure to the study participants.

Example 5.1. Investigators wish to evaluate whether smoking causes premature failure of a kidney transplant. They recruit 200 patients with a functioning kidney transplant who are receiving care at a single transplant clinic. The investigators use a questionnaire to ascertain smoking status for all study subjects, and then divide the study population into a cohort of smokers, and another cohort of nonsmokers. They observe each cohort over time and compare the risks of incident kidney transplant failure between the cohorts.

Example 5.2. Investigators wish to ascertain whether a new antibiotic, supramycin, is associated with the development of a skin rash. They identify a group of patients who are prescribed antibiotic therapy for community-acquired pneumonia. Some patients are prescribed the new antibiotic supramycin, whereas others receive a different antibiotic. The investigators divide the study population into a cohort of supramycin users and a cohort of supramycin nonusers, and then compare the incidence of skin rash between the cohorts.

Cohort studies typically focus on incident, or new cases of disease that occur during follow-up. To accomplish this goal, investigators typically *exclude individuals who have prevalent disease at the beginning of a cohort study.* The evaluation of incident disease outcomes helps to establish that the exposure of interest *preceded* the outcome, and therefore might represent a cause of the disease. For the smoking and kidney transplant example, investigators would likely measure kidney function in all subjects at the beginning of the study, and then exclude those who have evidence of transplant failure. For the supramycin and rash example, investigators may require a baseline skin examination to exclude subjects who have a rash at the beginning of the study.

In cohort studies, researchers *observe* the exposure of interest. As a result, exposed and unexposed individuals may differ by characteristics other than the

exposure. For example, exposed subjects in the supramycin study (supramycin users) may differ from unexposed subjects (nonusers) by characteristics that influenced the decision to prescribe supramycin, such as the severity of pneumonia, practice patterns of the prescribing physician, and health insurance status. If these potential differences also influenced the risk of developing a rash, they could distort the study findings. In other words, a greater risk of drug rash among supramycin users might be caused by supramycin itself, or might be due to other characteristics that are linked with supramycin use, obscuring causal inference. The concept that factors other than the exposure may influence the study results is called confounding. Confounding is a major limitation of observational studies and is discussed in detail in Chaps. 9 and 10.

5.2 Ascertainment of Study Data

Once investigators decide on the specific exposures and outcomes to study, they should attempt to measure these characteristics using the most accurate methods available within their resources. For the smoking and kidney transplant failure study, possible choices to ascertain kidney transplant failure include the use of medical records from the transplant clinic, serum and urine markers of kidney function, a kidney transplant biopsy, and/or data from a national registry that tracks kidney transplant failures. Investigators must consider which of these sources provide the most valid measurements, and whether these data are uniformly available for all study participants. Kidney biopsy data may be the most accurate method for detecting kidney transplant failure; however, kidney biopsy data may be available for only a small number of study participants. For the supramycin and drug rash example, possible methods to ascertain supramycin use include the use of questionnaires, review of medical charts, and/or query of an electronic pharmacy database, if such data are readily available.

Important considerations in measuring study data are the *validity* of the measurements, *timing* of the measurements, and availability of *uniform* measurements among the study population.

5.2.1 Validity of Measurements

The validity of a measurement refers to how closely the measured data represent the true data. For the supramycin study, one possible method to ascertain whether or not a subject was using supramycin would be to transcribe the prescription bottle labels from all medications that were brought to a study examination (also known as the inventory method). The inventory method may not always yield a valid assessment of medication use, as demonstrated by the results from a study that compared medication information transcribed from study participants' prescription medication bottles to measured serum levels of four commonly used drugs (Table 5.1).[16]

Table 5.1 Agreement between inventory and serum detection of common medications

	Medication inventory report			
	Yes		No	
	Serum detection		Serum detection	
	Yes	No	Yes	No
Aspirin	15	36	7	44
Propanolol	21	28	0	49
Hydrochlorothiazide	37	13	1	49
Digoxin	47	3	0	50

These data demonstrate good agreement between medication inventory and serum detection of some medications, such as digoxin, but somewhat poor agreement for other drugs, such as aspirin. Many participants who were classified as aspirin users by the medication inventory method did *not* have a detectable serum level of aspirin, possibly because participants who use aspirin do not take it regularly. A possible response to these findings would be to add additional questions regarding aspirin use to study questionnaires in order to improve the validity of ascertaining aspirin use.

In some studies, the exposure of interest may be a biomarker, such as the serum cholesterol level or quantification of viral DNA. The validity of these exposure measurements will depend on the characteristics of the particular laboratory assay that was used to measure them.

5.2.2 Timing of Measurements

Optimal timing of exposure and outcome measurements will depend on the scientific question of interest. In some instances, investigators may be interested in the association between *recent* exposure and disease, particularly if there is concern that the exposure might change during the study. For example, nonsteroidal anti-inflammatory medications can cause rapid constriction of the renal arteries, leading to an increased risk of acute kidney injury. Investigators studying the relationship between nonsteroidal anti-inflammatory medication use and acute kidney injury might focus their efforts on obtaining a valid estimate of *recent* nonsteroidal anti-inflammatory medication use in relation to the occurrence of acute kidney injury. Ascertaining recent medication use may be accomplished by frequently updating each participant's medication status during the study. There are other examples in which the association of interest is the relationship between *long-term* exposure and disease. For example, smoking may damage a kidney transplant by slowly thickening the intimal and medial layers of the arteries within the transplant. Investigators in this study may therefore choose to evaluate *long-term smoking history*, such as the number of pack-years smoked after the transplant.

Cohort studies may not be ideal for studying very recent or very distant exposures. Distant exposures may change over time, obscuring their relationship with the outcome of interest. Very recent exposures may confuse the presumed temporal relationship between exposure and outcome, as demonstrated by the following example.

Example 5.3. Investigators wish to study whether smoking increases the risk of bacterial pneumonia. They recruit 1000 smokers and 1000 non-smokers and follow them for the future development of hospitalized bacterial pneumonia. Because smoking habits may change over time, investigators decide to update subjects' smoking status by querying admission records at the time of pneumonia hospitalization. Using this strategy to define the exposure, they observe paradoxically low rates of bacterial pneumonia among smokers.

How does the use of very recent smoking status in this example impact the study results? On one hand, smoking is suspected to *cause* bacterial pneumonia, because smoking destroys the natural ciliary defense mechanisms within bronchioles. On the other hand, symptoms of bacterial pneumonia may cause some smokers to quit, at least temporarily. In this example, it is likely that some of the smokers discontinued smoking when they developed pneumonia, leading to a distorted association when the most current assessment of smoking status is used. As a general rule, sufficient elapsed time between exposure and outcome are needed to prevent the outcome from being able to influence the exposure.

5.2.3 Uniform Measurements

The methods used to ascertain study data should not only be valid and timely, but also fair. That is, uniform measurement procedures should be utilized to ascertain study data across all subjects. The use of nonuniform data collection methods may not always be apparent in a clinical research study as illustrated by the following example.

Example 5.4. Investigators wish to study whether hepatitis C infection increases the risk for hospitalized major infections among HIV positive men. They recruit 1,000 men with HIV, measure their hepatitis C viral status using a sensitive and specific laboratory assay, and then maintain contact with three local hospitals to determine hospitalizations for major infections.

In this example, surveillance of only local hospitals may result in preferential capture of major infections among participants who are hepatitis C negative. Individuals who are hepatitis C positive may have additional problems such as homelessness and intravenous drug use that increase their tendency to migrate, and not be hospitalized locally. Preferential collection of fewer outcomes in hepatitis C positive individuals could lead to a spurious association of hepatitis C virus with *a lower* rate of hospitalization for major infections.

5.2.4 Retrospective Versus Prospective Data Collection

The terms 'retrospective' and 'prospective' typically refer to when the study data were collected in relation to the study investigators. For example, investigators conduct a study to examine risk factors for stroke. They abstract the medical records of individuals from a health care system who did not have a history of

stroke prior to the year 2000. They collect stroke risk factors as of January 1, 2000 and follow subjects for the development of incident stroke between the years 2000 and 2004. This study proceeds forward in time according to the principles of a cohort study; however, the data were collected retrospectively.

A prospective study involves the collection of new data, often for the purpose of addressing a specific question. For example, investigators recruit participants who do not have a prior history of stroke, collect blood for the measurement of a panel of novel serologic stroke markers, and then follow subjects for the development of incident stroke. This study also proceeds forward in time, according to the principles of a cohort study, but this time the data are collected prospectively. The distinction between retrospective and prospective studies is essentially descriptive; it does not impact the analysis plan or the study results.

5.3 Advantages of Cohort Studies

5.3.1 Study of Multiple Outcomes

Once investigators assemble the study cohorts, they are free to study more than one outcome, provided that the study subjects are free of each outcome of interest when the study begins. For the smoking and kidney transplant failure example, once cohorts of smokers and non-smokers are identified, investigators could also study whether smoking was associated with the development of kidney or bladder cancer, provided they exclude individuals who have evidence of these cancers at the beginning of the study.

One of the largest cohort studies even conducted was the Nurses Health Study, which recruited 127,000 nurses between the ages of 30 and 55.[17] The nurses completed questionnaires every 2 years that queried medical conditions, prescription and nonprescription medication use, social habits, dietary patterns, and physical activity. This open-ended cohort study design allowed for the creation of multiple cohort studies that evaluated risk factors for a wide range of diseases, including cancer, heart disease, and fractures.

5.3.2 Ability to Discern Temporal Relationship Between Exposure and Outcome

Unlike cross-sectional studies, cohort studies *can* reveal temporal relationships between the exposure and outcome, provided that a reasonable amount of time elapses between measurement of the exposure and the occurrence of the outcome. The existence of a temporal relationship strengthens evidence for the exposure to be a possible cause of the disease.

Recall the cross-sectional study from Chap. 4, in which higher serum homocysteine levels were associated with prevalent peripheral arterial disease. The cross-

sectional study design hindered interpretation of study findings because we were unable to distinguish whether higher serum homocysteine levels preceded the development of peripheral arterial disease or whether peripheral arterial disease preceded the higher serum homocysteine levels.

The cohort study approach to this problem would be first exclude subjects who have peripheral arterial disease at the beginning of the study, measure serum homocysteine levels, and then follow subjects for the development of incident peripheral arterial disease during follow-up. An association of higher serum homocysteine levels with incident peripheral arterial disease using a cohort study approach would strengthen the evidence for homocysteine as a *cause* of peripheral arterial disease. Note that establishing a temporal relationship between exposure and outcome is *inherent in the study design* and cannot be addressed by any statistical methodology.

5.4 Disadvantages of Cohort Studies

5.4.1 Confounding

Cohort studies are observational studies and are therefore subject to confounding. That is, other factors that are linked with the exposure of interest could account for some or all of the associations that are observed. For the homocysteine example, it is possible that other characteristics of people who have higher homocysteine levels will distort the association of higher homocysteine levels with incident peripheral arterial disease.

5.4.2 Inability to Examine Diseases That Are Rare or Have a Long Latency

Cohort studies may be inefficient if the *outcome is rare* or the *disease has a long latency period*.

Example 5.5. Investigators wish to investigate whether childhood seizures increase the risk of developing migraine headaches later in life. Using a cohort study approach they could recruit cohorts of children with and without seizure disorder, and then follow these cohorts for the subsequent development of migraine. However, many years of follow-up would likely be needed to observe migraine headache cases, which may not develop until the teenage years or early adulthood. In this example, the long latency period between exposure and outcome would make this study vastly expensive and logistically challenging.

Example 5.6. Investigators wish to investigate whether the prone sleeping position is a potential cause of sudden infant death syndrome (SIDS) in infants.

Using a cohort study approach they could identify a cohort of infants who sleep in the prone position and a cohort of infants who sleep in the supine position, and then compare the risks of SIDS between the cohorts. However, SIDS is a relatively rare outcome, necessitating recruitment of thousands of newborns to evaluate enough SIDS cases to draw meaningful conclusions. Rare outcomes are generally better evaluated using a case-control study design, which will be described in Chap. 6.

It is possible to overcome cohort study limitations of rare outcomes and long latency by using data sources that facilitate the study of large populations, as demonstrated in following example.

Example 5.7. Kidney disease leads to metabolic disturbances that cause bone disease. To investigate whether chronic kidney disease is associated with the occurrence of hip fracture, investigators used the electronic medical record system of a large health care network to identify three large cohorts of patients: one with normal kidney function, one with mild chronic kidney disease, and one with severe chronic kidney disease.[8] A total of 33,000 patients were available in the database for analysis; subjects with a previous history of hip fracture were excluded. During long-term follow-up, hip fracture incident rates were 1.0, 1.1, and 3.0 fractures per 1,000 person-years among patients with normal kidney function, mild kidney disease, and severe kidney disease, respectively. Although hip fracture is rare, the large number of individuals available for analysis permitted conduct of a successful cohort study.

5.5 Cohort Studies for Evaluating Medication Use

Cohort studies can be an important tool for studying the risks and benefits of medication use. Pharmacoepidemiology (observational studies of medication use) is gaining popularity as electronic pharmacy data systems become more widely accessible. These studies can provide specific information about medications that may not be easily obtained from randomized trials.[18]

First, cohort studies can assess the risks and benefits of medications among special populations that tend to be excluded from randomized trials, for example, patients with physical and mental disabilities, those with advanced kidney failure, or those with liver disease. Second, cohort studies can evaluate *clinical endpoints* and *unintended side effects* of approved medications by identifying very large cohorts of medication users from automated pharmacy data systems. For example, randomized trials initially demonstrated that peroxisome proliferator-activated receptor agonists effectively reduce glycosylated hemoglobin levels, a marker of glucose control, among people with diabetes.[19] Subsequent cohort studies that evaluated tens of thousands of existing peroxisome proliferator-activated receptor agonist users observed that these medications were associated with a *greater* risk of myocardial infarction.[20] The expense of randomized trials motivated the use of a biological endpoint, hemoglobin A1C levels, which may not have fully captured the diverse effects of the study medication.

The principal limitation of observational studies of medication use is difficulty separating the effect of a medication from the underlying characteristics of people

who tend to use that medication. For example, observational studies of hormone replacement therapy (HRT) found that women who used HRT experienced *lower* risks of cardiovascular events compared to women who did not. In contrast, when HRT or placebo was randomly assigned to participants in the Women's Health Initiative, HRT use caused *higher* risks of cardiovascular events.[21] A number of explanations have been proposed for these discrepant findings.[22] One possible explanation is that HRT users in the observational studies may have other characteristics that resulted in lower cardiovascular risk (better educated, more compliant with other prescribed medications).

A second disadvantage of pharmacoepidemiology studies is their tendency to evaluate chronic medication users rather than new users. This strategy can miss important early adverse reactions that occur after starting a new medication. For example, particularly high rates of acute thrombotic cardiovascular events were observed among HRT users at the beginning of the Women's Health Initiative study. It is hypothesized that HRT can trigger acute thrombotic events among a subset of women who have an underlying tendency toward blood clot formation. Evaluation of chronic HRT users in observational studies may have missed this important early phenomenon.

Despite their limitations, pharmacoepidemiology studies remain an important tool for evaluating medications that are used in clinical practice. Well-conducted pharmacoepidemiology studies use techniques to address potential differences between medication users and nonusers, focus on incident medication use, and rely on many of the analytic techniques that are used in clinical trials such as the principle of intention-to-treat analysis that is discussed in Chap. 6.

5.6 Analysis of Data From Cohort Studies

5.6.1 Incidence Proportion Versus Incidence Rate

Cohort studies usually involve two cohorts for comparison, but may have only one (with an informal comparison to other published cohorts or to historical data) or more than two (e.g., different groups of serum homocysteine levels or different doses of a medication). Regardless of the number of cohorts, we first calculate the incidence of disease in each cohort as either incidence proportion or incidence rate.

Hypothetical data for the first five smokers in the kidney transplant study are shown in Table 5.2.

Among this subsample of 5 smokers, there is 1 outcome (kidney transplant failure) that occurs in study subject number 4. The incidence proportion of kidney transplant failure in this subsample is 1 kidney transplant failure per 5 people, or 20%. This group contributes a total of 12.1 person-years at risk, obtained by calculating the sum of the follow-up times. The incidence rate of kidney transplant failure among this subsample of smokers is 1 kidney transplant failure per 12.1 person-years, or (1/12.1)*1,000 = 82.6 kidney transplant failures per 1,000 person-years.

Table 5.2 Follow-up data for first 5 smokers in the kidney transplant study

Subject #	Study start date	End of follow-up	Years followed	Reason for disenrollment
1	Jan 1, 2000	Jun 30, 2003	3.5	Study ended
2	Jan 5, 2000	Feb 1, 2002	2.1	Died
3	Feb 14, 2000	Jun 30, 2003	3.4	Study ended
4	Mar 2, 2000	Nov 15, 2000	0.7	Transplant failure
5	Mar 5, 2000	Aug 4, 2002	2.4	Lost to follow-up

Failure to account for person-time may lead to bias, which describes a systematic distortion of the true result. For example, suppose that smokers drop out of the study more frequently than nonsmokers (smoking may be an indication of noncompliance) such that smokers are followed for less time than nonsmokers. Under these conditions, it would be possible to observe a *lower* incidence proportion of disease among smokers simply because they are followed for less time. The inclusion of follow-up time in the denominator of incidence rate standardizes disease rates by time across cohorts, eliminating this potential problem.

5.6.2 Relative Risk

Once we determine the incidence of disease in each study cohort, we can next compare these incidences. Many types of comparisons are possible; the two most common are *relative risk* and *attributable risk.*

The relative risk is defined as the incidence (proportion or rate) in one cohort *divided* by the incidence in the *reference cohort*. For example, the incidence rate of kidney transplant failure is 82.6 failures per 1,000 person-years among smokers and 55.3 failures per 1,000 person-years among nonsmokers. Using nonsmokers as the reference cohort, the relative risk would be:

$$\frac{82.6\,\text{failures per}\,1{,}000\,\text{person - years}}{55.3\,\text{failures per}\,1{,}000\,\text{person - years}} = 1.49(\text{comparing smokers to nonsmokers})$$

The relative risk can also be expressed using smokers as the reference cohort:

$$\frac{55.3\,\text{failures per}\,1{,}000\,\text{person - years}}{82.6\,\text{failures per}\,1{,}000\,\text{person - years}} = 0.67(\text{comparing nonsmokers to smokers})$$

Both relative risks are correct. You can pick either cohort to be the reference group. Note that relative risk has no units.

The interpretation of the first relative risk of 1.49 would be, “kidney transplant recipients who smoke are 49% more likely to develop transplant failure compared to nonsmokers.” We say “49% more likely” because 1.49 is 49% higher than 1.0,

which is the value that indicates equal risk between groups. Another equally correct interpretation of these findings would be, "smoking is associated with a 1.49-fold greater risk of kidney transplant failure."

An interpretation of the second relative risk would be, "nonsmokers with a kidney transplant are 33% *less* likely to develop transplant failure compared to smokers." The second relative risk demonstrates an association with a lower risk of disease, because the value is less than 1.0.

Why is smoking associated with a 49% greater risk of transplant failure, but nonsmoking with only a 33% lower risk of transplant failure? In other words, why are the relative risks not additively symmetrical for the same exposure? Note that relative risks, like all ratios, can assume possible values ranging from 0 to infinity; however, 1.0 defines the unity value. It is harder to obtain relative risks that are much less than 1.0 because they are bounded at 0. For this reason, relative risks that are less than 1.0 indicate stronger associations than symmetrical associations that are greater than 1.0.

Note that we have not attempted to account for potential differences in characteristics between smokers and nonsmokers, such as age or race, but instead calculated a "raw" relative risk from the data. This "raw" relative risk is called a "crude relative risk," or "unadjusted relative risk."

How are relative risks calculated for multiple cohorts? Consider the following data:

Cohort	Incidence rate of kidney transplant failure
Nonsmoker	55.3 per 1,000 person-years
Smoke 1–10 cigarettes per day	60.1 per 1,000 person-years
Smoke 11–20 cigarettes per day	78.7 per 1,000 person-years
Smoke more than 20 cigarettes per day	123.0 per 1,000 person-years

To obtain relative risks for a study with multiple cohorts, a reference cohort must be selected. The reference cohort that makes the most sense in this case is the nonsmokers. We can calculate a relative risk for each exposed cohort in relation to the reference cohort as follows:

Cohort	Crude relative risk of kidney transplant failure
Nonsmoker	Reference cohort: relative risk = 1.0
Smoke 1–10 cigarettes per day	60.1/55.3 = 1.09
Smoke 11–20 cigarettes per day	78.7/55.3 = 1.42
Smoke more than 20 cigarettes per day	123.0/55.3 = 2.22

In some instances, the reference cohort may be selected to be intermediate values of a particular exposure. For example, a cohort study evaluated the association of serum retinol levels with fracture risk among European men.[23] Study investigators considered both retinol deficiency and retinol excess to be potentially harmful, motivating selection of individuals with intermediate serum retinol values as the reference cohort.

Serum retinol level (μmol/liter)	Unadjusted relative risk of fracture
<1.95	1.1
1.95–2.16	0.8
2.17–2.36	reference (1.0)
2.37–2.64	1.0
>2.64	1.7

On the basis of these data, serum retinol values >2.65 μmol/liter are associated with an estimated 70% greater risk of fracture *compared to serum retinol values between 2.17 and 2.36 μmol/liter.*

5.6.3 Attributable Risk (also Called "Risk Difference" or "Excess Risk")

The attributable risk is defined as the incidence (proportion or rate) in one cohort *minus* the incidence in another cohort. For the smoking and kidney transplant example, the attributable risk of transplant failure comparing smokers to nonsmokers would be:

82.6 transplant failures per 1,000 person-years (smokers)–55.3 transplant failures per 1,00 person-years (nonsmokers)=27.3 transplant failures per 1,000 person-years

Note that attributable risk, unlike relative risk, maintains the same units as the incidence values that are being compared.

If the exposure of interest causes the disease, then the attributable risk describes the amount of *additional* or *extra* risk that is due to the exposure. For example, if smoking truly causes kidney transplant failure, then the interpretation of the attributable risk for smoking would be, "there are 27.3 additional kidney transplant failures per 1,000 person-years among transplant recipients who smoke." Stated another way, "smokers with a kidney transplant incur an estimated 27.3 extra kidney transplant failures per 1000 person-years."

Relative and attributable risks provide complementary information. The relative risk is one tool that is used to evaluate whether the exposure may be a *cause* of the outcome. In the above example, smoking more than 20 cigarettes per day is associated with a 2.2-fold greater relative risk of kidney transplant failure. This *strong association*, combined with the observed temporal sequence (smoking precedes the development of transplant failure), dose–response (graded increased risk of transplant failure with more cigarettes smoked), and biological plausibility (smoking damages blood vessels of transplanted organs) help support the case for smoking as a *cause* of kidney transplant failure. If smoking does actually cause kidney transplant failure, then the attributable risk describes the *impact of smoking* on the kidney transplant population. Both measures of risk are correct, but have different interpretations.

Chapter 6
Case-Control Studies

Learning Objectives

1. Case-control studies work backwards, first identifying diseased and nondiseased individuals, and then ascertaining the frequency of previous exposures.
2. Ideal characteristics for selecting cases:
   a. Use a specific definition of the disease
   b. Select individuals who have incident disease
3. Ideal characteristics for selecting controls:
   a. Controls should derive from the same underlying population as the cases
   b. Controls should have the opportunity to be counted as cases if they develop disease
4. Case-control studies have certain advantages:
   a. Can be useful for studying rare outcomes and those with long latency periods
   b. Can be useful for evaluating multiple risk factors for a disease
5. Case-control studies have certain disadvantages:
   a. Observational design: other factors may be responsible for observed associations
   b. Recall bias may distort associations
6. Case-control studies can estimate only the relative risk of disease; the incidence and attributable risk cannot be determined from case-control data alone.
7. The primary measure of effect in a case-control study is the odds ratio.
8. Odds ratios approximate relative risks when the outcome is rare.

To introduce the concept of case-control study design, we consider a new question: Does the measles-mumps-rubella (MMR) vaccine cause pervasive developmental disorders such as autism in children? This question arose from case reports describing children who presented with a major developmental disorder following MMR vaccination. The case series data influenced public behavior; MMR vaccination rates declined modestly and measles outbreaks occurred.

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_6, © Springer Science+Business Media, LLC 2009

First consider the cohort study approach to this problem. Investigators could identify a cohort of 1,000 children who receive the MMR vaccine and another cohort of 1,000 children who do not receive the vaccine. After excluding children with developmental disorder at the beginning of the study, investigators would observe each cohort for the development of incident developmental disorder during follow-up. Data obtained using this approach are presented in Table 6.1 below.

Table 6.1 MMR vaccination status and developmental disorder: Cohort study approach

	Pervasive developmental disorder		
MMR vaccination	Yes	No	Total
Yes	5	995	1,000
No	6	994	1,000

These data do *not* suggest a difference in pervasive developmental disorder, comparing children who receive MMR vaccination to those who do not; however, there are too few cases of developmental disorder in this study to draw meaningful conclusions. In this example, the cohort study design is inefficient because the outcome occurs infrequently.

What if the cohort study design was turned around? Instead of starting with cohorts of exposed (MMR-vaccinated) and unexposed (not MMR-vaccinated) children, what if the investigators began by identifying children with and without the *outcome of interest*, developmental disorder?

In one such study, investigators identified 1,300 children with pervasive developmental disorder using a national health care database.[24] They used the same database to determine whether these case children had received MMR vaccination *prior to the development of developmental disorder*.

Table 6.2 MMR vaccination status among 1,300 children with developmental disorder

Previous MMR vaccination	
Yes	1,000 (77%)
No	300 (23%)
Total	1,300

The data in Table 6.2 demonstrate a seemingly high proportion (77%) of previous MMR vaccination among children who have pervasive developmental disorder. The next step is to compare this proportion to that of similar children who do *not* have developmental disorder. Investigators used the same national health database to ascertain the following data.

The data in Table 6.3 demonstrate that the proportion of MMR vaccination among children without pervasive developmental disorder (79%) is similar, in fact

Table 6.3 MMR vaccination status among 4,500 children without developmental disorder

Previous MMR vaccination	
Yes	3,550 (79%)
No	950 (21%)
Total	4,500

slightly higher, than that of children with developmental disorder. The study findings are best appreciated by presenting the disease and nondisease data together.

The data in Table 6.4 illustrates a case-control study design, which begins with diseased and nondiseased individuals and then ascertains their exposure status *prior to developing the disease*. Compared to cohort studies, case-control studies seem to work backwards; however, properly conducted case-control studies can be used to suggest important causal relationships.

Table 6.4 MMR vaccination status and developmental disorder: Case-control approach

Previous MMR vaccination	Developmental disorder (N = 1,300)	No developmental disorder (N = 4,500)
Yes	1,000 (77%)	3,550 (79%)
No	300 (23%)	950 (21%)

6.1 Case-Control Study Design

6.1.1 Overview

The fundamental design of a case-control study is:

1. Identify cases and controls
2. Ascertain the frequency of exposure(s) among each group

Cases refer to people who have developed the *disease or outcome* in question.

Controls refer to people who do not have the *disease or outcome* in question, and who are selected to estimate the frequency of the exposure in the population.

The main distinction between a case-control study and a cohort study is that the *cohort studies identify subjects based on their exposure status*, whereas *case-control studies identify subjects based on their outcome status.*

Like cohort studies, case-control studies can reveal temporal relationships between exposure and outcome, strengthening the case for a causal relationship. Temporality in case-control studies is demonstrated by assuring that an exposure was present prior to the development of the disease. In the MMR example, investigators determined whether study children had *previously received* the MMR vaccine *before the development of pervasive developmental disorder*.

Careful selection of cases and controls is a critical first step for successful conduct of case-control studies.

6.1.2 Selection of Cases

6.1.2.1 Use a Specific Definition of Disease

A highly *specific* definition of disease is generally desired in case-control studies to ensure that the disease in question is truly present among individuals who are defined as cases. This strategy may require exclusion of individuals with milder forms of a disease in order to focus on more advanced cases, which can be diagnosed with greater certainty. For example, a highly specific definition of developmental disorder might require at least two diagnosis codes for developmental disorder plus confirmatory results of formal testing (if available). The use of a specific definition increases the likelihood that case children actually have pervasive developmental disorder, at the expense of possibly missing some children who have milder forms of the disease. Turning to a different example, investigators conduct a case-control study to investigate whether air travel increases the risk of developing deep venous thrombosis. The investigators may require definitive ultrasound evidence of deep venous thrombosis to define a case and exclude individuals who have only intermediate or suggestive ultrasound findings, because inadvertent inclusion of nondiseased individuals as cases will greatly diminish the ability of a case-control study to detect associations.

In contrast, it is generally less important to devote extraordinary study resources toward confirming that control subjects are truly free of the disease in question. Case-control studies usually investigate rare diseases such as vasculitis, brain cancer, or aplastic anemia, which are unlikely to be present in randomly selected control individuals. Even in the absence of screening, the vast majority of controls are likely to be free of these diseases based on their absolute rarity.

6.1.2.2 Selection of Incident Cases

Typically, the goal of case-control studies is to study the *development of disease*; therefore, incident (new) cases of disease are generally preferred to chronic or longstanding cases. One reason for focusing on incident disease is to establish that the exposure of interest was clearly present before the disease occurred. For example, selection of children with longstanding developmental disorder would complicate documentation of MMR vaccination prior to the development of the disorder. A second reason for choosing incident cases is that the alternative, selecting chronic cases, will intertwine the development of disease with survivorship. To illustrate this concept, consider a case-control study that evaluates whether serum oxidative stress markers are related to the development of stroke. Investigators begin by identifying

individuals with stroke (cases) and those without stroke (controls), and then measure oxidative stress markers from serum samples that were collected 10 years earlier. If investigators select cases to be individuals with chronic stroke, they would be studying *stroke survivors*, whose survival may be linked with healthy characteristics, including lower levels of oxidative stress markers. The result may be an artificial association of oxidative stress markers with a *lower* risk of stroke.

6.1.3 Selection of Controls

6.1.3.1 Select Controls from the Same Underlying Population as the Cases

Case-control studies compare the frequency of exposure among individuals who have a disease to that of individuals who do not have the disease. Interpretation of case-control data hinges on the assumption that a fair control population was selected to estimate the frequency of exposure.

The general goal is to obtain control subjects who derive from the *same underlying population* as the case subjects. In the MMR vaccine example, children with developmental disorder were selected from the national health system in Great Britain, where MMR vaccination is relatively common. If investigators instead selected control children from a different country, where MMR vaccination is less common, they would have observed a higher proportion of MMR vaccination among the case children, leading to the spurious conclusion that MMR vaccination is more common among children who have pervasive developmental disorder.

6.1.3.2 Controls Should Have the Same Opportunity to Be Counted as Cases should the Disease Occur

It is important to ensure that control subjects have the opportunity to be counted as cases should the disease or outcome in question develop. In the MMR study, control children had the same opportunity to be diagnosed with developmental disorder as case children because they were part of the same health system that captured developmental disorder using diagnosis codes.

Equal access to diagnosis is not always a simple consideration. Consider a different example, in which investigators explore whether Chinese herbal medications are a novel cause of kidney dysfunction. Investigators examine the laboratory database from a University hospital to identify patients who have evidence of kidney dysfunction. Investigators then contact potential case subjects, administer a questionnaire, and determine that 5% used Chinese herbs within the previous year. How should the investigators select a comparable control population?

Because kidney dysfunction may not cause symptoms, laboratory testing is needed to confirm the diagnosis. For this reason, more frequent laboratory testing will lead to a greater opportunity to be diagnosed with kidney disease. Choosing control individuals

from the general population may not be ideal, because many people in the general population do not undergo routine laboratory testing, and therefore would not have the opportunity to be diagnosed with kidney disease if it were present. A more suitable control population in this example might be people who visit the same University hospital and who also receive laboratory testing.

One method to ensure that cases and controls derive from the same underlying population and have the same opportunity to be diagnosed with a disease is a *nested case-control study*. This study design selects cases and controls from within a larger cohort study. For example, the Cardiovascular Health Study recruited 5,800 older adults from four communities and obtained serological measurements of kidney function among all participants.[25] A nested case-control study within CHS could readily identify case subjects with kidney dysfunction and control subjects with normal kidney function based on laboratory data that were obtained using identical methods. Investigators could then try to estimate the frequency of previous Chinese herb use among cases and controls using previously collected non-prescription medication data.

6.1.3.3 Matching

Case-control studies often use matching to increase the degree of similarity between case and control subjects. In the MMR study, an example of matching would be to first select a child who has pervasive developmental disorder, and then identify a control child who does not have developmental disorder, but is the same age and sex as the case child. Using appropriate analytic techniques, matching can reduce the possibility that other factors account for the association between exposure and outcome (confounding).

However, it is sometimes possible to "overmatch" in case-control studies if the matching factor happens to be related to the exposure of interest. For example, if case children with developmental disorder are additionally matched to control children according to parental beliefs toward vaccination (support vaccination *versus* suspicious of vaccination) then this matching process could artificially diminish contrasts in the proportion of MMR vaccination between children with and without developmental disorder.

6.1.3.4 Number of Controls

In a case-control study, the disease of interest is usually rare, so finding cases is often the rate-limiting step. There are no specific rules regarding the number of controls that are needed per case; however, more controls will generally provide a more accurate estimation of exposure frequency in the control group and can increase study power (the ability to detect an association if one is truly present). Study resources usually determine the number of controls that can be selected per case. There is a fairly steep increase in study power as more controls are added until reaching about three to four controls per case, at which point adding more controls has little further effect on study power.

6.2 Advantages of Case-Control Studies

6.2.1 Case Control Studies Can Be Ideal for the Study of Rare Diseases or Those with a Long Latency

Cohort studies and randomized trials may be difficult to perform when the outcome of interest is rare or the latency period between exposure and outcome is long. For the MMR vaccine example, there were only 11 cases of pervasive developmental disorder among 2,000 children using the cohort approach described at the beginning of this chapter. The case-control approach, which directly targets developmental disorder, quickly identified 1,300 children with the disease.

Case-control studies may be useful for studying processes in which the time period between exposure and disease development is particularly long, if previous exposure data is available or can be readily ascertained. For example, it may take years for certain dietary factors, such as fish oil, to produce cardiovascular benefits. A case-control study could readily identify individuals with and without incident coronary heart disease, and then question them regarding the frequency and amount of previous fish oil consumption.

6.2.2 Case-Control Studies Allow for the Study of Multiple Exposures

Recall that cohort studies identify subjects based on their exposure status, and then follow the cohorts for the development of different outcomes. In contrast, case-control studies identify subjects based on their disease status, permitting the study of multiple *exposures* within a predefined group of cases and controls. For example, once cases of pervasive developmental disorder and controls without developmental disorder are identified, study investigators could explore other risk factors for developmental disorder, if such exposures can be accurately ascertained. Some examples of additional exposure variables are presented in Table 6.5.

These data demonstrate a greater frequency of a family history of autism, but not prematurity, or exposure to pesticides, comparing children with developmental disorder to those without it.

Table 6.5 Association of multiple exposures with developmental disorder

	Developmental disorder (%) (N = 1,300)	No developmental disorder (%) (N = 4,500)
Previous MMR vaccination	77	79
Family history of autism	27	13
Prematurity	11	12
Exposure to pesticides	2.0	1.4

6.3 Disadvantages of Case-Control Studies

6.3.1 Observational Study Design

Like cohort studies, case-control studies are observational study designs and may be subject to confounding. Cases may differ from controls by factors other than the exposure of interest. Confounding occurs when a factor other than the exposure of interest distorts the association between exposure and outcome, thereby limiting inference that the exposure causes the disease.

6.3.2 Recall Bias

Like cohort studies, case-control studies may ascertain study data using a variety of sources, including medical records, questionnaires, interviews, and laboratory measurements. Like cohort studies, case-control studies strive for valid, precise, and uniform measurements of the exposure and outcome. An important additional consideration for case-control study measurements is the use of interviews or questionnaires to ascertain previous exposure status, because these procedures can lead to a specific type of bias known as *recall bias*. Recall bias occurs when case individuals, who tend to be sick, and control individuals, who tend to be generally well, recall their exposure status differently. For example, if MMR vaccination status was ascertained by interviewing the parents of children with and without pervasive developmental disorder, it is possible that the parents of children recently diagnosed with developmental disorder may overreport previous MMR vaccination, particularly if they have preconceived ideas about vaccination safety. Systematic overreporting of MMR vaccination among the cases, but not controls, could lead to a spurious association of MMR vaccination with developmental disorder.

The ideal solution to recall bias in case-control studies is to use data that were collected systematically, prior to the development of disease. For example, in the MMR study, investigators used data from a national health care database to ascertain MMR vaccination status prior to the occurrence of developmental disorder. For a second example, consider a case-control study of HIV seroconversion following occupational exposure to HIV infected blood.[26] Study investigators identified case health care workers who seroconverted to HIV following accidental needlestick injury and control health care workers who did not seroconvert following a needlestick injury. If investigators contacted these individuals to inquire about previous characteristics of their needlestick exposure, such as “were you wearing gloves?” or “did you see a large amount of blood on the needle?,” case individuals with newly diagnosed HIV may recall exposures differently than non-HIV controls. Instead, the study investigators collected exposure data from mandatory occupational injury reports that were completed by cases and controls at the time of their injury, before they were aware of their future HIV seroconversion status.

6.3.3 Case Control Studies only Provide Information Regarding the Relative Risk (Odds) of Disease

Recall that cohort studies can determine the incidence of disease among exposed and unexposed individuals, and then compare these incidences using a ratio (relative risk) or a difference (attributable risk). As we well see in Sect. 6.4, *case-control studies can provide only an estimate of relative risk*. They cannot be used to calculate attributable risk, nor can they be used to calculate the specific incidence of disease in any group.

6.4 Analysis of Case-Control Data

6.4.1 Theory of the Odds Ratio

The MMR case-control data demonstrated that 77% of children with pervasive developmental disorder previously received MMR vaccination, and that 79% of children without developmental disorder previously received MMR vaccination. However, we are not usually interested in determining exposure frequencies among people with or without a particular disease; a parent of a child with developmental disorder would generally not be interested in knowing that their child has a 77% chance of previously receiving MMR vaccination. Instead, we are interested in the opposite association, "what is the risk of developmental disorder comparing children who receive MMR vaccination to those who do not?" We can estimate this information from case-control data using a term called the *odds ratio*.

To develop the concept of the odds ratio, imagine that we have unlimited funding and resources to study children in the British healthcare system. We conduct a massive *cohort study* of MMR vaccination that obviates the rare outcome problem. Hypothetical data are presented in Table 6.6.

After recruiting more than 200,000 children into the hypothetical cohort study, we can calculate the incidence proportion of pervasive developmental disorder in each cohort:

$$\text{Incidence proportion, MMR-vaccinated children} = \frac{1000}{161,000} * 100\% = 0.62\%$$

$$\text{Incidence proportion, non-MMR - vaccinated children} = \frac{300}{45,300} * 100\% = 0.66\%$$

Since we are looking at a hypothetical cohort study, we next use the incidence proportion data to calculate the *relative risk of developmental disorder* = (0.62/0.67) = 0.93. An interpretation of the relative risk is, "MMR vaccination is associated

Table 6.6 Cohort study of MMR vaccination status with developmental disorder

Previous MMR vaccination	Developmental disorder	No developmental disorder	Total
Yes	1,000	160,000	161,000
No	300	45,000	45,300
Total	1,300	205,000	206,300

Table 6.7 Case-control study of MMR vaccination status with developmental disorder

Previous MMR vaccination	Developmental disorder (N = 1,300)	No developmental disorder (N = 4,500)
Yes	1,000 (77%)	3,550 (79%)
No	300 (23%)	950 (21%)

with a 7% *lower* risk of developmental disorder." Now let's return to the original case-control data, presented in Table 6.7.

Comparing the cohort and case-control data, we see that the 4,500 controls represent only a fraction of the 205,000 children who do not have developmental disorder. In the case-control study, investigators selected 4,500 control children based on their study resources, power considerations, and practicality. This "arbitrary" selection of 4,500 controls makes it impossible to calculate the true incidence of developmental disorder among children who are exposed or not exposed to the MMR vaccine. We *cannot* state that the incidence proportion of pervasive developmental disorder among MMR-vaccinated children is 1,000/4,550 = 22%; this proportion is many times greater than the actual incidence proportion of 0.63%, and would differ depending on the investigator's particular choice of the number of controls. Similarly, we cannot state that the incidence proportion of developmental disorder among the non-MMR-vaccinated children is 300/1,250 = 24% for the same reason.

To estimate relative risk in case-control studies, we make the assumption that although the specific incidences of disease for exposed and unexposed individuals are (often blatantly) wrong, they are proportionally wrong. The "false" incidence proportion of developmental disorder for MMR-vaccinated children using the case-control data is 22%, which is about 35-fold higher than the true incidence proportion of 0.63% derived from the cohort data. The "false" incidence proportion of developmental disorder for non-MMR-vaccinated children using the case-control data is 24%, which is about 36-fold higher than the true incidence proportion of 0.66%. Because each incidence estimate errs by about the same degree, *the ratio of these false incidences will approximate the relative risk of disease*. This ratio of false incidences is not really a relative risk, and instead has a different name, the *odds ratio*. *The odds ratio is the principal measure of risk in a case-control study.*

6.4.2 Practical Calculation of the Odds Ratio

In reality, the odds ratio is calculated using the number of controls (people without disease) as the denominator. Given the MMR case-control data:

		Developmental disorder (N = 1300)	No developmental disorder (N = 4500)
Previous MMR vaccination	Yes	1000 (77%)	3550 (79%)
	No	300 (23%)	950 (21%)

The odds ratio is calculated as (1,000/3,550)/(300/950) = 0.89.

Equally correct interpretations of this odds ratio include:

1. MMR vaccination is associated with 11% lower odds of developmental disorder.
2. The odds of developmental disorder are 11% lower among children who received MMR vaccination compared to those who did not.
3. The odds ratio of developmental disorder is 0.89, comparing children who received MMR vaccination to those who did not receive MMR vaccination.

A simple method exists for calculating the odds ratio from case-control data. First, set up a data table with disease versus no disease on the top and exposed versus unexposed on the left side. Given this setup, the cells of the table are referred to as *a*, *b*, *c*, and *d*, as shown in the table below. The odds ratio is calculated from this table as (*a***d*)/(*b***c*).

		Developmental disorder (N=1300)	No developmental disorder (N=4500)
Previous MMR vaccination	YES	*a.* 1000	*b.* 3550
	NO	*c.* 300	*d.* 950

The odds ratio is calculated as (1,000*950)/(3,550*300) = 0.89.

Note that we have not performed any adjustment for other study factors such as age, race, or sex. As a result, this odds ratio is also called a "crude odds ratio," or "unadjusted odds ratio."

6.4.3 Odds Ratios and Relative Risk

Using the cohort study approach, we found that the *relative risk* of pervasive developmental disorder, comparing MMR-vaccinated with non-MMR-vaccinated children, was 0.93. Using the case-control approach, we found that the *odds ratio* of pervasive

developmental disorder, comparing MMR-vaccinated with non-MMR-vaccinated children, was 0.89. These estimates are similar, but not exactly the same.

The major factor that determines the amount of agreement between the relative risk and the odds ratio is the rarity of the outcome in question. The *lower* the prevalence of the outcome in the population, the closer the agreement between the relative risk and the odds ratio. In the MMR example, pervasive developmental disorder is relatively rare in the population; there are only 1,300 cases among 206,300 children (prevalence = 0.63%). There is no specific cutoff value to define "rare," but generally case-control studies of diseases with a prevalence <5% will yield odds ratios that closely approximate the relative risk.

If the disease in question is rare, such that the odds ratio approximates the relative risk, we can substitute the term "risk" for "odds" and the term "relative risk" for "odds ratio" when interpreting odds ratios in case-control studies. For the MMR example, we can state that "the *relative risk* of developmental disorder is 0.89, comparing children who received MMR vaccination to those who did not receive the vaccine."

6.4.4 Case-Control Studies Cannot Estimate the Actual Incidence of a Disease or Outcome

The MMR case-control data demonstrate that *relative to a child who does not receive MMR vaccination*, a child who receives MMR vaccination has an estimated 11% lower risk of developing pervasive development disorder. Case-control data can inform only the *relative chance of an outcome comparing one group to another.* Case-control data alone *cannot* be used to calculate the specific risks of pervasive development disorder, whether a child is vaccinated or not. Data from other studies are needed to provide information about the specific incidence of pervasive developmental disorder among vaccinated and nonvaccinated children.

To further illustrate this point, consider another case-control study that evaluated whether unprofessional behavior during medical school is related to future disciplinary action by a state medical board.[27] Discipline by a state medical board tends to be a fairly serious action that can result from drug or alcohol abuse, negligent professional behavior, or worse.

Study investigators selected 235 case physicians who were disciplined by a state medical board and 469 control physicians who were not disciplined, and then went back and examined their medical school records. The primary finding was that an episode of unprofessional behavior documented during medical school was associated with an odds ratio of 3.0 for future disciplinary action by a state medical board.

The strict interpretation of these case-control data is that physicians who are disciplined by a state medical board are 3 times more likely to have committed an act of unprofessional behavior during medical school. Because the odds ratio can be transposed, these data can also be interpreted as, "a student who commits an act of unprofessional

behavior during medical school has a 3-fold greater chance of being disciplined by a state medical board *compared to another medical student who does not commit an act of unprofessional behavior*." Notice that we substitute the word 'chance' for 'odds' in this example because the outcome, disciplinary action by a state medical board, is rare.

How should this finding be applied to an individual medical student who is cited for an episode of unprofessional behavior? On the one hand, the case-control data demonstrate that this student is three times more likely, on average, to incur future disciplinary action by a medical board, relative to his/her classmate who is not cited for unprofessional behavior. On the other hand, discipline by a state medical board is an extraordinarily rare occurrence for physicians. Other studies indicate that disciplinary action by a state medical board occurs in <0.3% of physicians. Even a threefold greater risk due to an episode of unprofessional behavior during medical school yields a less than 1% chance that this rare outcome will actually occur.

Chapter 7
Randomized Trials

Learning Objectives

1. Randomized trials should be considered when:
   a. There is uncertainty regarding the effect of an exposure or treatment
   b. The exposure can be modified in a trial setting
2. Phase I and phase II studies evaluate the tolerability and biological activity of a drug; phase III and phase IV studies are randomized trials that evaluate clinical endpoints.
3. Potential limitations of randomized trials include:
   a. Limited generalizability of the study population
   b. Limited generalizability of the study environment
   c. Randomized trials address a narrow study question
   d. Randomized design accounts only for confounding
4. Common measures of effect in randomized trials are relative risk and risk difference.
5. The number needed to treat or harm = 1/risk difference.
6. The intention-to-treat analysis predictably leads to bias toward the null.
7. Criteria used to judge whether results of a subgroup analysis are valid include:
   a. Biological plausibility for a particularly strong effect in the subgroup
   b. The subgroup analysis was pre-specified
   c. Reasonably large number of outcomes in the subgroup

A randomized trial is a *prospective* study *in humans* comparing the effect and value of an *intervention* against a *control*.

7.1 Rationale for Randomized Trials

Observational studies are often plagued by the problem of confounding. Confounding occurs when differences in the characteristics of exposed and unexposed individuals obscure the central question of whether an exposure *causes* a disease. Implications of confounding range from annoyance to a fatal flaw. Consider the following two examples are 7.1.1 and 7.1.2.

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_7, © Springer Science+Business Media, LLC 2009

7.1.1 Kidney Transplant and Mortality

The progressive loss of kidney function leads to end stage kidney disease, which requires chronic dialysis or a kidney transplant for survival. Kidney transplantation is considerably more effective than dialysis for clearing metabolic waste products and retained fluid caused by kidney failure. In one observational study, investigators evaluated a group of chronic dialysis patients who had been placed on the kidney transplant waiting list.[28] During the study, some of these patients received a transplant, whereas others remained on dialysis. By the end of the study, patients who received a kidney transplant had a 70% lower mortality rate compared to those who remained on dialysis.

The hypothesis that kidney transplantation favorably influences survival is supported by this strong observed association and biological plausibility. However, it remains possible that other characteristics of patients receiving a kidney transplant, and not transplantation itself, are responsible for the observed survival advantage. To disentangle the effects of transplantation from the health status of transplanted individuals, the authors cleverly limited their analyses to patients who were on the kidney transplant waiting list, and therefore were healthy enough to receive a transplant. Still, the best evidence to support the hypothesis that kidney transplantation itself improves survival would come from a clinical trial that randomly assigns chronic dialysis patients to receive a kidney transplant or to remain on dialysis. Such a trial would be unethical because the observational data already provide compelling evidence for the benefits of kidney transplantation.

7.1.2 Angioplasty versus Fibrinolysis for Patients with Acute Myocardial Infarction

Two effective therapies for heart attack (acute myocardial infarction) are angioplasty, a procedure that quickly opens blocked arteries within the heart, and fibrinolysis, the administration of a specialized medication that rapidly dissolves blood clots. Patients with acute myocardial infarction typically receive angioplasty if they present to a hospital that has the technical capacity to perform this procedure, or fibrinolysis if the hospital is less technically equipped.

Observational studies that compared early angioplasty versus fibrinolysis for acute myocardial infarction reported inconsistent results. These studies were limited by potential differences in the underlying characteristics of patients who received angioplasty versus those who received fibrinolysis, for example, differences in access to health care, health insurance status, general compliance with medical care, socioeconomic background, and lifestyle factors. Although observational studies can attempt to measure and adjust for all of the possible differences between patients who received angioplasty versus those who received fibrinolysis, many of these characteristics will be difficult to ascertain reliably. Given ambiguous

clinical implications of these two effective therapies, and difficulty separating the effects of treatment from the characteristics of people who receive these treatments, observational studies will leave lingering uncertainty as to whether a specific treatment itself is responsible for potential differences in clinical outcomes. To address this question, investigators recruited more than 1000 patients with acute myocardial infarction who presented to a hospital without the technical capacity to perform angioplasty. Half of these patients were randomly assigned to immediate transfer to an outside facility for angioplasty, and the other half received fibrinolysis.

As a general rule, randomized trials should be considered when (1) there is continued uncertainty as to the effect of an exposure or treatment (equipoise, see below) and (2) the exposure can be modified in a trial setting. Nevertheless, evidence for a causal effect may be inferred from well-conducted observational studies that include a reasonably large number of outcomes, use appropriate analyses that account for relevant, well-measured risk factors, and observe a large magnitude of effect (relative risks that are >2.0 or <0.5). While these are useful general rules, there are many examples of exposures that have been associated with benefit in well-conducted observational studies, but not in randomized trials. Therefore, caution should always be used when interpreting results of observational studies with regard to treatment effect.

7.1.3 Equipoise

In order to conduct a randomized trial ethically, one cannot assign people to one that is known to be inferior to another treatment. There must be true uncertainty as to which treatment is actually more beneficial. *Equipoise* is defined as the point in which a rational, informed person has no preference between two (or more) available treatments. For example, investigators cannot ethically randomize a person with high cholesterol levels to receive a cholesterol-lowering medication versus placebo because there is already a body of evidence demonstrating that (some) cholesterol medications improve survival. An ethical randomized trial might compare a new cholesterol medication with the current standard of care, such as another cholesterol medication that is approved and in use.

7.2 Phases of Drug Development

Randomized trials of medications often represent the result of a long drug development process that proceeds in an organized series of steps or phases. Phase I and II studies are typically *not* randomized trials but are used to develop randomized phase III and IV studies.

7.2.1 Phase I Studies

The first step in developing a new drug is to understand *how well the drug is tolerated* in a small number of people. Although not a clinical trial, these types of studies are referred to as *phase I studies*. Participants in phase I studies are either healthy adults or people with the specific disease that the drug is meant to modify. Occasionally, phase I studies cannot be performed in healthy adults because the drug has unacceptable adverse effects, such as a chemotherapeutic agent. Phase I studies seek to determine how large a dose can be given before unacceptable toxicities occur. Phase I studies start with low doses in limited numbers of people and then increase the dosage incrementally. Ultimately, phase I data are used to inform dosing in phase II studies.

7.2.2 Phase II Studies

Phase II studies are designed to evaluate whether a drug has *biological activity* and to determine safety and tolerability. In phase II studies, participants are assigned to one of at least two different medication doses, determined from phase I studies. Phase II studies often use *surrogate markers of biological activity*, for example, LDL cholesterol levels or bone mineral density, to determine drug efficacy, rather than clinical events, such as myocardial infarction or fracture. Safety is carefully monitored. Phase II data are used to inform phase III studies.

7.2.3 Phase III/IV Studies

Phase III studies are *randomized trials* designed to assess the *effectiveness and safety* of an intervention. Outcomes of phase III studies are typically *clinical events*, such as death or tumor-free survival. Safety assessments occur over a longer period compared with phase II studies.

Phase IIIb studies are randomized trials that occur after a drug has been submitted for approval, but before approval has been granted. Phase IV studies occur after approval. Both phase IIIb and phase IV studies typically focus on *long-term safety surveillance*; phase IV studies evaluate outcomes associated with a drug or intervention as it is used in clinical practice.

7.3 Conduct of Randomized Trials

7.3.1 Comparison Group

In randomized trials, a treatment or procedure can be compared to no therapy, a similar therapy, a placebo, or a preexisting standard of care. The choice of comparison group will depend on the specific question of interest and the requirement for equipoise.

Sometimes problems with the comparison group can lead to erroneous findings in a randomized trial:

Example 7.1. In a randomized trial, investigators compared two different dialysis treatments among patients who developed acute kidney injury in the intensive care unit.[29] One group of patients received daily dialysis, while another group received dialysis every other day, the existing standard of care. The trial results suggested that patients who received daily dialysis had improved survival compared to those who received dialysis every other day.

In this particular trial, further scrutiny of the "standard of care" group revealed somewhat less vigorous dialysis therapy than expected. Specifically, the delivered intensity of dialysis in the every other day control group was considerably less than that prescribed. Although the daily dialysis group had improved survival under these study conditions, the benefit was relative to a control group that might have received suboptimal treatment. A subsequent trial that achieved comprehensive treatment targets in conventional, every other day dialysis patients found no survival difference when compared to daily dialysis.[30]

7.3.2 Placebo

A placebo is a substance or procedure that is perceived as therapy, but has no biologic or therapeutic activity. A placebo can be a pill that appears identical to a study medication, but lacks the active ingredient, or can be a "sham procedure," such as creating a skin incision under sedation to mimic a surgical procedure. Placebos are used in randomized trials to attempt to separate the *effects* of a particular treatment from the *context* of the treatment. In other words, it is possible that the setting of a trial and the administration of a specific treatment by itself produce benefit, particularly when the study outcome is subjective, such as change in pain or mood. For example, one study assigned subjects to one of two different pain relievers. Participants were told that the first compound was an inexpensive (10 cents per pill) pain reliever and that the other compound was an expensive ($2.50 per pill) pain reliever. In reality, both pills were placebos, that is, neither had biological activity. Participants in both groups reported considerable pain relief; the extent of pain relief was significantly greater in the expensive placebo group.[31]

Another important reason to use a placebo is to *blind* study participants (and, when possible, study investigators) to the treatment assignment. For example, participants in a randomized EPO trial were assigned to biweekly subcutaneous injections of EPO or biweekly subcutaneous injections of saline, which was prepared in syringes that appeared identical to EPO and could not be distinguished by the study participants, coordinators, or investigators. Blinding study participants to the treatment assignment attempts to make the intervention and control groups as similar as possible, including subjects' expectations of therapy. Blinding study investigators attempts to remove potential biases that may occur in study measurements and analysis.

7.3.3 Block Randomization

Randomizing a large number of people to different study treatments should result in balanced characteristics between the treatment groups. However, if the number of randomized subjects is relatively small, bad luck may result in unbalanced characteristics between the treatment groups. For example, one randomized trial compared a high-tech form of dialysis called "CVVH" to traditional dialysis in patients with acute kidney injury.[32] The goal of the study was to evaluate whether the newer dialysis treatment could prolong survival, compared to the traditional treatment. Baseline characteristics of the randomized subjects are shown in Table 7.1 below.

Table 7.1 Baseline characteristics from a randomized trial of CVVH versus dialysis

	CVVH (n = 84)	Traditional dialysis (n = 82)
Age (years)	54.5	56.3
Male (%)	83.3	68.3
Oliguric (minimal urine output) (%)	20.2	24.4
Liver disease (%)	42.9	29.3

In this relatively small trial, the distributions of baseline characteristics are only coarsely balanced between treatment groups. One characteristic of particular concern is liver disease, which is considerably more prevalent in the CVVH group (most likely due to chance). The relatively high proportion of liver disease, which itself is strongly related to mortality, may result in a higher mortality rate in the CVVH treatment group, even if CVVH treatment is truly beneficial.

One approach to this type of problem is called *block randomization*. To use this method, investigators first identify a particular characteristic, or group of characteristics, that may influence the study outcome. In this case, the investigators may decide from the outset that the frequency of liver disease should be balanced between treatment groups to avoid confounding. Before the study begins, investigators would decide on a total number of liver disease patients to enroll. Randomization would then be carried out separately for enrolled patients with and without liver disease so that exactly half receive each treatment. For example, if the investigators decide to enroll a total of 30 liver disease patients into the study, they could print 30 pieces of paper, half with a "0" and half with a "1," and then assign a piece of paper to each liver disease patient who enters the study. This method will distribute exactly 15 liver disease patients to each treatment group, thereby ensuring a balanced distribution of this important characteristic. A much larger randomized trial will achieve a balanced distribution of characteristics naturally.

7.3.4 Biological Versus Clinical Endpoints

Randomized trials tend to be expensive, and the costs increase with longer follow-up. To conduct randomized trials more expediently, and with lower cost, investigators often study *surrogate endpoints* such as the result of a blood test or procedure. For example, investigators wish to evaluate whether a new osteoporosis drug can reduce the risk of hip fracture. The incidence of hip fracture is very low, necessitating recruitment of thousands of subjects with many years of follow-up in order to observe enough hip fractures to conduct a meaningful comparison. An alternative approach would be to change the study outcome from hip fracture to bone mineral density after 2 years. Previous studies have established that bone mineral density predicts subsequent fracture and can be measured relatively easily and inexpensively. In other words, investigators could use bone mineral density as a *surrogate marker* of fracture. Other examples of surrogate markers include the use of cardiac ejection fraction, measured by echocardiogram, as a surrogate marker of heart failure, and the use of carbon dioxide levels, measured by arterial blood gas, as a surrogate marker of the severity of chronic obstructive pulmonary disease.

Surrogate markers are not always good indicators of clinical events. A classic example is the Cardiac Arrhythmia Suppression Trial (CAST) trial, in which patients were randomized to receive either a novel antiarrhythmic medication or placebo.[33] The CAST trial was motivated by preliminary data indicating that the new antiarrhythmic drug effectively suppressed ectopic cardiac beats, an early sign of cardiac arrhythmia. The results of the CAST trial demonstrated that the new antiarrhythmic medication substantially decreased the risk of the surrogate endpoint, ectopic beats, as predicted, but actually *increased* the risk of cardiac death, leading to premature termination of the trial. This example and others like it has led to increased demand for trials that include clinical, rather than biologic endpoints.

7.4 Limitations of Randomized Controlled Trials

7.4.1 Generalizability of the Study Population

To achieve credible results, randomized trials frequently enroll relatively healthy participants who are likely to remain in the study and comply with study procedures. These measures can increase the *internal validity* or accuracy of the results. However, the price paid for greater internal validity can be reduced *generalizability* (*external validity*), the degree to which study results are valid in other people who have the same disease.

Example 7.2. Investigators conduct a randomized trial to evaluate whether EPO treatment can reduce mortality in patients with chronic kidney disease. They recruit 4000 people who have chronic kidney disease and anemia and randomize them to

receive EPO versus placebo. To increase the likelihood of adequate follow-up, investigators exclude subjects who plan to move from the area during the upcoming year, those who have missed more than two recent clinic appointments, and those who have been enrolled in their health plan for less than 6 months.

These exclusions are applied to minimize the likelihood of subject dropout, thereby increasing the internal validity of the results. It remains possible that excluded subjects will differ from those who entered the study; however, unless there is strong reason to suspect that EPO acts differently in people who move or miss clinic appointments, results from this trial should reasonably generalize to people with anemia and chronic kidney disease.

Example 7.3. To further focus on the effects of EPO and to lessen the impact of other factors that might influence mortality in this same trial, investigators also exclude people who have high blood pressure, diabetes, coronary heart disease, valvular heart disease, heart failure, or stroke.

These exclusions may increase internal validity by removing other factors that can influence mortality. However, generalizability of results from this trial will be compromised. EPO treatment may exacerbate high blood pressure and increase the risk of heart disease in susceptible individuals; therefore, EPO might have different effects in this highly restricted study population, compared to a general population of patients with kidney disease and anemia.

7.4.2 Generalizability of the Study Environment

Randomized trials are conducted under controlled conditions to ensure that the study hypothesis is tested in a reproducible fashion. Protocols of randomized trials often cannot be duplicated in clinical practice. The specialized study environment of a randomized trial may result in internally valid findings, but results that cannot be generalized to patients who have the same condition.[34]

Example 7.4. In a community-based case-control study, investigators identified patients hospitalized for an elevated serum level of potassium (hyperkalemia, which can be life-threatening) and control individuals without hyperkalemia who resided in the same community.[35] After adjustment, the use of a potassium-sparing diuretic medication was associated with 20-fold greater odds of hospitalization for hyperkalemia.

Example 7.5. In a clinical trial, investigators randomly assigned patients with heart failure to receive a potassium-sparing diuretic or a placebo in addition to their existing heart failure medications.[36] The goal of the study was to evaluate whether the potassium-sparing diuretic could reduce mortality from heart failure. An important safety measure was hyperkalemia, which was carefully monitored. The investigators found that the potassium-sparing diuretic significantly decreased mortality compared to placebo; there was no difference in the risk of hospitalization for hyperkalemia.

How can these vastly discrepant relationships between potassium-sparing diuretic use and hyperkalemia by reconciled? It is possible that methodological differences between these studies played some role; case and control subjects in the observational study may have differed by characteristics that were not measured. The study populations were different; it is possible that potassium-sparing diuretics cause less risk of hyperkalemia in heart failure patients, who frequently take other diuretic medications that lower potassium. However, it is unlikely that these differences explain the 20-fold discrepancy in results between the studies.

The most likely source of the discrepancy is the difference in study environments. The randomized trial frequently monitored serum potassium levels in all study participants to maximize safety. Even a small increase in the serum potassium level triggered a reduction in dose or temporary discontinuation of the potassium-sparing diuretic. These intensive monitoring procedures essentially eliminated the risk of hospitalization for hyperkalemia in the randomized trial. In contrast, monitoring procedures are far less stringent in normal practice. The likely conclusion is that both study results are correct. Potassium-sparing diuretics do cause hyperkalemia in general practice, but not in the environment of a randomized clinical trial.

7.4.3 Limited Question

Randomized trials are designed to definitively answer a specific, narrowly focused research hypothesis by isolating the effect of one or a small number of therapies. Randomized trials are not designed to evaluate the mechanisms by which a therapy may produce benefit or harm. In the randomized trial of potassium-sparing diuretics in heart failure, this medication substantially improved survival with minimal side effects. Results of this trial provide a strong rationale for prescribing potassium-sparing diuretics to patients with heart failure; other studies are needed to understand *why* potassium-sparing diuretics might produce this survival benefit. Although multiarm trials can evaluate a small number of treatment groups, randomized trials are generally not the ideal environment to explore the effects of a large number of different therapies, dosages, or combinations of therapies, as this would notably dilute the size of each treatment group.

7.4.4 Limited Clinical Applicability

Randomized trials are limited to situations in which the exposure of interest can be modified in a trial setting, for example, medication use or lifestyle change. There are countless exposures that are highly relevant for study, but cannot be easily modified, such as genes, serum markers, socioeconomic conditions, and childhood history. Randomized trials are limited to specific clinical situations in which the exposure of interest can be readily modified.

7.4.5 *Randomized Design Accounts only for Confounding*

Large randomized trials will balance participant characteristics between treatment groups, thereby reducing or eliminating confounding as a potential study flaw. However, randomized trials are still prone to other problems inherent in epidemiology/clinical research studies, such as misclassification (Chap. 8) and sampling variation (Chaps. 14–16).

7.5 Analysis of Randomized Controlled Trial Data

7.5.1 *Measures of Effect*

Analysis of randomized trial data is analogous to the analysis of cohort study data. We first calculate the incidences (proportion or density) of the outcome for each treatment group, and then compare these incidences using the relative risk or the risk difference.

Example 7.6. Blood clots are a serious complication of orthopedic surgery. Postoperatively, orthopedic surgery patients are frequently prescribed enoxaparin, a subcutaneous medication that inhibits clotting. A randomized trial compared enoxaparin, the standard of care, to a new oral anticlotting medication called rivaroxaban in 1,702 patients following knee arthroplasty.[37] The primary endpoint was "major thrombotic complication," which was defined by the occurrence of deep venous thrombosis, pulmonary embolism, or death. Results of this trial appear below.

	Number of people	Number of events (%)
Rivaroxaban	824	79 (9.6%)
Enoxaparin	878	166 (18.9%)

In this example, follow-up time was relatively short (measured in days) and was not appreciably different between the treatment groups. Therefore, incidence proportion is a reasonable measure of risk in each group. *Incidence rate* would be preferred for studies with longer follow-up time or unbalanced follow-up times between the treatment groups. The individual incidence proportions can be used to calculate relative risk and risk difference.

$$\text{Relative risk of major thrombotic complication} = \frac{9.6\%}{18.9\%} = 0.5$$

$$\text{Risk difference of major thrombotic complication} = 9.6\% - 18.9\% = -9.3\%$$

If the rivaroxaban versus enoxaparin trial were a cohort study, we would interpret the above relative risk as, "rivaroxaban *is associated* with a 50% lower relative risk of a major thrombotic complication, compared to enoxaparin." Since this is a large

randomized trial, we are reasonably confident that the rivaroxaban and enoxaparin groups differ only by the use of these medications. We can therefore be bolder and interpret the relative risk as "rivaroxaban *causes* a 50% lower relative risk of a major thrombotic complication, compared to enoxaparin." We replace the term "association" with "cause" because we are confident that other characteristics of rivaroxaban and enoxaparin-treated patients are not distorting this result.

7.5.2 Numbers Needed to Treat/Harm

Emboldened by the idea that the exposure *causes* the outcome in a randomized trial, we can estimate how many patients would need to be treated with rivaroxaban instead of enoxaparin to prevent a single major thrombotic complication. The interpretation of the risk difference finding above is, "treatment of 100 knee arthroplasty patients with rivaroxaban, compared to enoxaparin, will result in 9.3 fewer major thrombotic complications." We can rearrange the risk difference data to estimate the number of knee arthroplasty patients that would need to be treated with rivaroxaban to prevent a single major thrombotic complication. Mathematically,

$$\text{Number needed to treat} = \frac{1}{\text{risk difference}} = \frac{1}{0.093} = 10.8 \text{ people}$$

A similar calculation can be performed for *adverse events* that occur in a randomized trial. For example, 50.7% of rivaroxaban-treated patients required a blood transfusion during the study, compared with 46.4% of enoxaparin-treated patients; risk difference = 50.7% − 46.4% = 4.3%. The *number needed to harm* in this case would be 1/0.043 = 23.3 people. Treatment of approximately 23 people with rivaroxaban, instead of enoxaparin, will result in one extra blood transfusion.

7.5.3 Measures of Effect in Journal Articles

Because the randomized trial design balances characteristics between treatment groups, minimizing the chance that another factor is confounding the results, the "crude," or "unadjusted" relative risk represents a valid measure of effect. However, randomized trials sometimes include adjustment for statistical reasons. Moreover, randomized trials often utilize a statistical method called *proportional hazards*, which will be discussed in Chap. 20. Proportional hazards models yield a measure of effect called the "hazard ratio," which is very similar to the relative risk. So, the terms "*relative risk*," "*adjusted relative risk*," "*hazard ratio*," and "*adjusted hazard ratio*" may be used to describe the results of randomized trials in journal articles. These measures of effect are similar.

7.5.4 Intention-to Treat-Analysis

Although study participants are assigned to a particular treatment at the beginning of a randomized trial, they may discontinue therapy or change therapy during the study. For example, some subjects initially assigned to rivaroxaban may discontinue treatment due to side effects, or may switch to enoxaparin. How should the data be analyzed in the presence of changing or switching therapies *after the study begins*? There are three general approaches.

1. Remove subjects from the study when they discontinue or switch therapy (censoring).
2. Analyze subjects according to the treatment they are receiving at the time they develop the study outcome (as-treated analysis).
3. Analyze subjects according to their initial treatment assignment and ignore switching that occurs during the trial (intention-to treat-analysis).

Choice 3, the *intention-to-treat analysis*, is generally the preferred method for analyzing randomized trial data.

To understand this concept, consider a hypothetical situation in which 200 subjects who are assigned to rivaroxaban develop postoperative bleeding, prompting a switch to enoxaparin (possibly due to the study investigators' preconceived notion that enoxaparin is safer). Analysis choice 1 would remove these 200 subjects from the analysis, analysis choice 2 would consider them to be in the enoxaparin group after they switch therapy, and analysis choice 3 would consider them to remain in the rivaroxaban group, per their original treatment assignment.

Choices 1 and 2 undo the randomization process. Randomized trials are designed to balance participant characteristics *at the beginning of a study*. Removing some subjects from one treatment group, or preferentially switching some subjects from one treatment group to another after a study begins may disrupt this initial balance. The disruption will create uncertainty as to whether study findings are due to the effect of the treatment, or result from differences in the characteristics of the treatment groups.

How would analysis choices 1 and 2 impact the study results in this example? Subjects who develop postoperative bleeding may be inherently sicker than the remainder of the study population (postoperative bleeding may indicate comorbidity or may itself increase the risk of death). Preferentially removing sicker individuals from the rivaroxaban group may result in a falsely low incidence of the study outcome (which includes death) in the rivaroxaban group, and could lead to an inflated relative risk, comparing rivaroxaban to enoxaparin.

Switching subjects who develop postoperative bleeding from the rivaroxaban group to the enoxaparin group in the analysis will result in both an unusually low incidence of the study outcome in the rivaroxaban group *and* an unusually high incidence of the study outcome in the enoxaparin group. This will further inflate the relative difference between the two treatments.

As a general rule, analysis choices 1 and 2 will have unpredictable consequences on the study outcome in the presence of switching or stopping therapy during a trial. The relative risk may be falsely increased, falsely decreased, or unchanged; considerable scrutiny is often required to guess at which one of these biases might have occurred.

Choice 3, the intention-to-treat approach, maintains the initial randomization process. In the presence of switching therapy or stopping therapy during a trial, the intention-to-treat analysis will lead to a *predictable change* in the relative risk: it will move closer to 1.0. This predictable change is called *bias toward the null*.

Consider the intention-to-treat principle applied to an extreme example in which all 824 subjects assigned to rivaroxaban immediately switch to enoxaparin after the study begins. Under the intention-to-treat principle, the rivaroxaban group is analyzed according to the initial treatment assignment, despite the fact that all of these subjects actually receive enoxaparin during the trial. The observed incidence of the study outcome in the "rivaroxaban" group will be equal to that of the enoxaparin group (because all subjects in this group actually receive enoxaparin). Comparing the incidence of outcomes in this "rivaroxaban" group to that in the enoxaparin group will result in observing a relative risk of 1.0, and a risk difference of 0.

The consequences of the intention-to-treat analysis will be subtler for more realistic conditions of switching or stopping therapy. If 200 rivaroxaban-assigned subjects switch to enoxaparin during the study, the resulting rivaroxaban group in the intention-to-treat analysis will become tainted with enoxaparin use, producing an incidence of the study outcome that moves a little closer to that of enoxaparin.

The intention-to-treat analysis allows treatment groups to mix together, producing groups that more closely resemble one another. Comparing outcome rates between these mixed groups will yield relative risks that are close to 1.0 and risk differences that are closer to 0, as compared to study results that would be obtained if no switching occurred. Greater treatment discontinuation or greater switching during a trial causes more bias toward the null under an intention-to-treat analysis. This predictable "conservative" error using intention-to-treat may result in missing some true treatment effects, particularly if these effects are subtle, but intention-to-treat will generally not produce false associations, or overexaggerate the effects of a treatment.

Although there may be certain instances where non-intention-to-treat analyses are used, special care is required for describing the results and ensuring that bias has not occurred in these settings.

7.5.5 Subgroup Analyses

Investigators may be interested in using randomized trial data to explore whether a specific treatment might be particularly helpful or harmful to specific groups of subjects. For example, the risk difference of a major clotting outcome, comparing rivaroxaban to enoxaparin, was −9.3% among the entire knee arthroplasty study population. It is possible that rivaroxaban is particularly effective in people who have a family history of clotting disorder or those who have complicated surgeries. It is also possible that rivaroxaban is particularly harmful in certain subgroups. To explore these hypotheses, investigators would calculate the risk difference, comparing rivaroxaban to enoxaparin, specifically within the individual subgroups of interest.

However, examination of a treatment effect across a large number of subgroups can lead to spurious findings, especially if statistical significance is used for guidance. Dividing a study population into many subgroups often leaves few outcomes in any particular group, increasing the likelihood of a chance finding. The probability of finding a "false-positive" result within any individual subgroup will increase as more subgroups are tested.

To demonstrate the amount of natural variation in treatment effect that is expected across subgroups, consider the relative risk of major thrombotic complication, comparing rivaroxaban to enoxaparin, by subgroups defined by astrological sign of the study patient. The assumptions for Table 7.2 are a relative risk of 0.5 for the entire cohort, a sample size of 1700 subjects, and no true difference in the effect of rivaroxaban by astrological sign in the population.

Table 7.2 Natural variation across subgroups in the rivaroxaban versus enoxaparin trial

Subgroup	Relative risk of a major thrombotic complication
All subjects (n = 1700)	0.50
Capricorn, "The Goat"	0.62
Aquarius, "The Water Carrier"	**1.23**
Pisces, "The Fish"	0.40
Aries, "The Ram"	0.34
Taurus, "The Bull"	0.44
Gemini, "The Twins"	0.23
Cancer, "The Crab"	0.47
Leo, "The Lion"	**0.99**
Virgo, "The Virgin"	0.37
Libra, "The Scales"	0.53
Scorpio, "The Scorpion"	0.57
Sagittarius, "The Archer"	0.36

Bold numbers highlight results that are substantially different from the others.

Most subgroups demonstrate the protective affect of rivaroxaban that was observed in the full study population; however, rivaroxaban appears to be ineffective among subjects with an astrological sign of Leo and harmful among subjects with an astrological sign of Aquarius.

For general guidance, findings from subgroup analyses are more likely to be valid if:

1. There is biological plausibility for a particularly strong effect or harm in the subgroup.
2. The subgroup analysis was pre-specified at the beginning of the study.
3. There are a reasonably large number of outcomes in the subgroup.

For example, the kidneys clear enoxaparin but not rivaroxaban. A prespecified subgroup analysis might compare major bleeding risks of rivaroxaban versus enoxaparin in subgroups of participants with normal and impaired kidney function.

Participants with normal kidney function ($N = 1362$)	Incidence of major bleeding (%)
Rivaroxaban	1
Enoxaparin	1
Risk difference	0

Participants with impaired kidney function ($N = 340$)	Incidence of major bleeding (%)
Rivaroxaban	2
Enoxaparin	6
Risk difference	–4

Results of this subgroup analysis demonstrate that enoxaparin is more likely than rivaroxaban to cause major bleeding in subjects with impaired kidney function, but that the two drugs have equivalent bleeding risks in subjects with normal kidney function. Retention of enoxaparin in the setting of impaired kidney function provides a biological rationale for this subgroup finding.

Consider a second example of a subgroup analysis. Investigators in an HIV vaccine trial randomize high-risk individuals to an experimental HIV vaccine versus placebo. At the end of the trial, the incidences of HIV infection were identical comparing HIV vaccine-treated subjects to those given a placebo. Further analyses further revealed that the HIV vaccine had no effect on preventing HIV infection among subgroups of subjects who were young, those who did not have a recent major infection, and those who had a normal white blood cell count. However, a seemingly impressive vaccine effect was noted for African-American women in the study. Of the four African-American women who received the HIV vaccine, only one developed HIV, whereas three of six African-American women who received placebo developed HIV; risk difference = 50% − 25% = 25%, $p = 0.02$ for comparison. Should this statistically significant subgroup finding motivate further studies of the HIV vaccine among African-American women?

In this case, there is no biological rationale for the HIV vaccine to work only in African-American women, but not in men or people of other races. Moreover, the subgroup of African-American women in this study was small (only ten participants and only three HIV outcomes) and the subgroup analysis was not prespecified at the beginning of the study. This subgroup finding is likely to be spurious, resulting from a "fishing expedition" by the vaccine manufacturer to find a vaccine effect somewhere. Additional study of the HIV vaccine in African-American women based on this subgroup finding alone is *not* warranted.

Chapter 8
Misclassification

Learning Objectives

1. Misclassification arises from errors in measuring study data.
2. Nondifferential misclassification results from random errors that occur roughly equally within a study population.
3. Nondifferential misclassification usually yields measures of association that are closer to the null value of 1.0.
4. Differential misclassification results from systematic errors that occur preferentially within a subset of a study population.
5. Differential misclassification may obscure true associations, falsely amplify associations, or create false associations
6. Uniform data collection procedures help to minimize the possibility for differential misclassification.

While it may initially seem like clinical research studies could err in an infinite number of ways, these errors generally fall into one of a few major categories, each with distinct implications. An important goal is to recognize patterns of error in clinical/epidemiological studies and to understand the consequences that result from them.

8.1 Definition of Misclassification

Misclassification results from errors in measuring the study data. Any type of study data can potentially be misclassified. Recall the example from Chap. 5 in which researchers investigated whether a new antibiotic, supramycin, could cause a rash. The exposure in this example was supramycin use, which could be measured using a variety of possible methods, including telephone interview, query of automated pharmacy records, or transcription of prescription medication bottle labels. Each of these methods will have varying degrees of accuracy for capturing true supramycin use; however, none will be 100% perfect for determining whether a person actually takes supramycin. Directly observing each study subject swallowing the medication may be the only measurement technique that is truly free of error.

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_8, © Springer Science+Business Media, LLC 2009

The study outcome, rash, could be measured by reviewing medical charts, thereby relying on the treating physicians' expertise at diagnosing rash. A more valid method for measuring the occurrence of rash might be periodic skin examinations by a study dermatologist. Even this method may not be perfect, as some rashes can remit between study examinations.

Measurement considerations are not limited to the study exposure and outcome. Additional data elements of interest in the supramycin study might include a recent history of upper respiratory infection, the white blood cell count, and current aspirin use. All of these data elements could potentially be misclassified depending on the methods that are used to ascertain them.

8.2 Nondifferential Misclassification

8.2.1 Example of Nondifferential Misclassification of the Exposure

We begin with an idealized cohort study that evaluates the association of supramycin use with rash under the assumption that supramycin use can be measured perfectly (Table 8.1).

We will consider this relative risk of 2.44 to represent the "gold-standard" association of supramycin use with rash under the assumption of perfect ascertainment of supramycin use and perfect ascertainment of rash.

Next, consider a practical situation in which supramycin use is measured by transcribing the labels of prescription medication bottles. For the purpose of this example, we will consider the transcription method to have some degree of error in capturing true supramycin use; specifically, we will assume that 10% of participants classified as "supramycin users" by the transcription method actually do not take

Table 8.1 Association of supramycin with drug rash: Supramycin use measured perfectly

		Rash		
		YES	NO	Total
SUPRAMYCIN	YES	22	978	1000
	NO	18	1982	2000

Relative risk = (22 / 1000) / (18 / 2000) = 2.44

supramycin during the study, possibly due to noncompliance or medication side effects. What is the impact of this measurement error on the study results?

To reflect the fact that "supramycin users" are now tainted with some non-supramycin users, we will move 10% of nonsupramycin users into the supramycin group. We obtain nonsupramycin users in equal proportions from subjects who do and do not subsequently develop a rash, as shown in Table 8.2.

Study investigators will observe and report only the second part of Table 8.2, in which some "supramycin users" classified by the transcription method are actually nonusers. This 10% misclassification of supramycin use results in observing a new relative risk of 2.25. Stated another way, supramycin use is associated with 2.44-fold greater risk of rash in the idealized world, where medication use is ascertained perfectly, and a 2.25-fold greater risk of rash in the presence of 10% misclassification of the exposure.

Table 8.2 Association of supramycin with drug rash: 10% misclassification of supramycin

Investigators do not observe these idealized data

		Rash		
		YES	NO	Total
		22	978	1000
SUPRAMYCIN	YES	2 → 18	198 → 1982	200 → 2000
	NO			

Investigators observe these misclassified data

		Rash		
		YES	NO	Total
SUPRAMYCIN	YES	24	1176	1200
	NO	16	1784	1800

Relative risk = (24 / 1200) / (16 / 1800) = 2.25

In this example, we assumed that 10% misclassification of supramycin use occurred in equal proportions among subjects who develop or do not develop a future rash. This assumption is probably reasonable because errors in classifying supramycin use should not depend on a condition (rash) that has not yet developed at the time of measurement.

Now let's consider a second situation in which supramycin use is determined by telephone interview, which may be less accurate than direct transcription of prescription medication labels. For the purpose of this example, we will now consider that 30% of subjects who report supramycin use by telephone interview actually do not take supramycin during the study. What is the impact of even greater misclassification of supramycin use on the study results? To answer this question, we will again move nonsupramycin users into the supramycin group to reflect the fact that "supramycin users," classified by telephone interview, actually include some nonsupramycin users. We will again assume that errors in measuring supramycin use are unlikely to be related to the future development of a rash, as depicted in Table 8.3.

Again, the relative risk is lowered, although this time to a larger degree.

What if misclassification of the nonsupramycin users were to also occur? We will now assume that the telephone interview method also misclassifies 30% of the nonsupramycin users, such that 30% of subjects who report "no supramycin use" by telephone interview actually use supramycin during the study (subjects may have difficulty remembering all of their medications during a telephone interview). What are the impact of 30% misclassification of the supramycin users *and* 30% misclassification of the nonsupramycin users (Table 8.4)?

In these examples, progressive misclassification of the exposure, supramycin use, yields relative risk estimates that are progressively lower than those observed when the exposure was measured perfectly. As more misclassification occurs, the relative risk is altered to a greater degree.

Actually, what occurs is movement of the relative risk toward 1.0. The predictable *attenuation* of relative risk estimates toward 1.0 is called *bias toward the null*. Random misclassification of the exposure leads to bias toward the null because the process mixes exposed and unexposed individuals, obscuring true differences that may exist between them. In other words, as the "supramycin group" becomes tainted with nonsupramycin users, the groups will become increasingly similar to one another, diminishing contrasts in their incidence of rash.

8.2.2 Definition and Impact of Nondifferential Misclassification of the Exposure

Misclassification that occurs randomly or roughly equally throughout a study population is called *nondifferential or nonselective misclassification*. Examples of nondifferential misclassification include use of a standard cuff to measure blood pressure, administration of a questionnaire to determine the presence of diabetes, and use of a variable laboratory assay to measure serum triglyceride levels.

Table 8.3 Association of supramycin with drug rash: 10% misclassification of supramycin

Investigators do not observe these idealized data

		Rash		
		YES	NO	Total
		22	978	1000
SUPRAMYCIN	YES	5	595	600
	NO	18	1982	2000

Investigators observe these misclassified data

		Rash		
		YES	NO	Total
SUPRAMYCIN	YES	27	1573	1600
	NO	13	1387	1400

Relative risk = (27 / 1600) / (13 / 1400) = 1.82

Consider the use of one of these measurement tools, the blood pressure cuff, in a study designed to evaluate the association of systolic blood pressure with stroke. The cuff method is known to produce variable blood pressure readings within an individual; misclassification of blood pressure is likely to occur. There is no reason to expect that the *degree or direction of inaccuracy* of the blood pressure cuff will differ among subjects who do or do not develop a future stroke; therefore, misclassification of blood pressure by the cuff method is likely to be nondifferential. The consequence will be an observed association of blood pressure with stroke that is closer to 1.0 compared to that which would be observed if blood pressure were measured perfectly.

How could the amount of misclassification of blood pressure be reduced in such as study? Investigators could consider more invasive methods to measure blood

Table 8.4 Association of supramycin with rash: bi-directional exposure misclassification

Investigators do not observe these idealized data

		Rash		
		YES	NO	Total
		22	978	1000
SUPRAMYCIN	YES	5	595	600
	NO	18	1982	2000

			7	293	300

Investigators observe these misclassified data

		Rash		
		YES	NO	Total
SUPRAMYCIN	YES	20	1280	1300
	NO	20	1680	1700

Relative risk = (20 / 1300) / (20 / 1700) = 1.31

pressure, such as the use of an arterial catheter. This method measures blood pressure more accurately than a standard cuff, but requires arterial puncture and expensive monitoring equipment. A relatively simple method to reduce random error from the blood pressure cuff is to perform *repeated measurements*. For example, many health studies will average the results from three consecutive blood pressure cuff readings obtained 5 min apart. The use of repeated measurements can be a highly effective method for reducing random error in study measurements.

On the basis of above discussion of exposure misclassification, we can formally state that:

1. Nondifferential misclassification of the exposure occurs when error in measuring the exposure is similar among subjects who experience or do not experience the outcome.

2. Nondifferential misclassification of the exposure results in bias toward the null. The greater the misclassification, the closer the observed relative risk moves toward 1.0.

8.2.3 Nondifferential Misclassification of the Outcome

We now turn to misclassification of the study outcome, rash. First, we will examine the impact of *overdiagnosing* rash, such that some rashes classified by the study procedures are in fact false. To reflect false diagnoses of rash, we will move some subjects without a rash into the "rash" group as shown in Table 8.5.

An incorrect diagnosis of rash in 1% of the study population will result in a total of 30 false "rash" diagnoses. We will assume that the overdiagnosis of rash is *not* related to whether a study subject uses supramycin. Under this assumption, false rashes will occur in equal proportions among supramycin users and nonsupramycin users.

Equal misdiagnosis of rash has again resulted in bias toward the null; the observed relative risk is closer to 1.0 compared to the true relative risk of 2.44 that was obtained when drug rash was diagnosed perfectly in Table 8.1. The reason that such a small degree of outcome misclassification produced so much conservative bias in this particular example is that the outcome, drug rash, is relatively rare. Note again that study investigators observe, analyze, and present only the data from the second part of the table that is misclassified. Investigators may not be aware that some "rashes" classified by their study methods actually represent false diagnoses.

The impact of nondifferential misclassification of the study outcome differs when there is *underdiagnosis* of the disease. For example, study investigators may decide to query medical charts to ascertain the occurrence of rash. The chart review method might miss some rashes, either because a skin exam was not performed, results of the exam were not documented in the medical chart, or the rash was not present at the time of exam. For the purposes of this example, we will consider the medical chart review to miss half of the rashes, such that 50% of subjects who are classified as having "no rash" by medical record review actually *had* a rash during the study. The impact of underdiagnosing rash is depicted in Table 8.6.

The observed relative risk of 2.44 in the presence of underdiagnosis of rash is equal to the true relative risk. Nondifferential misclassification of the outcome that results from underdiagnosis of the disease or condition *has no impact on the relative risk.*

8.2.4 Definition and Impact of Nondifferential Misclassification of the Outcome

On the basis of above discussion of outcome misclassification, we can formally state that:

Table 8.5 Association of supramycin use and rash: 1% overdiagnosis of rash

Investigators do not observe these idealized data

		Rash		
		YES	NO	Total
SUPRAMYCIN	YES	22 ← 10	978	1000
	NO	18 ← 20	1982	2000
	Total	40 ← 30	2960	3000

Investigators observe these misclassified data

		Rash		
		YES	NO	Total
SUPRAMYCIN	YES	32	968	1000
	NO	38	1962	2000

Relative risk = (32 / 1000) / (38 / 2000) = 1.68

1. Nondifferential misclassification of the outcome occurs when error in measuring the outcome is similar among exposed and unexposed subjects.
2. Nondifferential misclassification of the outcome may result in bias toward the null or no change in the observed relative risk.

In summary, nondifferential misclassification of either the exposure or the outcome typically results in observing a relative risk that is attenuated, or closer to 1.0, than

Table 8.6 Association of supramycin use and rash: 20% underdiagnosis of rash

Investigators do not observe these idealized data

		Rash		
		YES	NO	Total
SUPRAMYCIN	YES	22 → 11	978	1000
	NO	18 → 9	1982	2000
	Total	40 → 20	2960	3000

Investigators observe these misclassified data

		Rash		
		YES	NO	Total
SUPRAMYCIN	YES	11	989	1000
	NO	9	1991	2000

Relative risk = (11 / 1000) / (9 / 2000) = 2.44

the true relative risk that would be obtained if the study data were measured perfectly. The usual consequence of nondifferential misclassification is to obscure detection of true associations, especially if the size of the association of interest is modest. In contrast, nondifferential misclassification alone will not create false associations, nor exaggerate the strength of an association.

8.3 Differential Misclassification

Thus far, we have considered measurement error that occurs at random. We now turn to examples of measurement error that occur systematically. Returning to the use of medical chart review to define rash, we will now presume that the examining physician *is* aware of whether a study subject is using supramycin at the time of their examination. Preconceived concerns about the side effects of supramycin may motivate the examining physician to perform a particularly detailed skin exam, or to overdiagnose a nonsignificant skin lesion as a rash. Consider the impact of falsely classifying ten nonrashes as rash *only among the supramycin users* (Table 8.7).

In the above example, preferential misclassification of the outcome, rash, among only exposed (supramycin-treated) subjects resulted in an observed relative risk that is *further from 1.0* compared to the true relative risk of 2.44. Misclassification that occurs preferentially or systemically within a particular subset of the study population is called *differential* misclassification. Differential misclassification can alter relative risk estimates toward or away from 1.0, depending on the particular situation. Careful scrutiny of the study methods and results are required to estimate the direction of bias that may result from differential misclassification.

Based on the above discussion, we can formally state that:

1. Differential misclassification occurs when:
   a. there is misclassification of the exposure that differs among subjects who experience or do not experience the outcome or
   b. there is misclassification of the outcome that differs among subjects who are exposed or unexposed
2. Differential misclassification can result in observing a relative risk that is closer to or further from 1.0, depending on the particular situation.

In contrast to nondifferential misclassification, which can obscure detection of true associations, differential misclassification can falsely amplify the strength of an existing association, or even create false associations. These wide-ranging consequences motivate study investigators to strive for the use of *uniform data collection methods* in clinical/epidemiological studies.

Differential misclassification of the exposure may be of particular concern in case-control studies that query diseased and nondiseased individuals as to their previous exposure status.

Table 8.7 Differential misclassification of rash in ten supramycin users

Investigators do not observe these idealized data

		Rash		
		YES	NO	Total
SUPRAMYCIN	YES	22	978	1000
	NO	18	1982	2000

10

Investigators observe these misclassified data

		Rash		
		YES	NO	Total
SUPRAMYCIN	YES	32	968	1000
	NO	18	1982	2000

Relative risk = (32 / 1000) / (18 / 2000) = 3.56

Example 8.1 To evaluate whether alcohol use during pregnancy affects the risk of fetal birth defects, investigators identify 90 case infants with a birth defect and 280 control infants without a birth defect. They query mothers of each infant as to their use of alcohol during pregnancy. The investigators find that any prenatal alcohol consumption is associated with a 12% *lower risk* of a major birth defect. Could misclassification have impacted these paradoxical findings?

To investigate possible effects of misclassification, we start by examining hypothetical data that demonstrate an association of prenatal alcohol consumption with a *greater risk* of birth defect, as expected when maternal alcohol use and birth defect are measured perfectly (Table 8.8).

We next consider real-world data, in which maternal alcohol use is ascertained by administering a questionnaire to all study mothers after they deliver. Some general underreporting of prenatal alcohol use is expected because alcohol consumption

Table 8.8 Maternal alcohol use and birth defect: Maternal alcohol use measured perfectly

Investigators do not observe these idealized data

		Birth defect	
		YES (cases)	NO (controls)
Prenatal alcohol use	YES	20	30
	NO	70	250
	Total	90	280

Odds ratio = (20*250) / (30*70) = 2.38

during pregnancy is known to be harmful to the fetus. For the purposes of this example, we will assume that 30% of study mothers who consumed alcohol during pregnancy will falsely report "no alcohol use" on the questionnaire. As a result of this misclassification, the "no alcohol use" group, defined by questionnaire, will include some mothers who actually used alcohol during pregnancy, as demonstrated in Table 8.9. Assuming that an equal proportion of case and control mothers underreport prenatal alcohol use, nondifferential misclassification of prenatal alcohol use results in an odds ratio that is (modestly) closer to 1.0 than that obtained when maternal alcohol use was measured perfectly.

Although nondifferential misclassification is an expected limitation of this study, a major concern here is that mothers of an infants who have a birth defect will be *particularly likely* to underreport alcohol use during pregnancy, due to feelings of guilt. We now consider study data after eight additional mothers of infants who have a birth defect fail to report their true prenatal alcohol consumption, as shown in Table 8.10.

The original observation that prenatal alcohol use was associated with a 12% lower risk of birth defect may be explained by differential misclassification of the exposure. In this somewhat extreme example, differential misclassification resulted in a spurious association in the opposite direction of the true association.

Differential misclassification of the exposure that occurs in case-control studies is also called *recall bias*. Recall bias occurs because diseased (case) and nondiseased (control) individuals may recall previous events differently. One method to reduce

Table 8.9 Maternal alcohol use and birth defect: Nondifferential misclassification

Investigators do not observe these idealized data

		Birth defect: YES	Birth defect: NO
Prenatal alcohol use	YES	20 → 6	30 → 9
	NO	70	250

Investigators observe these misclassified data

		Birth defect: YES	Birth defect: NO
Prenatal alcohol use	YES	14	21
	NO	76	259

Odds ratio = (14*259) / (21*76) = 2.27

recall bias in case-control studies is to identify exposure data that were *collected before study subjects became aware of their disease status*, if such data are available. For example, information regarding maternal alcohol use might have been collected during prenatal medical visits, before study mothers were aware of their infant's subsequent birth defect status. Prenatal alcohol use is still subject to nondifferential misclassification; however, the impact of nondifferential misclassification is predictable. As a general rule, differential misclassification should be suspected when study subjects or study personnel are aware of the exposure or the outcome status at the time the study data are measured.

Differential misclassification need not be limited to case-control studies. Consider the impact of misclassification in the following cohort study.

Table 8.10 Maternal alcohol use and birth defect: differential misclassification

Investigators do not observe these idealized data

		Birth defect	
		YES	NO
Prenatal alcohol use	YES	14 → 8	21
	NO	76	259

Investigators observe these misclassified data

		Birth defect	
		YES	NO
Prenatal alcohol use	YES	6	21
	NO	84	259

Odds ratio = (6*259) / (21*84) = 0.88

Example 8.2 Laparoscopic cholecystectomy is a common surgical procedure that results in less postoperative pain and a shorter hospital stay. Investigators compared the risks of major infections between patients who underwent laparoscopic cholecystectomy versus those who underwent open cholecystectomy. An inpatient infection control team diagnosed in-hospital infections based on daily wound examinations. Outpatient postoperative infections within 30 days of the surgery were assessed by telephone interview. Postoperative infection rates were found to be considerably lower in the laparoscopic group (2%) compared to the open surgery group (5%). What is the potential impact of misclassification on these study findings?

In this example, two different methods were used to diagnose surgical site infections: physical examination for hospital inpatients and telephone surveillance

for outpatients. One might suspect that the first method might have greater sensitivity for detecting infections, particularly those that are early or subtle in nature. Since patients undergoing laparoscopic surgery are typically discharged from the hospital earlier than those undergoing open surgery, laparoscopic surgery patients may experience lower infection rates simply due to the use of a less accurate method (telephone surveillance) for detecting infections in the outpatient setting. Differential misclassification of the outcome in this example could create a spurious association of laparoscopic surgery with a lower risk of surgical site infection. The use of more than one method to assess the study outcome represents an initial clue that differential misclassification may have occurred in the study.

8.4 Assessment of Misclassification in Clinical Research Articles

The methods used to ascertain the study exposure, outcome, and other data are typically described in the *methods* section of a clinical research article, following description of the study population. Once the study authors define the methods used to measure the study data, they generally take the liberty of applying the idealized descriptions throughout a research article. For example, authors of a study that defines appendectomy as "all persons reporting to have previously undergone an appendectomy via telephone interview" will freely use the term "appendectomy," throughout their paper, rather than the reality, "all persons claiming to have undergone appendectomy by telephone interview." When the analyses are performed, it is the potentially misclassified data that are used, the impact of which has been discussed in this chapter.

Assessment of misclassification in a clinical/epidemiological research article is *subjective* because there is often no way to go back and obtain the actual data. No study data are perfect and therefore some degree of measurement error is likely. We have to use common and clinical sense to decide on the nature and magnitude of misclassification that might have occurred in a research study, and then apply the rules outlined in this chapter to predict the consequences of misclassification. A useful approach to judging misclassification in a clinical research article is to ask the following questions:

(1) Based on the data collection methods, what study elements may have been misclassified?
(2) Is misclassification likely to be differential or nondifferential?
(3) What is the expected impact of misclassification on the study results?

Chapter 9
Introduction to Confounding

Learning Objectives

1. Confounding is an important limitation of observational studies.
2. Confounding alters the interpretation of study results, obscuring whether the exposure is a cause of the outcome.
3. A confounder is classically defined as a factor that is:
   a. associated with the exposure,
   b. associated with the outcome, and
   c. not in the causal pathway of association.
4. Study data are used to judge whether a potential confounder is associated with the exposure and the outcome.
5. Biological and clinical knowledge are used to judge whether a potential confounder is in the causal pathway of association.
6. Confounding-by-indication occurs when the specific indication for a medication confounds the association between the use of that medication and the study outcome.

9.1 Confounding and the Interpretation of Clinical Data

We begin by again considering the question of whether a new antibiotic, supramycin, causes a rash. The most direct method to address this question would be a randomized trial. For example, investigators could identify a group of adults with acute bronchitis, and then randomly assign them to receive supramycin or amoxicillin, which is an established antibiotic for bronchitis. Both antibiotics could be prepared in identical appearing capsules to blind study subjects and investigators to the treatment assignment. Baseline data from this hypothetical randomized trial are presented in Table 9.1 and results are presented in Table 9.2.

The data in Table 9.2 indicate that participants who are randomized to supramycin have a 60% greater incidence of rash, compared to those who are randomized to amoxicillin (0.71% versus 0.44%). In the context of a large randomized trial, these results can be interpreted as, "supramycin *causes* a 60%

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_9, © Springer Science+Business Media, LLC 2009

Table 9.1 Characteristics from randomized trial comparing supramycin to amoxicillin

	Supramycin (N = 5,110)	Amoxicillin (N = 4,998)
Age (years)	44.9 (13.2)	44.7 (12.1)
Female sex	2,480 (48.5)	2,512 (50.3)
Race		
Caucasian	2,432 (47.6)	400 (48.0)
African-American	1,623 (31.8)	640 (32.8)
Other	1,055 (20.6)	58 (19.2)
Current smoker	620 (12.1)	04 (12.1)
History of cardiovascular disease	75 (19.1)	68 (19.4)
History of eczema	99 (9.8)	71 (9.4)

Data presented as mean (standard deviation) or number of subjects (percent)

Table 9.2 Results from randomized trial comparing supramycin to amoxicillin

	Number of subjects	Number of rashes	Incidence proportion (%)	Relative risk
Supramycin	5,110	36	0.71	1.60
Amoxicillin	4,998	22	0.44	p = 0.001

greater risk of rash compared to amoxicillin," because the supramycin and amoxicillin groups differ only by the use of these antibiotic medications. Alternative explanations to explain the study findings such as "supramycin-treated patients are younger, and younger people tend to get rashes" or "supramycin is usually prescribed by physicians who like to prescribe new medications, and these physicians are also better at diagnosing rashes," are *not* valid explanations for the study results, because the randomized design balances such attributes between the treatment groups. Moreover, blinding study participants and study investigators to the treatment status safeguard against potential differential errors in ascertaining rash (misclassification, *see* Chap. 8).

The above trial would likely require multiple sites to recruit 10,000 subjects, cost millions of dollars, involve a lengthy recruitment period, and may be subject to ethical issues if one of the antibiotics is thought to be superior for treating bronchitis.

The question of whether supramycin causes a rash could be addressed more expediently and less expensively using observational data. Investigators could identify data from a large health maintenance network that includes supramycin and amoxicillin on the formulary. Using the computerized medical record system, the investigators could select 5,000 patients with acute bronchitis who were treated with supramycin and another 5,000 patients with acute bronchitis who were treated with amoxicillin. They would then observe these cohorts for the development of rash; these data for this study are presented in Tables 9.3 and 9.4.

These results indicate that supramycin use is *associated* with a 2.53-fold greater risk of rash, compared to amoxicillin. In contrast to the randomized trial example,

Table 9.3 Characteristics from observational study comparing supramycin to amoxicillin

	Supramycin users (N = 5,000)	Amoxicillin users (N = 5,000)
Age (years)	57.2 (12.9)	44.3 (13.1)
Female	2,755 (55.1)	2,900 (50.3)
Race		
Caucasian	2,005 (40.1)	2,560 (51.2)
African-American	1,920 (38.4)	1,340 (26.8)
Other	1,075 (21.5)	1,100 (22.0)
Smoker	590 (11.8)	630 (12.6)
History of cardiovascular disease	770 (15.4)	795 (15.9)
History of eczema	701 (14.0)	435 (8.7)

Data presented as mean (standard deviation) or number of patients (percent)

Table 9.4 Results from observational study comparing supramycin to amoxicillin

	Number of patients	Number of rashes	Incidence proportion (%)	Relative risk
Supramycin	5,000	43	0.86	2.53
Amoxicillin	5,000	17	0.34	p = 0.001

we *cannot* readily infer from these observational data that supramycin use *causes* a rash, because supramycin and amoxicillin groups differ by characteristics other than the use of these medications. For example, supramycin-treated patients are older, more likely to be African-American, and more likely to have a previous history of eczema, compared to amoxicillin-treated patients. Supramycin- and amoxicillin-treated patients may also differ by characteristics that are not presented in the table. Because of these differences, it is possible that other characteristics of supramycin users, and not supramycin itself, are responsible for the greater incidence of rash.

The idea that a third factor may be responsible for an observed association between an exposure and an outcome is called *confounding*. Confounding is essentially a problem that is intrinsic to observational studies and is well addressed by large randomized trials. Smaller randomized trials can occasionally run into problems with confounding if the treatment groups are unbalanced with respect to participant characteristics, simply due to chance.

The presence of confounding does not imply that the observed study data are false. The observational data above demonstrate that supramycin-treated patients are more likely to develop a rash than amoxicillin-treated patients. This finding is true regardless of whether or not another factor, other than supramycin, is *causing* the rash. Confounding alters *the interpretation* of study results. If another factor is in fact responsible for the observed association of supramycin use with rash, then the case for supramycin use as a *cause* of rash is weakened.

Consider two other examples of potential confounding in observational studies.

Example 9.1. Researchers reporting from the Nurses Health Study observed that women who consumed the highest quantities of trans fats were at greatest risk for developing future coronary heart disease.[38] On first glance, these data appear to suggest that trans fats may be a cause of coronary heart disease; however, women who eat high quantities of trans fats may also consume high quantities of other atherogenic fats and be less likely to exercise compared with women who eat low quantities of trans fats. It is possible that other dietary factors and/or lower exercise levels, and not trans fat intake, are the true cause of the increased risk of coronary heart disease.

Example 9.2. Investigators reporting from a cancer registry observe an association of mononucleosis infection with the development of Hodgkin's lymphoma later in life. These observational data do not clarify whether it is the mononucleosis, or other characteristics of individuals who contract mononucleosis, that cause Hodgkin's lymphoma. It is possible that people who are susceptible to mononucleosis are also susceptible to other viral infections that increase lymphoma risk, or have underlying differences in B-cell function that may be linked with lymphoma. A randomized trial is not possible in this situation because people cannot ethically be randomized to mononucleosis. Additional work to support the hypothesis that mononucleosis causes Hodgkin's lymphoma might include clinical studies that demonstrate a graded association of higher Epstein-Barr viral titers with increased Hodgkin's lymphoma risk, and laboratory studies that document Epstein-Barr virus within lymphoma cells.

In summary, the relationship between exposure and outcome in observational studies may be ambiguous due to the lingering possibility of confounding. The remainder of this chapter will focus on how to identify confounding factors in clinical research studies. Chapter 10 will describe methods used to mitigate the effect of confounding factors in an attempt to isolate the independent association between an exposure and an outcome, thereby suggesting causation.

9.2 Formal Evaluation of a Potential Confounding Factor

Returning to the results of the observational supramycin study, supramycin use was associated with a 2.53-fold greater risk of rash, compared to amoxicillin use. Examining the observational baseline characteristics table, one concern is that supramycin-treated patients were more likely to have a previous history of eczema, which might itself increase the risk of rash. In other words, previous eczema might be confounding the observed association of supramycin use with rash. How do we evaluate whether previous eczema is indeed a confounder in the study?

9.2.1 Evaluation of a Confounder: Association with Exposure

The first step is to explicitly identify the three factors in question: the exposure, the outcome, and the potential confounding factor. In this case, the exposure is supramycin use, the outcome is rash, and the potential confounder under evaluation is previous eczema. We begin by evaluating whether the confounder in question is associated with the exposure. An association between the confounder (previous eczema) and the exposure (supramycin use) means that the proportion of subjects with previous eczema differs among supramycin and amoxicillin users.

	Supramycin users (N = 5,000)	Amoxicillin users (N = 5,000)
History of eczema	701 (14.0)	435 (8.7)

The proportion of patients with a previous history of eczema *is* considerably different between the supramycin and the amoxicillin groups (14.0% vs 8.7%). Because of this discrepancy, we can state that the potential confounder (previous eczema) is "associated" with the exposure (supramycin use).

Deciding whether a factor is "considerably different" across levels of an exposure is subjective, and does *not* involve statistical considerations. Clinical/scientific knowledge and common sense are used to decide whether the observed difference is important. In this case, previous eczema is nearly twice as common among supramycin-treated patients.

9.2.2 Evaluation of a Confounder: Association with Outcome

The next step is to evaluate whether the potential confounder, previous eczema, is associated with the study outcome, rash. An association of previous eczema with rash would mean that the incidence of rash is different among subjects with and without previous eczema. Information needed to address this question has not yet been presented. Table 9.5 describes risk factors for developing a rash during the study.

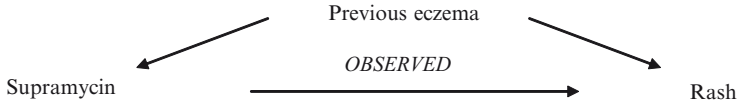
Table 9.5 Risk factors for the development of rash

	Relative risk of rash (95% confidence interval)
Supramycin use	2.53 (2.02, 3.01)
Age (per 10 year increase)	1.04 (0.92, 1.16)
African-American race	0.97 (0.87, 1.07)
Male	1.65 (1.10, 2.20)
Previous eczema	2.12 (1.77, 2.47)

Implicit in this table is that supramycin is being compared with amoxicillin, African-American race with Caucasian race, men with women, and previous eczema with no previous eczema

These data indicate that study subjects with a previous history of eczema are more than twice as likely to develop a rash, compared with those without previous eczema (relative risk = 2.12). These data establish an association between the potential confounder, previous eczema, and the outcome, rash. Combining this finding with the data from baseline characteristics table, we can now state that a third factor, previous eczema, is associated with *both* the exposure and the outcome in this study. Stated another way, study subjects who used supramycin are more likely to have previous eczema, *and* study subjects with previous eczema are more likely to develop a rash. Because of these linkages, it is possible that previous eczema, and not supramycin use, is causing the excess risk of rash observed among the supramycin users.

It can be useful to view the association between confounder, exposure, and outcome using a simple diagram. In this case:



To be classified as a confounder, a factor must be associated with both the exposure and the outcome; *association with just one of these elements is not sufficient*. If previous eczema is more common among supramycin-treated patients, but has nothing to do with developing a rash, then previous eczema could not confound the observed association of supramycin use with rash. Similarly, if previous eczema is linked with developing a rash, but the prevalence of previous eczema is similar among users and nonusers of supramycin, then previous eczema could not confound the observed association of supramycin use with rash.

9.2.3 Evaluation of a Confounder: Not in the Causal Pathway of Association

To fulfill the classical definition of a confounder, a factor must not only be associated with both the exposure and the outcome but also must *not* reside along the causal pathway of association. The causal pathway of association represents the mechanisms connecting an exposure with an outcome. In this case, supramycin might initiate an autoantibody response or trigger mast cells to degranulate prior to the appearance of a rash. These characteristics could be linked with both exposure and outcome if they are measured after the initiation of supramycin, as shown below.

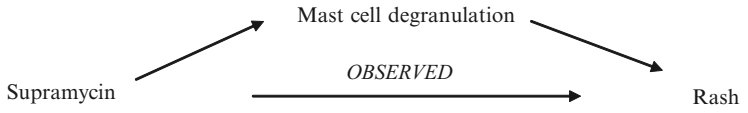
Autoantibody response and mast cell degranulation by exposure

	Supramycin users (N = 5,000)	Amoxicillin users (N = 5,000)
Autoantibody response	1,525 (30.5)	610 (12.2)
Mast cell degranulation	945 (18.9)	200 (4.0)

Autoantibody response and mast cell degranulation by outcome

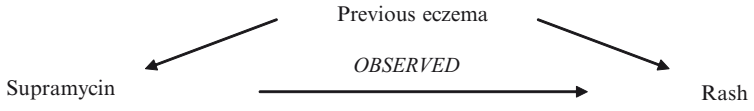
	Relative risk of rash (95% confidence interval)
Autoantibody response	9.5 (6.1, 12.9)
Mast cell degranulation	6.0 (4.2, 7.8)

It is tempting to conclude from these associations that autoantibodies and mast cell degranulation are confounding the observed association between supramycin use and rash. However, autoantibodies and mast cell degranulation *result* from supramycin use, in part explaining the *mechanism* by which supramycin is suspected to cause a rash. These factors do not confound the association between supramycin use and rash; rather, they explain it. The appropriate direction of causality between these factors can be appreciated in the following diagram:



Note the direction of the arrow connecting supramycin use with mast cell degranulation. In this example, supramycin use leads to mast cell degranulation, which then leads to a rash. Mast cell degranulation therefore lies *on the causal pathway of association* and would *not* be considered as a confounder, despite associations with the exposure and the outcome. No study data are used to judge whether a particular factor might reside on the causal pathway of association. This criterion is decided purely on scientific/biological and clinical knowledge.

In contrast to mast cell degranulation and autoantibodies, previous eczema does *not* reside on the causal pathway of association because previous eczema *precedes* supramycin use. As a result, previous eczema meets all three criteria for a confounding factor.

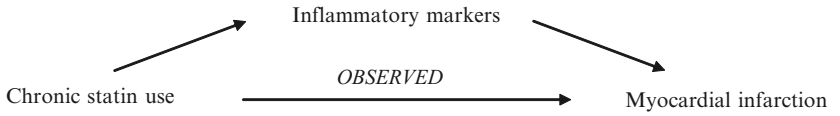


To summarize, a factor meets the classic definition of confounding in epidemiology when that factor is (1) associated with the exposure, (2) associated with the outcome, and (3) not on the causal pathway of association linking the exposure with the outcome.

9.2.4 *Other Examples of Factors That Reside on the Causal Pathway of Association*

Example 9.3. In an observational study, investigators found that chronic statin users experienced a lower risk of acute myocardial infarction compared with nonstatin users. The investigators also observed that chronic statin users had lower circulating levels of inflammatory markers, such as interleukin-6 and fibrinogen, which are independently linked with coronary disease risk. Are circulating inflammatory markers confounding the observed association between long-term statin use and myocardial infarction?

In addition to lowering lipid levels, statins also possess anti-inflammatory properties and can reduce circulating levels of inflammatory markers. Therefore, inflammatory markers are likely to reside in the causal pathway of association between statin use and myocardial infarction, representing a potential mechanism for the association and not confounding.



9.3 Scientifically Meaningful Versus Statistical Associations

Assessment of whether a potential confounding factor is associated with the exposure or the outcome should be based on subjective assessment. The use of statistical measures to identify these associations can result in misleading conclusions.

Example 9.4. In an observational study, investigators compared pharmacological therapy with psychotherapy as treatments for anxiety disorder. They observed more rapid improvement in anxiety symptoms among people who received pharmacological therapy; however, individuals who received pharmacological therapy were younger than those who received psychotherapy.

	Pharmacological therapy ($N = 100$)	Psychotherapy ($N = 100$)
Average age (years)	50.2	60.1

This 10-year average age difference is likely to be important; older people may have additional co-morbid illnesses that hamper their response to therapy, or may be less open to the idea of psychotherapy. It seems reasonable to conclude that age is "different" between the two therapy types, or that age is "associated" with the

exposure in this example. Whether 50.2 is "statistically different" from 60.1 is not of concern when evaluating age as a potential confounder.

Example 9.5. A small amount of protein in the urine (microalbuminuria) indicates dysfunction of vascular endothelial cells and predicts future cardiovascular risk. Using data from a large population-based study, investigators observed an association of microalbuminuria with greater carotid intimal-medial thickness, a marker of atherosclerosis. However, subjects with albuminuria were more likely to be male.

	Microalbuminuria (N = 4,554)	No microalbuminuria (N = 17,501)	p-Value
Male (%)	55.4	53.7	0.04

While it is true that the proportion of male subjects is "statistically" different between participants with and without microalbuminuria (p-value < 0.05), this difference is quantitatively small and unlikely to be meaningful in terms of confounding. The large sample size in this example explains the statistically significant result; however, statistical significance addresses only whether the observed differences may be due to chance, and not whether male sex may be confounding the association of microalbuminuria with carotid intimal-medial thickness to an important degree. As a general rule, confounding factors should be evaluated for *scientifically meaningful* associations with the exposure and outcome. The statistical significance of these associations (measured by p-values and confidence intervals) does not directly address the question of confounding.

9.4 Evaluation of a Confounder in Clinical Research Articles

The baseline characteristics table (usually the first table of a clinical/epidemiological study) is used to convey the distributions of subject characteristics with respect to the exposure or the outcome. In a cohort study or randomized trial, baseline characteristics are usually tabulated with respect to the exposure; in case-control studies, these characteristics are usually tabulated with respect to the outcome. The baseline characteristics table provides an opportunity for the reader to screen for potentially important confounding factors, recall the data presented in Table 9.3.

These baseline characteristics data demonstrate that age, race, and previous eczema are possible candidates for confounding because they differ by exposure status, whereas sex, smoking, and cardiovascular disease are unlikely to be important confounders. The next step for evaluating age, race, and previous eczema as possible confounders would be to determine whether age, race, and previous eczema are also associated with the outcome, rash.

9.5 Confounding-by-Indication

Confounding-by-indication is a specific type of confounding intrinsic to *observational studies of medication use*.[39] The concept is that the specific indication(s) for a particular medication, and not the effect of the medication itself, may be responsible for an observed association between the use of that medication and the study outcome. For example, previous observational studies have reported that loop diuretics, which increase urinary output, are associated with a greater risk of death and prolonged hospitalization among patients with acute kidney injury.[40] Importantly, the indication, or reason to prescribe a loop diuretic, is volume overload, which may indicate the presence of one or more serious underlying medical conditions such as heart failure or liver disease. It is possible that these medical conditions, and not potential adverse effects of loop diuretics, explain the observed association of loop diuretic use with death. In this example, the *indication* for prescribing a loop diuretic specifically confounded the association between loop diuretic use and adverse outcomes, as depicted in Fig. 9.1.

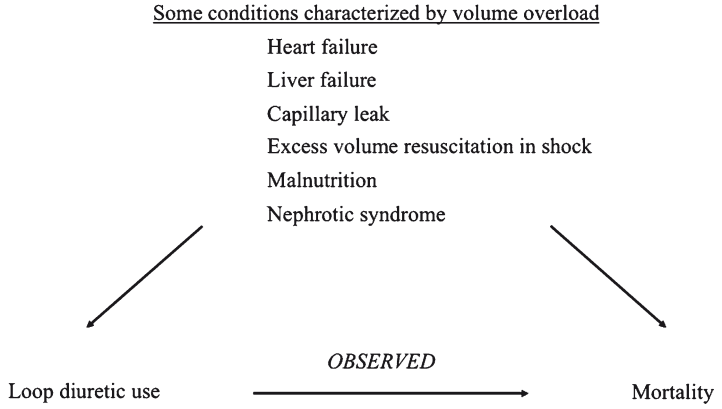


Fig. 9.1 Confounding-by-indication in observational studies of loop diuretic use. One or more conditions of volume overload represent the indication for prescribing a loop diuretic. These conditions, and not the effects of the diuretic, then influence the risk of mortality

Chapter 10
Methods to Control for Confounding

Learning Objectives

1. Methods used to control for confounding include:
   a. Restriction
   b. Stratification
   c. Matching
   d. Regression
   e. Randomization
2. Restriction can be a powerful method to address a limited number of confounders.
3. Restriction by the indication for a drug can be used to address confounding-by-indication in observational studies of medication use.
4. Stratification involves dividing the study population into strata, and then weighing and combining the stratum-specific results.
5. Restriction, stratification, and matching may not be appropriate for dealing with multiple confounding variables.
6. Randomization balances both measured and unmeasured characteristics.
7. A factor is likely to be confounding a given association if adjustment for that factor substantially changes the strength of the association.

In Chap. 9, a new antibiotic, supramycin, was found to be associated with a 2.53-fold greater risk of developing a rash, compared with amoxicillin. Further inspection of the study data revealed that supramycin-treated subjects were more likely to have a previous diagnosis of eczema, and that subjects with previous eczema were more likely to develop a rash. Because of these linkages, we concluded that previous eczema was confounding the observed association of supramycin use with rash. In this chapter, we focus on methods that can be used to control for (or *adjust* for) confounding, provided that the confounding factor(s) can be identified. *If* previous eczema is the only factor that is confounding the supramycin–rash association, then *adjusting* for previous eczema will yield the independent association of supramycin use with rash, strengthening inference for supramycin as a cause of rash.

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_10, © Springer Science+Business Media, LLC 2009

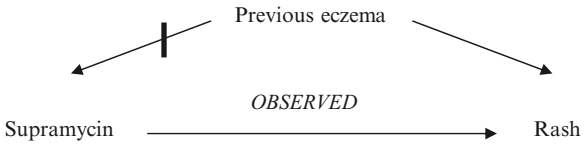
10.1 Restriction

10.1.1 Method of Restriction

In Chap. 9, the baseline characteristics table (Table 9.3) revealed that the prevalence of previous eczema was greater among supramycin users, compared to amoxicillin users.

	Supramycin users (*N* = 5,000)	Amoxicillin users (*N* = 5,000)
History of eczema	701 (14.0)	435 (8.7)

A simple approach would be to remove all 1,136 subjects in the study who have a previous history of eczema. This approach would "break" the link between previous eczema and supramycin use by restricting the study to only subjects without a history of eczema.



Once we remove subjects with previous eczema from the study, previous eczema can no longer confound the association between supramycin use and rash. Breaking the link between a factor and *either the exposure or the outcome* will eliminate the confounding influence of that particular factor, because a confounding factor must be associated with *both* the exposure and the outcome.

10.1.2 Pros and Cons of Restriction as a Means to Control for Confounding

Restriction is simple, easy to understand, and can completely eliminate the confounding influence of a particular factor. There are two important limitations of using restriction as a method to control for confounding: loss of study power and loss of generalizability. Removing the 1,136 subjects with previous eczema might leave too few remaining subjects to detect a statistically significant association. What if restriction was also used to deal with other potential confounding factors, for example, a previous medication allergy and a family history of eczema? Excluding patients who have a previous history of eczema, *or* a previous medication allergy, *or* a family history of eczema may remove a substantial proportion

of the study population. Moreover, even if the number of available subjects is adequate, the resulting study findings will apply only to people without *any* of the excluded conditions. Such results may not readily generalize to clinical practice where all types of patients are encountered. In general, restriction can be a powerful and convincing tool to control for one, or possibly a few of the most important confounding factors in a study.

10.1.3 Restriction to Control for Confounding-by-Indication

Restricting a study population to individuals who have an indication for a specific medication is an established method used to address confounding-by-indication in observational studies of medication use.

Example 10.1. Investigators wish to evaluate whether β-blockers, a class of medications that block activity of the sympathetic nervous system, can prevent the development of heart failure. They identify a group of β-blocker users and a group of nonusers from an electronic medical record system. All study subjects are free of heart failure at the beginning of the study. The investigators follow subjects for the development of incident heart failure. Contrary to their expectations, they find β-blocker use to be associated with a *greater* risk of future heart failure.

The paradoxical association of β-blocker use with heart failure may be confounded by the underlying characteristics of people who are prescribed β-blockers. This problem of confounding-by-indication, described in Chap. 9, may be addressed by *restricting the entire study population to individuals who have an indication for a β-blocker*. The most common indications for a β-blocker include hypertension, atrial fibrillation, and a previous history of myocardial infarction. Restricting the study population (β-blocker users and nonusers too) to individuals who have at least one of these conditions will help to minimize differences in characteristics between β-blocker users and nonusers.

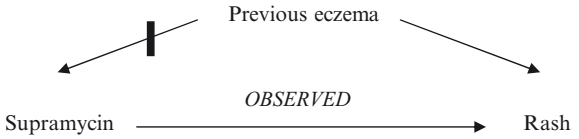
10.2 Stratification

10.2.1 Method of Stratification

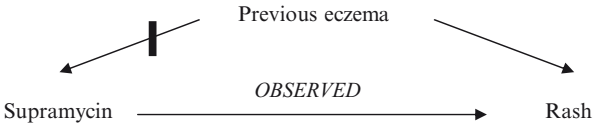
Returning to the supramycin study, an alternative strategy to excluding all individuals who have previous eczema is to separate study subjects according to their previous eczema status. The result would be two subpopulations, or *strata*. One stratum would consist exclusively of subjects with previous eczema. The other stratum would consist exclusively of subjects without previous eczema. Within each stratum,

the link between previous eczema and supramycin use will be broken, thereby eliminating confounding-by-previous eczema:

Stratum without previous eczema:



Stratum with previous eczema:



The next step is to calculate the relative risk of rash associated with supramycin use separately within each stratum, as demonstrated in Tables 10.1 and 10.2.

Because each stratum-specific relative risk cannot be confounded by previous eczema, it follows that these relative risks can be combined to obtain a summary relative risk that is also not confounded by previous eczema. A simple average of the two relative risks, (1.68 + 2.98)/2, would be inaccurate because most subjects belong to the stratum without previous eczema. A more precise method is to weight each stratum-specific relative risk by the proportion of subjects in the particular stratum and then combine the two weighted risks:

$$\begin{aligned}&\text{Combined relative risk}\\&= \text{relative risk}_{\text{no previous eczema}} * \text{weight}_1 + \text{relative risk}_{\text{previous eczema}} * \text{weight}_2\\&= 1.68 * (8,864/10,000) + 2.98 * (1,136/10,000) = 1.83\end{aligned}$$

Weighing the strata by the relative number of study subjects is an oversimplification of the actual weighting schemes that are used, but illustrates the concept of weighing and combining risks from separate strata. The combined relative risk of 1.83

Table 10.1 Stratum without previous eczema

	Number of patients	Number of rashes	Incidence proportion (%)	Relative risk
Supramycin	4,299	19	0.44	1.68
Amoxicillin	4,565	12	0.26	

Table 10.2 Stratum with previous eczema

	Number of patients	Number of rashes	Incidence proportion (%)	Relative risk
Supramycin	701	24	3.42	2.98
Amoxicillin	435	5	1.15	

represents the relative risk of rash, comparing supramycin to amoxicillin, *after adjustment for previous eczema.*

10.2.2 Pros and Cons of Stratification as a Means to Control for Confounding

The advantage of stratification, compared to restriction, is that the full study population is analyzed, preserving study power and maintaining generalizability. Stratification is particularly effective when dealing with *dichotomous* confounding variables (those that can take on only two distinct values), such as previous eczema, because the data can be completely separated into two discrete strata. Stratification is more difficult for continuous variables, such as age, because numerous strata must be created, for example, age <50 years, 50–60 years, 70–80 years, and >80 years.

The primary disadvantage of stratification is the inability to deal with multiple confounding factors simultaneously. To adjust for previous eczema, a previous medication allergy, *and* a family history of eczema, separate strata must be created for each combination of factors:

Strata 1: previous eczema, no previous medication allergy, no family eczema history
Strata 2: previous eczema, no previous medication allergy, family eczema history
Strata 3: previous eczema, previous medication allergy, family eczema history
Strata 4: previous eczema, previous medication allergy, no family eczema history, and so on…

If multiple confounding factors are considered, the individual stratum may become very small or disappear altogether (no patients in the stratum).

10.2.3 Stratum-Specific Associations

It is tempting to look within particular strata to examine how the association of interest varies across different groups of people within a study. In the above example, the association of supramycin use with rash was considerably stronger in people with a previous history of eczema (relative risk = 2.98) compared with those without previous eczema (relative risk = 1.68). However, natural variation across subgroups may account for observed differences in subgroup associations. The creation of multiple small subgroups will increase the likelihood of a spurious chance finding. The concept that the size of an association between exposure and outcome differs according to a third factor is called effect modification. *Effect modification is a different concept than confounding* and will be discussed in Chap. 11.

10.3 Matching

10.3.1 Method of Matching

Another method that is used to control for confounding is matching. Matching involves identifying groups of subjects within a study population who are the same *with respect to a confounder of interest.* To perform matching in the supramycin study, investigators would first identify a supramycin user and then ascertain their previous eczema status, for example, subject #1 is a supramycin user who had no previous eczema.

Next, they would identify one or more amoxicillin users who also do not have previous eczema. Within this matched group of only two subjects, previous eczema status is held constant, preventing confounding-by-previous eczema.

This matching process is then repeated, matching as many subjects as possible.

Supramycin users	Matched controls
Subject #1: Supramycin user, no previous eczema	Amoxicillin user, no previous eczema
Subject #2: Supramycin user, previous eczema	Amoxicillin user, previous eczema
Subject #3: Supramycin user, no previous eczema	Amoxicillin user, no previous eczema
Subject #4: Supramycin user, no previous eczema	Amoxicillin user, no previous eczema
Subject #5: Supramycin user, previous eczema	Amoxicillin user, previous eczema and so on

After identifying matches for the first 1,000 supramycin users, the following data are obtained:

	Supramycin users (N= 1,000)	Amoxicillin users (N = 1,000)
History of eczema	98 (9.8)	98 (9.8)

Since in this example each supramycin user is individually matched to one amoxicillin user who has the identical previous eczema status (yes vs no), the distribution of previous eczema is identical between the matched groups. The table of baseline characteristics appears similar to that of a randomized trial in terms of the matching factor, a history of eczema. The matching process has broken the link between the exposure and the confounder thereby eliminating previous eczema as a confounder.

Number of Matched Subjects

In the above example, each supramycin user was matched to a single amoxicillin user. More than one amoxicillin user could be matched to each supramycin user, if suitable amoxicillin users were available. Increasing the number of amoxicillin users in each matched group maintains the function of matching (to control for confounding) and increases study power.

Matching on Multiple Confounders

Matching permits adjustment for multiple confounding factors, provided that appropriate control subjects can be identified. For example, matching could be used to simultaneously address a previous history of eczema, a previous medication allergy, and a family history of eczema. To accomplish this procedure, investigators would first identify a supramycin user and ascertain their previous eczema status, medication allergy history, and family history of eczema: subject #1: supramycin user, no previous eczema, no medication allergy, no family history.

They would next attempt to match this supramycin user to one or more amoxicillin users who also have no previous history of eczema, no previous medication allergy, and no family history of eczema. Repeating this process for as many supramycin users as possible will balance the proportion of previous eczema, medication allergy, and family eczema history between the supramycin and amoxicillin groups.

Matching in Cohort Versus Case Control Studies

In the above examples, matching is performed according to the exposure status, supramycin use. In other words, exposed subjects (supramycin users) were matched to unexposed subjects (amoxicillin users) by the potential confounding variables. In general *matched cohort studies match on the exposure.*

In contrast, *matched case-control studies match on the outcome*. Imagine a case-control version of the supramycin study, in which investigators begin by identifying case individuals who develop a rash and control individuals without a rash. They would next ascertain and compare the proportion of supramycin use within cases and controls to make an inference regarding the association of supramycin use with rash. To perform matching in this case-control example, investigators would first identify a case patient and then ascertain their previous eczema status: subject #1: rash, no previous eczema.

They would next match this case subject to one or more control subjects without a rash who has the same previous eczema status. This matching process is then repeated, matching as many case subjects as possible. The result is a balanced proportion of previous eczema among the cases and controls. Importantly, matching potential confounding variables according to *either the exposure or the outcome* will adequately address confounding, because breaking only one of these links is all that is necessary to eliminate the confounding influence a variable.

10.3.2 Pros and Cons of Matching as a Means to Control Confounding

Advantages to matching are that it is an intuitive process and that it can address several confounders simultaneously, provided that a large enough population is available to find suitable matches.

One limitation of matching is that the matching procedure must be specified as part of the initial study design. In the supramycin case-control study, investigators matched case subjects who had a rash to control subjects without rash who were the same with respect to their previous eczema status. Once this matched cohort is created, there is no easy way to "go back" and further match on additional factors such as socioeconomic status.

A second disadvantage of matching is that it can be difficult to find suitable matches for multiple confounding factors. For example, given a person with rash who has previous eczema, a previous medication allergy, and a family history of eczema, it may be difficult to find many control subjects without rash who have the exact same characteristics, depending on the size of the available population.

A third limitation of matching is that factors selected to be matching variables can no longer be evaluated as disease risk factors. For example, matching on previous eczema status will create identical proportions of previous eczema between the cases and controls. This process effectively controls for previous eczema as a confounder, but prevents evaluation of previous eczema as a possible risk factor for the development of rash in this study.

10.4 Regression

A fourth method used to control for confounding is called regression, which is covered in Chaps. 18–19. Briefly, regression is a mathematical model that can estimate the independent association between many exposure variables and an outcome variable. Regression utilizes all of the study data, can account for multiple confounders simultaneously, and can deal with different types of potential confounding variables, such as those that are continuous and those that are dichotomous. Because of its flexibility, regression is the most commonly used method to deal with confounding in the medical literature. Disadvantages of regression are that the methods can sometimes be difficult to explain to a general audience, and that the results may be inaccurate if assumptions of the mathematical models are not satisfied.

10.5 Randomization

As previously discussed in Chap. 7, randomizing subjects to an exposure at the beginning of a study effectively breaks the link between the exposure and the potential confounding factors. Randomizing study subjects to receive either supramycin or amoxicillin will balance participant characteristics between these treatment groups. *Randomization has the important advantage of balancing both measured and unmeasured characteristics*, removing uncertainty as to whether the observed associations might be confounded by factors that were not measured in the study.

10.6 Interpreting Study Results After Adjustment for Confounding

Having discussed different methods available to control or adjust for confounding, we now turn to the task of interpreting study results after adjustment. Consider the following data in Table 10.3.

The result from the first row of this table can be interpreted as the association of supramycin use with rash *prior to any adjustment*. This result is also called the 'crude' or "unadjusted" relative risk. The result from the second row can be interpreted as the association of supramycin use with rash *independent of previous eczema*, and the result from the third row can be interpreted as the association of supramycin use with rash *independent of previous eczema and age*. If we can identify and adjust for *all* of the important confounders in a study, then we will be left with the same independent association between the exposure and the outcome that would be observed in a randomized trial.

10.7 Unadjusted Versus Adjusted Associations: Confounding

From the above table, the association of supramycin use with rash changes considerably after adjustment for previous eczema. In Chap. 9, we went through the laborious process of defining a confounder as a factor that is associated with the exposure, associated with the outcome, and not in the causal pathway of association. An alternative method to evaluate a potential confounding factor is to compare associations before and after adjustment for that factor. In table 10.3, adjustment for previous eczema substantially changes the relative risk from 2.53 to 1.83. This change implies that the original association of supramycin use with rash was confounded in part by previous eczema. On the other hand, further adjustment for age does not appreciably change the relative risk (from 1.83 to 1.80), implying that once previous eczema is controlled, age has minimal impact on the association of supramycin use with rash.

There is no general agreement as to how much change in the strength of an association is required for a factor to be dealt with as a confounder; some experts have argued for a 5–10% change. In the above example, the association of supramycin use with rash was substantially reduced by adjustment for previous eczema, implying that previous eczema was in fact confounding the crude association, and that analyses should adjust for previous eczema.

Table 10.3 Association of supramycin use with rash after adjustment

	Relative risk of rash
Supramycin (unadjusted)	2.53
Supramycin (adjusted for previous eczema)	1.83
Supramycin (adjusted for previous eczema, and age)	1.80

Relative risks in the table are comparing supramycin to amoxicillin

10.8 Confounding: An Advanced Example

Example 10.2. Ghrelin is a hormone that is produced in the stomach and stimulates appetite. Researchers at a local weight loss clinic are interested in whether high plasma ghrelin levels are related to late night snacking behavior. They recruit 200 participants from their weight loss clinic, measure plasma ghrelin levels, and administer a questionnaire to estimate the number of kilocalories consumed after 10 PM. A high plasma ghrelin level is defined as >600 pg/ml, and late-night snacking is defined by the consumption of more than 500 kilocalories after 10 PM. Study results are presented in Table 10.4.

Compared to participants who have normal plasma ghrelin levels, participants who have high ghrelin levels are more likely to snack at night (relative risk = 1.44). These findings are limited by the cross-sectional study design and by the possibility that another factor may be confounding the association of ghrelin levels with late-night snacking. One possible confounding factor is smoking. The researchers decide to use the *method of stratification* to adjust for smoking, as described in Tables 10.5 and 10.6.

Although high plasma ghrelin levels are associated with a 44% greater unadjusted risk of late night snacking among the full study population, high plasma ghrelin levels are *not* associated with late night snacking among smokers *or* nonsmokers in the study. The study population is divided into only smokers and nonsmokers; no participants were missed

Table 10.4 Association of ghrelin levels with late night snacking

	Late night snacking				
Plasma ghrelin level	Yes	No	Total	Proportion	Relative risk
High	18	32	50	0.36	1.44
Normal	30	90	120	0.25	

Table 10.5 Stratum one: smokers only

Plasma ghrelin level	Plasma ghrelin level		Plasma ghrelin level	Plasma ghrelin level	Plasma ghrelin level
	Yes	No			
High	16	4	20	0.8	1.0
Normal	24	6	30	0.8	

Table 10.6 Stratum two: nonsmokers only

	Late night snacking				
Plasma ghrelin level	Yes	No	Total	Proportion	Relative risk
High	2	28	30	0.067	1.0
Normal	6	84	90	0.067	

The *adjusted* relative risk of late night snacking can be calculated by weighing and combining the two stratum-specific relative risks. Since each stratum-specific relative risk is 1.0, the weighted and combined relative risk must also be 1.0. Therefore, there is no association of plasma ghrelin levels with late night snacking *after adjustment for smoking*. Because of the substantial change in relative risk from 1.44 (unadjusted) to 1.0 (adjusted for smoking), we can conclude that smoking strongly confounds the association of ghrelin levels with late night snacking. As a general rule, confounding will be present if *all* of the stratum-specific relative risks are substantially greater than or substantially less than the unadjusted relative risk.

Chapter 11
Effect Modification

Learning Objectives

1. Effect modification is present when the size of an association differs by another factor.
2. The presence of effect modification can suggest synergy between exposure variables.
3. Effect modification is a different concept from confounding; a particular characteristic can function as a confounder, an effect modifier, both, or neither in a given study.
4. The likelihood ratio test evaluates whether the size of an association is statistically different across two or more categories of another factor.
5. The *p*-value from the likelihood ratio test represents the probability of finding the observed difference in the size of an association across subgroups due to chance.

11.1 Concept of Effect Modification

Most clinical research studies report an *average measure of effect or association* across the entire study population. For example, in the supramycin study discussed in Chap. 10, supramycin use was associated with 1.80-fold greater relative risk of rash after adjustment. This 1.80-fold greater risk represents an average that was derived using data from all members of the study population.

Summary measures may conceal a marked variation in response among individuals within a study population. For example, stratum-specific analyses in Chap. 10 revealed that supramycin use was associated with a substantially increased risk of rash among individuals with previous eczema (relative risk = 2.98), but only a modestly increased risk of rash among individuals without previous eczema (relative risk = 1.68). *Effect modification*, also called interaction or synergy, is the concept that the *size of an effect or association differs according to another factor*. The presence

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_11, © Springer Science+Business Media, LLC 2009

of effect modification implies that different groups of people respond differently to a particular exposure or treatment. Another example of effect modification is presented below.

Example 11.1. Recombinant tissue plasminogen activator (rtPA) is an intravenous medication used to dissolve blood clots in patients with acute stroke. Investigators compared the effect of rtPA on stroke outcomes between men and women.[41] Among men with acute stroke, disability-free survival was 36.7% in the rtPA group and 38.5% in the placebo group, an absolute treatment difference of 1.8%. Among women with acute stroke, disability-free survival was 30.3% in the rtPA group and 40.5% in the placebo group, an absolute treatment difference of 10.2%.

These data demonstrate a considerably stronger treatment effect of rtPA among women with acute stroke compared with men. Stated another way, the *effect* of rtPA on stroke outcome is *modified* by sex. A number of possible mechanisms may explain the observed sex-specific response to treatment. Women tend to have higher circulating levels of plasminogen activator inhibitor-1, which is the target of rtPA therapy. Women may also present with less severe stroke symptoms than men, and therefore have lesions that may be more responsive to rtPA therapy.

11.2 Synergy Between Exposure Variables

The presence of effect modification may suggest that two or more factors are acting *in combination* (synergistically) to influence the outcome. A classic example of synergy is the association of smoking and heavy alcohol use with laryngeal cancer, as shown in Table 11.1.

Simply adding the laryngeal cancer rates for smokers and heavy drinkers does *not* yield the rate of laryngeal cancer for people who both smoke *and* drink. In this example, the *combination of smoking and heavy alcohol use* results in a substantially greater rate of laryngeal cancer than that predicted by adding the individual effects of these characteristics. Chronic alcohol use is thought to damage the protective mucosal layer of the larynx, thereby enhancing the carcinogenic effects of smoking.

Table 11.1 Association of smoking and heavy alcohol use with laryngeal cancer

Smoking	Heavy alcohol use	Rate of laryngeal cancer per 100,000 person years
No	No	6.0
No	Yes	14.3
Yes	No	31.2
Yes	Yes	117.4

11.3 Effect Modification Versus Confounding

Effect modification and confounding address distinct research questions. Confounding addresses the question of whether an observed association is likely to represent a *causal relationship*. If heavy alcohol use confounds the association of smoking with laryngeal cancer, then evidence for smoking as a *cause* of laryngeal cancer is weakened. Effect modification addresses the question of whether the size of an observed association or effect differs across a study population. If heavy alcohol use modifies the association of smoking with laryngeal cancer, as shown above, then smoking has different effects on laryngeal cancer depending on alcohol status. In a given study, a particular factor may act as a confounder, an effect modifier, both, or neither.

To demonstrate effect modification in tabular form, consider the association of supramycin use with rash stratified by previous eczema status, presented in Tables 11.2 and 11.3.

If supramycin does in fact *cause* a rash, then every 1,000 supramycin prescriptions would be expected to result in about 2 extra rashes among people without previous eczema. On the other hand, every 1,000 supramycin prescriptions would be expected to result in about 2 extra rashes among people with previous eczema, potentially motivating a different antibiotic choice or closer monitoring for these individuals. These subgroup findings help focus attention to a particular subgroup of patients that may be most susceptible to the adverse effects of a medication.

What about confounding? We noted previously in Chap. 10 that weighing and combining the stratum specific relative risks (1.68 and 2.98) yielded an adjusted relative risk (1.83) that differed considerably from the unadjusted relative risk. We therefore concluded that previous eczema was likely to be confounding the association of supramycin use with rash. In this example, previous eczema is functioning as both as a confounder and an effect modifier of the association between supramycin use with rash.

Table 11.2 Rashes among stratum without previous eczema

	Incidence proportion (%)	Expected number of rashes per 1,000 prescriptions
Supramycin	0.44	4.4
Amoxicillin	0.26	2.6
Attributable risk	0.18	1.8

Table 11.3 Rashes among stratum with previous eczema

	Incidence proportion (%)	Expected number of rashes per 1,000 prescriptions
Supramycin	3.42	34
Amoxicillin	1.15	12
Attributable risk	2.27	22

The expected number of rashes per 1,000 prescriptions is calculated from the incidence proportion data as: *expected number of rashes per 1,000 prescriptions = incidence proportion *1,000.*

11.4 Evaluation of Effect Modification

11.4.1 Epidemiologic Evaluation of Effect Modification

Effect modification should be considered when the size of an association is "substantially" different across levels of another factor. What exactly constitutes a "substantial" difference? The general principles used to assess subgroup findings were first presented in Chap. 7. A subgroup finding is more likely to be valid if:

(1) There is biological plausibility for a particularly strong effect or harm in the subgroup.
(2) The subgroup analysis was prespecified at the beginning of the study.
(3) There are a reasonably large number of outcomes in the subgroup.

For the supramycin example, there is some plausibility for an antibiotic to provoke an allergic rash more frequently among people who have an allergic skin condition (eczema). On the contrary, the subgroup analysis was not prespecified.

11.4.2 Statistical Evaluation of Effect Modification

In additional to the 3 general principles cited above a specific statistical test called the *likelihood ratio test* can be used to evaluate whether the size of an association is statistically different among subgroups. The *p*-value from the likelihood ratio test represents the probability of finding the observed subgroup results, or even more extremely different subgroups results, due to chance.

For example, consider the supramycin study data adding the statistical test for interaction:

	Attributable risk of rash
Stratum without previous eczema	0.18%
Stratum with previous eczema	2.27%
p-Value for interaction	0.03

The *p*-value for interaction can be interpreted as, "the probability of observing these different attributable risks, or even more different attributable risks, due to chance alone is 3%." Since this probability is low, we can reasonably conclude that the size of the association of supramycin use with rash *is* likely to differ according to previous eczema status. Stated another way, the *p*-value for interaction provides *statistical evidence* for effect modification. Some equivalent interpretations of the statistical test for interaction that may be encountered in clinical research articles are presented below.

(1) The association of supramycin use with rash was statistically different between people with and without previous eczema (*p*-value for interaction = 0.03).

(2) Previous eczema status statistically modified the association of supramycin use with rash.
(3) The association of supramycin use with rash was altered by previous eczema.

Specific definitions of *p*-values will be presented in Chaps. 15–17.

11.5 Effect Modification in Clinical Research Articles

In clinical/epidemiological research articles, tests for effect modification usually occur after presentation of the main summary data. Results of effect modification testing may be presented as figures, tables, or as text within the results section. Two examples of how effect modification data might be presented in clinical research articles appear below.

Example 11.2. Kidney disease and hypertension are frequently diagnosed simultaneously. Kidney dysfunction leads to salt retention and activation of hormones that can raise blood pressure; however, it remains unclear whether kidney disease predates the development of high blood pressure in the general population. To test this hypothesis, investigators studied a group of individuals who had normal blood pressure at the beginning of the study and used a blood test to estimate kidney function. They found that early kidney dysfunction was associated with an estimated 20% greater risk of developing hypertension later in life among the full study population. They next evaluated whether this 20% greater risk varied across the different subgroups within the study (Fig. 11.1).

Figure 11.1, typical of those found in clinical research articles, presents the incidence rate ratios (similar to relative risks) of hypertension associated with kidney dysfunction among subgroups defined by age, sex, and race/ethnicity. These findings demonstrate small variation in the size of the association between kidney dysfunction and hypertension across the subgroups tested. In other words, these data do *not* support effect modification by age, sex, or race/ethnicity in this particular study. The modestly stronger association of kidney dysfunction with hypertension among Hispanic participants is likely to represent natural variation within subgroups, though could motivate further scrutiny of kidney disease-hypertension relationships among Hispanic individuals in future studies.

Example 11.3. Adiposity leads to greater circulating concentrations of insulin-like growth factor-1, which may increase the risk of colon cancer. In postmenopausal women, adipose tissue is the main source of endogenous estrogen, which may counteract potentially carcinogenic effects of obesity. Investigators used long-term follow-up data from the Canadian National Breast Screening Study to explore possible menopause-related differences in the association of obesity with colon cancer.[42] Their findings are summarized in Table 11.4.

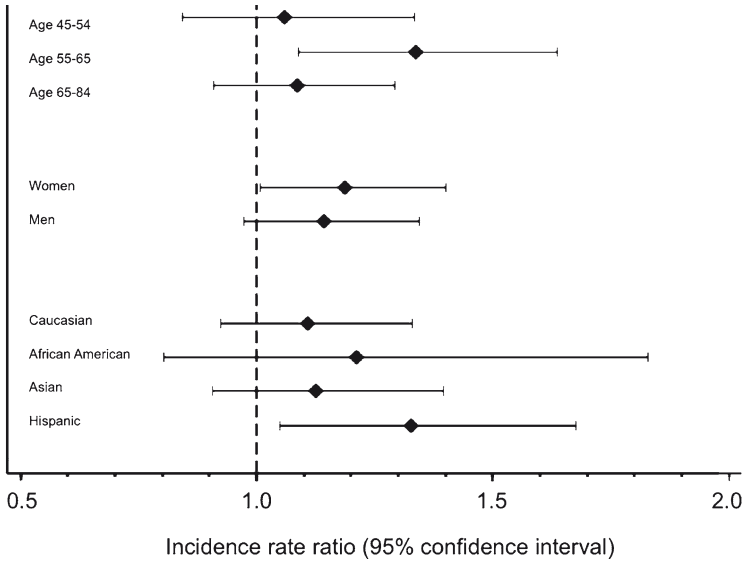


Fig. 11.1 Relative risk of developing hypertension associated with kidney dysfunction. *Diamonds represent relative risks. Gray lines represent 95% confidence intervals*

Among the entire cohort, obesity is not appreciably associated with colorectal cancer risk; relative risks of cancer comparing preobese and obese women to nonobese women are 1.03 and 1.08, respectively. When results are stratified by menopausal status, obesity is associated with a *greater* risk of colorectal cancer among premenopausal women (relative risk = 1.88), but a *lesser* risk of colorectal cancer among postmenopausal women (relative risk 0.73). The *p*-value for interaction of 0.01 can be interpreted as, "if the size of the association of obesity with colorectal cancer does not differ by menopausal status in the population, then the probability of observing these different subgroup findings, or even more different subgroup findings, is 1%." The notable differences in the size of the association, biological plausibility, prespecified analysis, and statistical evidence suggest that menopausal status *does* modify the association of obesity with colorectal cancer.

11.6 Effect Modification on the Relative and Absolute Scales

In this chapter, we have investigated effect modification by evaluating differences in both relative and attributable risks of disease. In the supramycin study example, we evaluated differences between the *attributable risks* of rash according to previous eczema status. In the colorectal cancer study example, we evaluated differences

Table 11.4 Association of and obesity with colorectal cancer by menopausal status

	Relative risk of colorectal cancer
Entire cohort (N = 500)	
Nonoverweight	*reference*
Overweight	1.03
Obese	1.08
Premenopausal (N = 200)	
Nonoverweight	*reference*
Overweight	1.06
Obese	1.88
Post menopausal (N = 300)	
Non-overweight	*reference*
Overweight	0.98
Obese	0.73
p-Value for interaction	0.01

Table 11.5 Hip fracture rates among men and women in a randomized trial

	Hip fracture rate (per 100,000 person-years)	Relative risk	Attributable risk (per 100,000 person-years)
Men (N = 1,000)			
Study drug	20	0.50	20
Placebo	40		
Women (N = 3,000)			
Study drug	230	0.79	60
Placebo	290		

between the *relative risks* of colorectal cancer according to menopausal status. Which procedure is correct?

Clinical research articles typically present analyses of effect modification using relative risks, similar to examples 11.2 and 11.3. However, the significance of effect modification is best appreciated by examining differences in attributable risks, as demonstrated in example 11.4.

Example 11.4. A new medication is developed that improves bone mineral density among individuals with osteoporosis. The medication is evaluated in a clinical trial of 4,000 men and women with clinical osteoporosis who are randomly assigned to receive either the new study drug or a placebo. Study participants are followed for the first occurrence of hip fracture.

Table 11.5 demonstrates that the new medication lowers the risk of hip fracture in both men and women. The difference in relative risks suggests that the medication may have a stronger protective effect among men (50% vs 21% lower relative risk of hip fracture). However, most of the fractures in the study occurred among

women, and the new medication prevented considerably more hip fractures in women compared with men. This distinction is appreciated by comparing the attributable risks, which directly relate to the number needed to treat (see Chap. 7). Treating 10,000 men with the new medication for 10 years would prevent, on average, 20 hip fractures. Treating 10,000 women with the new medication for 10 years would prevent, on average, 60 hip fractures. These results indicate that the drug would have the greatest clinical impact when prescribed to women.

Chapter 12
Screening

Learning Objectives

1. Important characteristics of diseases that are appropriate for screening include:
   (a) The disease should be important in the screened population.
   (b) The disease process should have a preclinical phase.
   (c) Treating the disease process at an early stage should provide benefit.
2. Reliability refers to the ability of a test to provide repeatable results.
3. Validity refers to the ability of a test to detect true disease, as defined by a gold standard.
4. Two important measures of validity are sensitivity and specificity:
   (a) Sensitivity is the probability of testing positive given the presence of disease.
   (b) Specificity is the probability of testing negative given the absence of disease.
5. The predictive values of a test are defined as:
   (a) Positive predictive value is the probability of true disease given a positive test.
   (b) Negative predictive value is the probability of no true disease given a negative test.
6. Disease prevalence is required for calculating the predictive values of a test.
7. High test specificity is needed to reduce false positives when screening for a rare disease.
8. ROC curves present sensitivity and specificity characteristics for all possible cutoff values of a continuous screening test.
9. Four potential biases in screening studies that may lead to spurious associations of a screening program with health outcomes are
   (a) Referral bias
   (b) Lead time bias
   (c) Length bias sampling
   (d) Overdiagnosis bias
10. The association of a factor with disease must be extraordinarily strong to qualify that factor as a potentially useful screening test.

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_12, © Springer Science+Business Media, LLC 2009

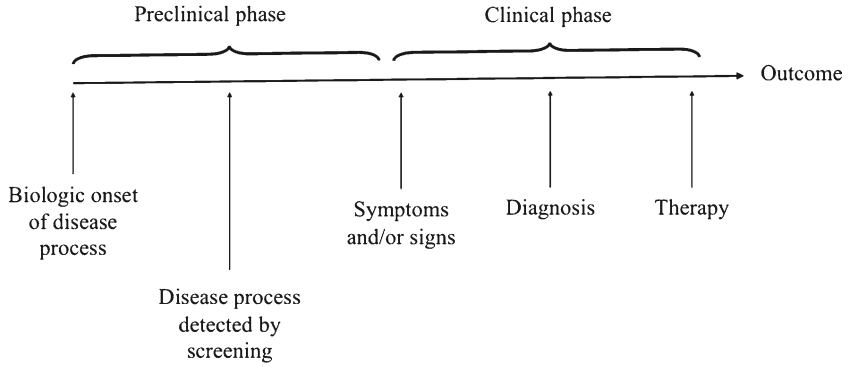


Fig. 12.1 Preclinical and clinical phases of a disease process

12.1 General Principles of Screening

Previously, we used epidemiologic principles to gain insight into the *causes* of disease. In the next two chapters, we will examine screening and diagnostic testing as clinical tools for *detecting* and treating disease.

Screening refers to the early detection of a disease or condition in the *preclinical phase*, defined as the period before clinical symptoms or signs are present. The identification of previously unrecognized disease can lead to subsequent interventions that impact the course of the disease, for example screening mammography can detect breast cancer at an early stage before it is clinically apparent and surgery plus chemotherapy given at an early stage can cure the disease.

Ideally, screening tests are applied to clinical conditions that progress in a series of ordered steps, as depicted in Fig. 12.1.

For example, colon cancer may begin with a cluster of abnormal cells, then progress to an adenomatous polyp, and then to carcinoma. Colonoscopy can detect (and remove) colonic polyps before the development of overt carcinoma.

12.2 Qualities of Diseases Appropriate for Screening

12.2.1 The Disease should be Important in the Screened Population

Generally, screening tests focus on *serious diseases*. Examples include screening for colon cancer in middle-aged adults, and screening for phenylketonuria in newborns. Detection of these potentially fatal conditions can lead to interventions that dramatically reduce mortality.

12.2.2 Early Recognition and Treatment of the Disease Should Prevent Clinical Outcomes

Detecting untreatable conditions earlier in their course can increase patient anxiety without influencing the disease process. For example, electron beam computed tomography (EBCT) is a specialized scanning procedure that is used to detect asymptomatic coronary artery disease. The EBCT scan can rapidly quantify the extent of coronary artery calcification, a marker of atherosclerosis. However, EBCT generally cannot distinguish high-grade coronary lesions that require surgical intervention from diffuse low-grade atherosclerotic plaques. Further, medications that are used to treat coronary artery disease, such as cholesterol-lowering drugs, have minimal impact on coronary calcification. While EBCT testing is attractive and uses modern technology, few proven treatment strategies are available for people who have a positive test.

12.2.3 The Disease Should have a Preclinical Phase

It would be difficult to screen for a condition like the common cold, because the time from biologic onset of disease to clinical symptoms is so short. On the other hand, other diseases, such as colon cancer, have an ordered preclinical phase that can be detected by the presence of histologic findings or specialized radiographic imaging studies, or specific biomarkers.

12.3 Qualities of Screening Tests

12.3.1 General Qualities

To achieve widespread use, a screening test ideally should be easy to administer, relatively inexpensive, and safe. Many blood tests and imaging studies satisfy these criteria, for example, the prenatal "triple screen" blood test that is used to screen for trisomy 21 during pregnancy and the chest X-ray that is used to screen for tuberculosis.

12.3.2 Reliability and Validity

Beyond these general qualities, screening tests are generally judged by their *reliability* and *validity. Reliability* refers to the ability of a test to provide *consistent* results. For example, the HIV antibody test is considered to be reliable, because the test will return a consistent result, positive or negative, within in a given individual on the same day. On the other hand, the potassium hydroxide test for diagnosing

cutaneous fungal infection may yield different results when repeated on the same individual, due to sampling variation, differences in specimen preparation, and the subjective opinion of the individual tester who is looking under the microscope.

Validity refers to the ability of a test to detect *true disease*, as defined by some *gold-standard* measurement technique. For example, the validity of mammography for detecting breast cancer is typically judged against the gold standard diagnosis, which is made by breast biopsy and pathologic examination. The validity of serum creatinine levels for detecting kidney function is judged against formal measurement of the glomerular filtration rate performed using an intravenous tracer. Other clinical conditions are best diagnosed by expert opinion. For example, the gold standard diagnosis of heart failure in many clinical studies is considered to be the expert opinion of a panel of cardiologists, who review each patient's medical chart. In general, gold standard testing is invasive, expensive, or not practical to apply to a large population for screening purposes.

Another perspective on reliability and validity is to think of an unreliable test as having *random error* and an invalid test as having *systematic error*, as illustrated in Fig. 12.2.

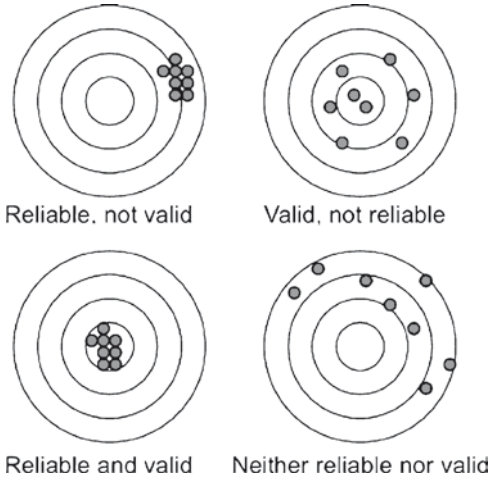


Fig. 12.2 Reliability versus validity of a screening test

12.4 Validity of Screening Tests

12.4.1 Sensitivity and Specificity

The *validity* of a screening test is typically described by the *sensitivity* and *specificity* of the test. These terms describe how well the test performs compared to a gold standard test.

Sensitivity and specificity can be explicitly defined using a 2×2 table, in which true disease status is presented across the top of the table and test result status is presented on the left-hand side of the table as presented in Table 12.1.

Table 12.1 Sensitivity and specificity of a screening test

	Disease	
	Yes	No
Test result		
Positive	*a*	*b*
Negative	*c*	*d*

$$\text{Sensitivity} = \frac{\text{Number who test positive with disease } (a)}{\text{Number with disease } (a+c)}$$

$$\text{Specificity} = \frac{\text{Number who test negative without disease } (d)}{\text{Number without disease } (b+d)}$$

Sensitivity is the probability of testing postive given the presence of disease.
Specificity is the probability of testing negative given the absence of disease.

The presence or absence of disease in these definitions refers to a gold standard method.

For example, the sensitivity of mammography for detecting breast cancer among women over 50 years old is about 85%. The interpretation of this sensitivity value is, "among women with biopsy proven breast cancer, the chance of having a positive mammogram is 85%."

The specificity of mammography for detecting breast cancer among women over 50 years old is about 95%. The interpretation of this specificity value is, "among women with biopsy proven absence of breast cancer, the chance of having a negative mammogram is 95%."

Sensitivity and specificity characteristics generally remain consistent across different populations, or may vary to only a small degree. However, sensitivity and specificity characteristics do not provide important clinical information for individual patients. In the mammography example, women typically would not be interested in their probability of having a positive mammogram after they are diagnosed with breast cancer. Instead, they would like to know the opposite information, specifically, what is their chance of having breast cancer given a positive or negative mammography result? To answer this more clinically relevant question, two additional characteristics of screening tests are needed.

12.4.2 Positive and Negative Predictive Value

Positive predictive value is the probability of disease given a positive test result.

$$\text{Positive predictive value} = \frac{\text{Number who test positive with disease } (a)}{\text{Number who test positive} (a+b)}$$

$$\text{Negative predictive value} = \frac{\text{Number who test negative without disease } (d)}{\text{Number who test negative} (c+d)}$$

Negative predictive value is the probability of no disease given a negative test result.

For example, the positive predictive value of mammography is 10% in a low-risk patient population. The interpretation of this positive predictive value is, "among low-risk women who have a positive mammogram, the probability of breast cancer is 10%."

The negative predictive value of mammography in this same population is 98%. The interpretation of this negative predictive value is, "among low-risk women with a negative mammogram, the probability of having breast cancer is 2%."

Unlike specificity and sensitivity, positive and negative predictive values depend on the prevalence of disease in the screened population. The relationships between sensitivity, specificity, positive and negative predictive values are illustrated in the following examples.

Example 12.1 The sensitivity and specificity of the hepatitis C antibody test for detecting hepatitis C infection are 99% and 95%, respectively. What is the positive predictive value of the hepatitis C antibody test for detecting hepatitis C infection?

Given only the sensitivity and specificity characteristics of a test (Table 12.2), it is *not possible* to determine the positive or negative predictive value. More information is needed.

Table 12.2 Hepatitis C antibody testing: disease prevalence unknown

	Disease		
	Yes	No	
Test result			
Positive	*a*	*b*	?
Negative	*c*	*d*	?
	Sensitivity = $a/(a + c) = 0.99$	Specificity = $d/(b + d) = 0.95$	

Positive predictive value = a / (a+b) = ?

Example 12.2 The sensitivity and specificity of the hepatitis C antibody test for detecting hepatitis C infection are 99% and 95%, respectively. Among United States veterans, the prevalence of hepatitis C infection is 10%. What is the positive predictive value of hepatitis C antibody testing for detecting hepatitis C infection among United States veterans?

The prevalence data, which indicate that 10% of the population has the disease, are needed to determine the positive and negative predictive values of the test. A useful method for calculating predictive values for these types of problems is to first create a hypothetical population of any size, 1,000 is usually a good round number, and then to use the prevalence data to first fill in the cells for disease and no disease as shown in Table 12.3.

In this example, the positive predictive value of the hepatitis C antibody test is 99/144 = 69%. The interpretation of this result is, "a U.S. veteran who tests positive for hepatitis C antibody has a 69% chance of having hepatitis C."

Table 12.3 Hepatitis C antibody testing: 10% prevalence of disease

	Disease		
	Yes	No	Total
Test result			
Positive	*a*	*b*	
Negative	*c*	*d*	
Total	100 (10%)	900 (90%)	1,000 total

Prevalence of disease = (a + c) / (a + b + c + d)

The next step is to use the sensitivity and specificity data to fill in cells *a* and *d*.

	Yes	No	Total
Test result			
Positive	Sensitivity = 99% 100 × 0.99 = 99		
Negative		specificity = 95% 900 × 0.95 = 855	
Total	100	900	1,000 total

Now there is enough information to complete the rest of the table.

	Yes	No	Total
Test result			
Positive	100 × 0.99 = 99	45	144
Negative	1	900 × 0.95 = 855	856
Total	100	900	1,000 total

Positive predictive value = a / (a + b) = 99 / 144 = 69%

The negative predictive value of the hepatitis C antibody test is 855/856 × 100% = 99.9%. The interpretation of this result is, "a U.S. veteran who tests negative for hepatitis C antibody has a 99.9% chance of not having hepatitis C." Note that the "true" diagnosis of hepatitis C refers to the use of a gold-standard method. The polymerase chain reaction, or PCR test for hepatitis C viral antigen is a gold-standard method that is used for detecting hepatitis C.

Example 12.3 The prevalence of hepatitis C infection among intravenous drug users is 30%. An intravenous drug user undergoes hepatitis C antibody testing and tests positive. What is the probability that this person has hepatitis C infection?

Setting up the table using the new disease prevalence of 30% and the same sensitivity and specificity information from the previous example yields the following data in Table 12.4:

The positive predictive value of the hepatitis C antibody test among intravenous drug users has now increased to 297/332 = 90%. The sensitivity and specificity of the test have remained fixed. Given a positive hepatitis C antibody test result, there is now a 90% chance that this person has hepatitis C. This is not surprising since this person had a higher "baseline" risk of hepatitis C prior to antibody testing, due to their use of intravenous drugs. The negative predictive value is now slightly lower than that of Example 13.2, again because the "baseline" risk of hepatitis C is higher prior to testing.

Table 12.4 Hepatitis C antibody testing: 30% prevalence of disease

	Disease		
	Yes	No	
Test result			
Positive	300 × 0.99 = 297	35	332
Negative	3	700 × 0.95 = 665	668
Total	300	700	1,000 total

Positive predictive value = a / (a +b) = 297 / 332 = 90%
Negative predictive value = d / (c + d) = 665 / 668 = 99.6%

Example 12.4 How many false negative results will occur from screening 10,000 intravenous drug users with the hepatitis C antibody test?

Notice that the top left and bottom right cells of the 2 × 2 screening test tables indicate agreement between test result and disease status, whereas the top right and bottom left cell indicate disagreement, and represent false positive and false negative test results, respectively, as shown in Table 12.5.

Table 12.5 Agreement between disease status and test result

	Disease	
	Yes	No
Test result		
Positive	True positive	False positive
Negative	False negative	True negative

Testing 10,000 intravenous drug users would yield 30 false negative hepatitis C test results, as illustrated in Table 12.6.

Table 12.6 False negative test results from screening 10,000 individuals for hepatitis C

	Disease		
	Yes	No	
Test result			
Positive	3,000 × 0.99 = 2970	350	3,320
Negative	**30**	7000 × 0.95 = 6650	6,680
Total	3,000	7,000	10000

Bold text indicates number of false negative results.

Example 12.5 A new rapid blood test for detecting pulmonary embolism (a blood clot in the lungs) is developed. The company reports the test to be "extremely accurate," with a sensitivity and specificity of 99% compared to gold-standard pulmonary angiography. Excited by the possibility of reducing missed diagnoses of pulmonary embolism, a busy emergency department decides to administer the new test routinely to all emergency room patients who have shortness of breath. The clinical decision plan is for a positive test to be followed by a pulmonary angiogram, which has an approximate 1% risk of major complications. If 10,000 patients present to this emergency room with shortness

Table 12.7 False positive test results from screening 10,000 individuals

	Disease		
	Yes	No	
Test result			
Positive	100 × 0.99 = 99	99	198
Negative	1	9,900 × 0.99 = 9,801	9,802
Total	100	9,900	10,000 total

of breath during the first year of testing, how many pulmonary angiograms will be performed on patients *who do not have* true pulmonary embolism?

To answer this question an estimate of disease prevalence is needed. Consulting the literature, the prevalence of pulmonary embolism among patients presenting to the emergency room with shortness of breath is estimated to be about 1%. Table 12. 7 demonstrates the consequences of administering the new screening test to 10,000 emergency room patients with shortness of breath.

The positive predictive value of the pulmonary embolism test is 99/198 = 50%; half of all positive test results will be false positives. There will be 99 pulmonary angiograms performed among patients without true pulmonary embolism, resulting in about one expected major complication. This example illustrates the difficulty of screening for a rare disease, even when using a supposedly "highly accurate" test. *The key test characteristic for avoiding false positives when screening for a rare disease is specificity*, which affects the vast majority of the screened population. In this case, a specificity of 99% is not high enough to avoid many false positive test results. An even higher specificity of 99.9% or 99.99% may be needed for this screening test to produce more benefit than harm.

12.4.3 Screening Tests with Continuous Values

Sensitivity, specificity, positive and negative predictive values apply to screening tests that return *dichotomous* values, meaning that the test result can either be positive or negative. Many screening tests yield naturally *continuous* results, in which a wide range of values is possible. One approach to dealing with continuous test results is to select some cutoff value for defining a "positive" versus "negative" test.

For example, consider the use of the serum prostate specific antigen (PSA) level to screen for prostate cancer in men. The PSA test does not naturally return a "positive" or "negative" result, but instead returns a continuous serum level, from 0 to infinity. What PSA cutoff value should be used to define a positive versus negative PSA test? This question can be addressed by calculating the sensitivity and specificity of several different possible cutoff values, as shown in Table 12. 8.

Selecting a PSA cutoff value of 0 ng/ml means defining a negative test to be a PSA level = 0 ng/ml and defining a positive test to be a PSA level >0 ng/ml. A PSA cutoff value of 0 ng/ml would yield a test with extremely high sensitivity, defined by the proportion of men with prostate cancer who test positive (PSA level >0 ng/ml).

Table 12.8 Prostate specific antigen screening for prostate cancer

PSA level cutoff value (ng/ml)	Sensitivity	Specificity
0	1.0	0
1	0.90	0.20
2	0.85	0.40
4	0.80	0.60
8	0.50	0.80
10	0	1.0

However, the cutoff value of 0 ng/ml also yields a test with extremely low specificity, defined by the proportion of men without prostate cancer who test negative (PSA level = 0 ng/ml).

Raising the cutoff value to define a positive PSA test result will increase the specificity of the test at the expense of sensitivity. The selection of a specific cutoff value depends on the intended use of the test. For example, a screening test with high specificity may be desired when screening for a rare disease in the general population, in order to reduce the frequency of false positives. For the PSA test, investigators would likely seek some reasonable balance between sensitivity and specificity; lower test sensitivity would result in missing some potentially important cancers, whereas lower test specificity would result in false positive tests that might trigger an unnecessary prostate biopsy with resultant complications. From the above table, a PSA cutoff level in the 2–4 ng/ml range seems to best balance sensitivity and specificity characteristics.

The tradeoff between sensitivity and specificity for a particular test can be presented graphically using a *receiver operating characteristic (ROC) curve*. These curves plot test sensitivity on the *Y*-axis vs. 1 − specificity on the *X*-axis. Figure 12.3 presents an ROC curve for the PSA test.

The solid line on the ROC curve represents the performance of the PSA test, and the dashed, diagonal line represents the performance of a *hypothetical test that is completely uninformative*. Another perspective is to think of the null, diagonal line on an ROC curve as representing sensitivity and specificity values that would be expected due to chance alone.

To illustrate the performance of an uninformative test, consider the use of patient height to screen for prostate cancer. If the cutoff value for defining a positive test is set very low, such as 5 ft., then the test will have high sensitivity, because virtually all patients with disease will test positive on the basis of being greater than 5-ft. tall. On the other hand the test will also have minimal specificity, since most, if not all patients without prostate cancer will also test positive. This low cutoff value of 5 ft. would be represented in the top right-hand corner on the dashed line, indicating high sensitivity, but low specificity.

As the cutoff value for defining a positive height test is raised, the specificity of the test will increase monotonically at the expense of reduced sensitivity, since patient height cannot really detect prostate cancer. Setting the cutoff value to the greatest height, say 8 ft., would lead to a negative test result in all patients with prostate cancer (sensitivity of 0%), and all patients without prostate cancer (specificity of

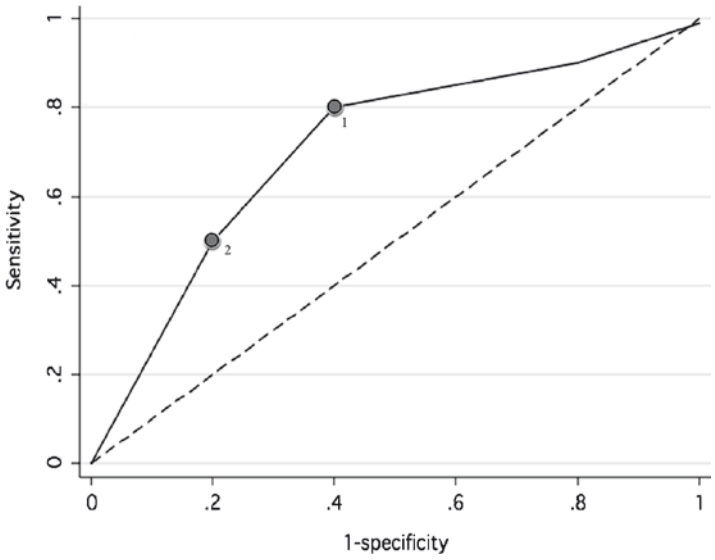


Fig. 12.3 Receiver operating characteristic curve for the PSA test

100%). This high cutoff value of 8 ft. would be represented in the lower left-hand corner on the dashed line, indicating low sensitivity, but high specificity.

In contrast, the PSA test demonstrates an increase in sensitivity without monotonic loss of specificity throughout the range of possible cut points. This quality indicates a good screening test. A perfect test would start at the lower left-hand corner of the graph, rapidly increase in near vertical fashion toward perfect sensitivity, and then move horizontally across the graph to the upper right-hand corner. Every continuous test will have its own a unique ROC curve.

As a general rule, cutoff values in the upper left-hand region of an ROC curve represent a reasonable trade-off between test sensitivity and specificity. For the PSA test, data point #1 on the graph, which has a sensitivity of 0.8 and specificity of 0.6, represents one reasonable choice.

Notice that the specific PSA values used as cutoff points are not displayed on the ROC curve. These values are typically found in a separate table or in the text of a journal article. Consulting Table 12.8, a sensitivity of 0.8 and specificity of 0.6 corresponds to a specific PSA cutoff value of 4 ng/ml. Investigators could choose a higher PSA cutoff value to increase test specificity, if their primary concern was to minimize false-positive tests. However, inspection of the ROC curve reveals that increasing PSA test specificity (moving to left along the PSA line to data point #2) comes at the expense of a considerable loss in test sensitivity, meaning that more cancers would be missed.

Importantly, positive and negative predictive values for the PSA test cannot be determined from the ROC data alone, because positive and negative predictive values for this test will depend on the prevalence of prostate cancer in the particular screened population. The ROC curve presents only sensitivity and specificity information.

ROC curves further allow for quantification of the overall quality of a continuous test by calculating the *area under the curve* between the specific test and the hypothetical uninformative test. A perfect test would yield an area under the curve of 0.5, which is exactly half of the graph area. The PSA test would probably yield an area under the curve that is closer to 0.2, which is still reasonable for a screening test. Using the area under the curve concept, it is possible to determine whether a screening test performs "significantly" better than chance alone using the *C-statistic*. A C-statistic *p*-value <0.05 would indicate that a continuous screening test performs better than chance alone across the full measured range of test values, or that the ROC curve is significantly different from the dashed diagonal line. The C-statistic does not however, inform the reader about which test values might be suitable cut-points.

Another important application of ROC curves is to compare different screening modalities. For example, the ROC curve for a new prostate cancer biomarker could be compared with the ROC curve for the standard PSA test. Superimposing these ROC curves on the same graph would provide a sense of which biomarker has the greater test validity.

12.5 Reliability of Screening Tests

12.5.1 Variation in Measurement Tools and Within an Individual

Sensitivity, specificity, positive and negative predictive characteristics address test validity, or how well a screening test compares to a gold standard. Another important consideration is the ability of a test to provide consistent results (reliability). In general, the ability of a screening test to detect disease will be diluted if the test yields highly variable results.

One component of test variability derives from inherent variation in the measurement tool itself. As described in Chap. 8, many data collection tools in clinical/epidemiological research are characterized by some degree of variability, for example the use of a questionnaire or a standard blood pressure cuff. For a blood test such as PSA, the measurement tool is the laboratory assay. The PSA assay has a variability of about 5%, meaning that slightly different PSA measurements will be obtained *from an identical blood sample.*

Another component of test variability derives from the day-to-day biologic variation in the characteristic under evaluation. Figure 12.4 presents PSA measurements obtained from the same individual on nine different days.

If the screening measurement of interest is the *mean PSA level*, represented by the dashed line, then a single PSA measurement on any particular day will imprecisely represent this value. Another term for this type of biological variation is *intra-individual variability*. As laboratory assay technology continues to advance, intra-individual variability has emerged as the major component of biomarker variation. An approach to this problem is to perform *repeated measurements within a given individual*, if such individuals are willing to undergo multiple tests.

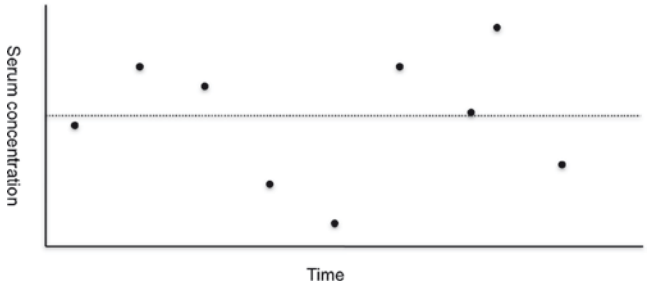


Fig. 12.4 Intra-individual variation in PSA test results

12.5.2 Measures of Reliability

The reliability of a dichotomous test can be described using a 2×2 table in which the first and second test results are presented on the left-hand side and top of the table, respectively. Consider the PSA test with a cutoff level of 4 ng/ml to define a positive test. The test is administered to 20 volunteers, and then repeated in the same individuals 1 month later to generate the data in Table 12.9:

Table 12.9 Reliability of the PSA test

		SECOND TEST		
		POSITIVE	NEGATIVE	TOTALS
FIRST TEST	POSITIVE	5 *tests agree*	3 *tests disagree*	8
	NEGATIVE	2 *tests disagree*	30 *tests agree*	32
	TOTALS	7	33	40

A simple way to summarize these findings is to calculate the proportion of tests that agree:

$$\text{Percent agreement} = \frac{\text{number of tests that agree}}{\text{total number of tests}} = 35/40 = 0.88 \times 100\% = 88\%$$

However, test agreement due to chance alone is possible, even with an imprecise test. The *Kappa statistic* is used to describe test agreement beyond that expected from chance alone. Calculation of the Kappa statistic is presented below.

$$\text{Kappa} = \frac{\left(\text{percent agreement} - \text{chance agreement}\right)}{\left(1 - \text{chance agreement}\right)}$$

where chance agreement = (a + b/total) * (a + c/total) + (b + d/total) * (c + d/total)

For this example,

Percent agreement = 0.88
Chance agreement = (8/40) × (7/40) + (33/40) × (32/40) =.035 + 0.66 = 0.7

$$\text{Kappa} = \frac{\left(0.88 - 0.7\right)}{\left(1 - 0.7\right)} = 0.6$$

The Kappa statistic ranges from +1 (perfect agreement), to 0 (no agreement beyond that expected from chance), to −1 (perfect disagreement). In general, a Kappa statistic <0.2 is considered poor agreement, 0.2 – 0.6 is considered fair agreement, and >0.6 is considered good agreement. A common application of the Kappa statistic is to measure agreement between two different testers who evaluate a subjective test characteristic, for example two independent radiologists scoring the same image or two independent pathologists interpreting the same biopsy.

The reliability of continuous tests can be described using a number of common summary measures. A simple and commonly used measure for continuous tests is the coefficient of variation, which is defined as the standard deviation of the test divided by its mean value. Lower values for the coefficient of variation indicate a more reliable test.

12.6 Types of Bias in Screening Studies

Once a particular screening test is found to be reasonably valid and reliable, the next step is to evaluate whether the test can provide useful information related to patient outcomes. Studies of screening tests are subject to the same considerations and biases as other epidemiological studies; however, there is a unique set of potential biases that are specific to studies of screening tests. The most definitive way to avoid these biases is to conduct a randomized clinical trial of the screening program.

12.6.1 Referral Bias

Observational studies that compare the outcomes of screened individuals to those of unscreened individuals are subject to confounding by characteristics that may be linked with referral for screening. On one hand, screened individuals may be highly educated, take a particular interest in their personal health, or receive care from a

physician that is particularly attentive to other health measures. These characteristics, and not the screening test itself, may explain a relative benefit observed in the screened group. On the other hand, individuals who undergo screening may have an inherently greater risk for a particular disease, due to a family history or high-risk behaviors. These characteristics have the potential to make a screening test appear harmful.

As with other observational study designs, randomization represents the ideal solution to the problem of confounding. If ethically possible, a randomized trial would assign individuals to receive either the specific screening program or no testing. If randomization is not possible, other approaches to confounding include regression, matching, and stratification, as described in Chap. 10. Central to these procedures is the consideration and accurate measurement of suspected characteristics that may differ between screened and unscreened individuals.

12.6.2 Lead Time Bias

A more complex problem in screening studies is that differences in the biologic duration of a disease may create the false impression that a screening test is protective. Consider the example of a hypothetical new blood test that is developed to detect early dementia. The new test is administered to 500 older adults who do not have dementia symptoms; 100 of them test positive. To demonstrate the value of the new dementia test, investigators compare mortality rates between individuals with dementia detected by the screening test to mortality of individuals with dementia detected by traditional clinical methods. The investigators find that mortality rates are considerably lower among adults whose dementia is detected by the new screening test, prompting them to conclude that early detection of dementia by screening prolongs survival.

However, in this example, depicted graphically in Fig. 12.5, it is possible that the only thing that happened is that the new screening test identified dementia earlier in its natural course, adding extra follow-up time to the screened group, and therefore lowering their mortality *rate*, which is defined as the number of deaths divided by

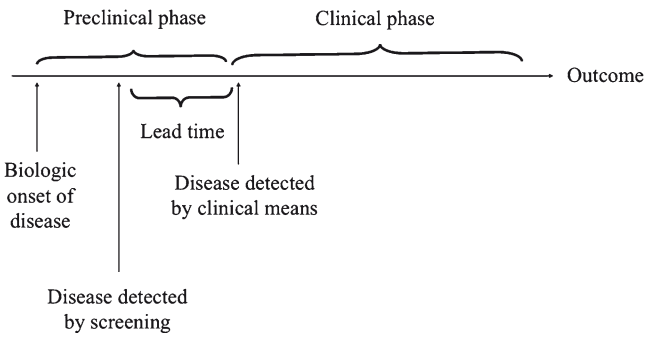


Fig. 12.5 Lead time bias in studies of screening

person-years at risk. Although the screening test had no impact on the course of the disease, the test appeared to be protective due to the preferential addition of risk time to the screened group. Lead-time bias can be addressed by a randomized trial that assigns comparable individuals to screening versus no screening, thereby starting the clock at the same time for screened and unscreened individuals.

12.6.3 Length Bias Sampling

A third issue with screening studies is the tendency of any screening test to *preferentially identify individuals who have a longer preclinical phase of disease*. Consider the above dementia screening test administered to five different individuals who will eventually develop dementia. There is variation in the natural history of dementia, such that some people develop aggressive, rapidly progressive disease with a short preclinical phase, whereas others develop a more slowly progressive disease. These differences may arise due to pathophysiologic variation in the causes of dementia, genetic differences, or interaction of the disease with characteristics of the individual, such as education level.

If the dementia test is given to these five individuals at a single time point, three of five cases of preclinical dementia will be detected preferentially among the people who have the slowest developing disease (Fig. 12.6). The two individuals who have the most rapidly progressive dementia have already developed clinical dementia at

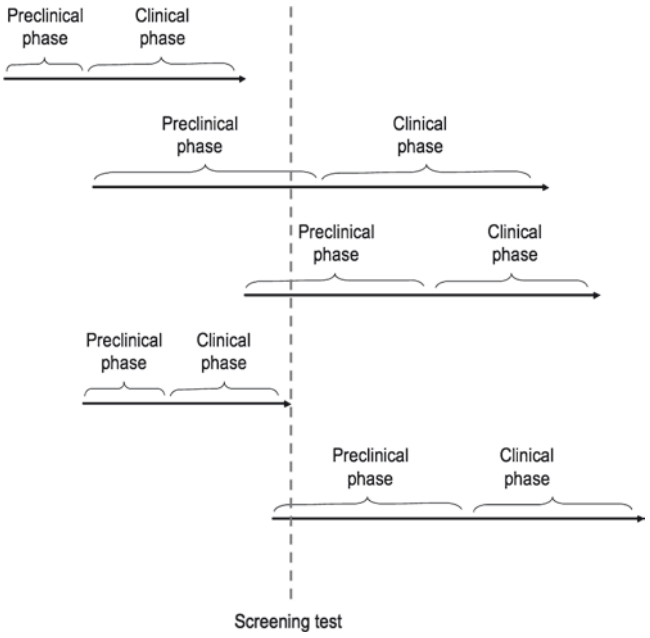


Fig. 12.6 Length-bias sampling in studies of screening

the time of testing, and therefore would not be eligible for screening. Comparing individuals whose dementia is detected by screening to those whose dementia is detected by usual clinical means may result in better outcomes among the screened group, because screened individuals will generally tend to have more mild disease.

12.6.4 Overdiagnosis Bias

Finally, it is possible for overdiagnosis of disease in a screening program to result in the appearance of a benefit from screening, when in fact no such benefit exists. This problem tends to occur in situations where blinding is not possible, and where subjective interpretation is required for interpreting the screening test. Consider the example of a screening program to detect colon cancer using magnetic resonance imaging (MRI). If this program is conducted at an institution that is enthusiastic about the superior image qualities of the newest MRI machines, then the readers may tend to overcall radiographic lesions as cancer. The result will be a "disease detected by screening" population that includes some people who do not actually have colon cancer. Comparing individuals who are diagnosed with colon cancer using the new screening test to a second group of individuals who are diagnosed by traditional, and more specific, means may yield a spurious finding of improved survival among people whose disease is detected by the new screening method.

12.7 Association versus Prediction

In previous chapters, we evaluated *associations* of risk factors with disease to decide whether a particular risk factor might be a *cause* of the disease. It is important to emphasize that many risk factors that are strongly associated with a particular disease often function poorly as screening tests, particularly when the disease is caused by multiple factors.[43]

Consider the association of C-reactive protein (CRP), an inflammatory biomarker, with incident coronary heart disease. In large cohort studies, higher CRP levels are "strongly" associated with the development of coronary heart disease with relative risks of 2.0–3.0 comparing individuals with higher versus lower CRP levels. These "strong" associations combined with temporality, dose-dependence, and biologic evidence documenting CRP within atherosclerotic plaques suggest CRP as a potential *cause* of coronary disease. The *association* of CRP levels with coronary heart disease in a typical cohort study is shown in Table 12.10.

Consider the identical data used to evaluate CRP levels as a screening test to *predict* coronary heart disease (Table 12.11).

These characteristics indicate that CRP levels would perform poorly as a screening test for coronary heart disease. Consider the results of CRP testing in the above population, in which the overall incidence of coronary heart disease is 10%. A person who tests positive for a high CRP level in this population would have a 17% chance of

Table 12.10 Association of CRP level with coronary heart disease

	Coronary heart disease		
CRP level	Yes	No	
High	60	300	360
Low	40	600	640
	100	900	1,000

Relative risk = (60/360)/(40/640) = 2.67

Table 12.11 CRP test characteristics for predicting coronary heart disease

	Coronary heart disease		
	Yes	No	
Test positive	60	300	360
Test negative	40	600	640
Total	100	900	1,000

Sensitivity = 60/100 = 60%
Specificity = 600/900 = 66%
Positive predictive value = 60/360 = 17%
Negative predictive value = 600/640 = 94%

developing coronary heart disease, based on the positive predictive value of the test. The change from a 10% background probability of coronary disease to a 17% probability of disease after CRP testing is unlikely to alter clinical decision-making, such as referring a patient for invasive diagnostic testing. A negative CRP test would indicate a coronary heart disease risk of 6%, which is not materially different than the 10% incidence that was already known prior to testing.

In general, the association of a factor with disease must be *extraordinarily strong* (i.e., at least tenfold) for that factor to perform well as a screening or diagnostic test in an individual patient. For example, when judged using the analytic approach of an association study, a positive mammogram result is *associated* with an approximate 25-fold greater risk of breast cancer, yet the mammogram still leaves much to be desired as a high-quality screening test.

Chapter 13
Diagnostic Testing

Learning Objectives

1. Clinical testing may be used for screening, diagnosis, and prognosis.
2. The following elements should be considered when ordering a diagnostic test:
   (a) Validity and repeatability characteristics of the test
   (b) Safety of the test
   (c) Pre-test probability of the disease
   (d) The natural history or seriousness of the disease without treatment
   (f) The benefits and risks of treatment
   (g) Patient preferences and costs
3. Diagnostic testing is most useful when the pre-test probability of disease is intermediate.
4. Likelihood ratios combine sensitivity and specificity characteristics for a specific test.

13.1 General Considerations in Medical Testing

In the previous chapter, we examined important characteristics of screening tests. In this chapter, we apply these concepts to the diagnosis and prognosis of disease.

We begin by considering a range of testing that is available for breast cancer. Mammography and self-examination represent initial *screening tests* that are used to detect breast cancer in asymptomatic women. A positive result from one of these screening tests typically prompts *diagnostic testing*, such as breast ultrasound, magnetic resonance imaging, or a biopsy, to look for the presence of the disease. If the diagnostic studies indicate breast cancer, then *prognostic tests* such as computed tomography and bone scanning might be used to stage the disease and to offer information regarding possible therapies and overall prognosis. This example illustrates that the clinical context for medical testing is constantly changing within a given individual, such that test results must be interpreted in the appropriate context.

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_13, © Springer Science+Business Media, LLC 2009

Before ordering any medical test, a number of general elements should be considered. The decision to order a screening, diagnostic, or prognostic test should be linked with a clear action plan for dealing with either a positive or negative result. In many cases, testing can increase clinical confusion, patient anxiety, cost, and lead to unnecessary complications from invasive procedures that are prompted by a positive test result.

General elements to consider when ordering a diagnostic test are

1. Validity and repeatability characteristics of the test
2. The safety of the test
3. The pre-test probability of the disease
4. The natural history or seriousness of the disease without treatment
5. The benefits and risks of treatment
6. Patient preferences and costs

These elements will be considered in three clinical examples.

Example 13.1 A 50 year-old woman presents to her outpatient primary care physician with a 3-day history of a painful swollen right leg. There is no previous history of cancer or trauma. On physical examination, the right leg is considerably swollen from the knee down, with mild redness, but there is no calf tenderness. The left leg appears normal.

The primary diagnostic concern here is deep venous thrombosis (DVT), which is a serious disease that can lead to pulmonary embolism and death if left untreated. Treatment for DVT is anticoagulation, which is highly effective, but increases the risk of major bleeding. Ultrasound is a common, noninvasive method for diagnosing DVT and has reasonably good repeatability and validity characteristics; test sensitivity and specificity are 95% and 98%, respectively. Finally, the patient is concerned about her symptoms and would like to know what is causing them.

Based on these considerations, an ultrasound test is obtained and comes back positive for DVT. Given the positive test result, what is the probability that this patient has a DVT?

This question refers to the positive predictive value of the ultrasound test. As discussed in Chap. 12, the predictive values for a test depend on the prevalence of disease; they cannot be determined from sensitivity and specificity data alone. For screening tests, we relied on the disease prevalence in the screened population to determine predictive values, because screening tests are usually administered to asymptomatic individuals. For diagnostic tests, individuals already have signs and symptoms of a disease. We substitute the *pre-test probability of disease* for prevalence in order to calculate the predictive values of diagnostic tests.

The pre-test probability represents a *clinical estimate*, based on an individual patient's signs and symptoms, the clinical judgment and experience of the physician, and published data – in other words, the art and science of medicine. Based on this patient's clinical presentation, the pre-test probability of DVT prior to any diagnostic testing is about 50%.

The intermediate (50%) probability of disease does not justify empirical use of anticoagulation, because the diagnosis of DVT is not certain and because treatment may have serious side effects. The intermediate probability also does not justify observation, because DVT is a serious disease that may cause pulmonary embolism or death if not treated. In general, *an intermediate pre-test probability of disease represents the most useful setting for ordering a diagnostic test.*

Given an estimated pre-test probability of 50%, we can construct a 2×2 table to calculate the positive predictive value using a hypothetical sample of 1,000 patients and known sensitivity and specificity values for the ultrasound test (Table 13.1).

Table 13.1 Predictive value of ultrasound testing for deep venous thombosis

	DVT		
Ultrasound	Yes	No	
Positive	500×0.95 = 475	10	485
Negative	25	500×0.98 = 490	515
Total	500	500	1,000 total

Positive predictive value = 475/485 = 98%
Negative predictive value = 490/515 = 95%

Given a positive test result, there is a 98% probability that this patient has a DVT. A positive ultrasound test would strongly rule in the diagnosis of DVT and motivate initiation of anticoagulation therapy. Additional "gold-standard" diagnostic tests, such as a venogram, could only marginally increase the likelihood of disease at this point, and are associated with additional risks.

If this patient has a negative ultrasound test, what is the probability of DVT?

From the Table 13.1, the negative predictive value = 490/515 = 95%. Given a negative ultrasound test result, there is a 5% probability that this patient has a DVT. This probability is somewhat low, and considerably different from the pre-test probability of 50%; however, DVT remains a concern given the seriousness of the disease. One potential strategy is to repeat the ultrasound test one week later. If the second test is independent of the first, what is the chance of DVT if this patient again tests negative by ultrasound?

Based on a negative first ultrasound test, this patient has a 5% pre-test probability of DVT prior to the second ultrasound test. Given fixed sensitivity and specificity characteristics of the ultrasound test and a 5% pre-test probability of disease, table 13.2 indicates the predictive values of repeat ultrasound testing for DVT.

Based on the data in Table 13.2, after the second negative ultrasound test we can be reasonably sure that this patient does not have a DVT. An appropriate clinical plan would be reassurance and to monitor her signs and symptoms closely.

The assumption that repeated test results are independent may not hold in practical application. The first and second ultrasound tests *are* likely to be related even if a different ultrasound technician and radiologist perform the test. The calculation of post-test probabilities for multiple non-independent tests is beyond the scope of this book.

Table 13.2 Repeat ultrasound testing for deep venous thrombosis

	DVT		
Ultrasound	Yes	No	
Positive	50×0.95 = 48	19	67
Negative	2	950×0.98 = 931	933
Total	50	950	1,000 total

Negative predictive value = 931/933 = 99.8%.

Example 13.2 A 20-year-old man presents to his outpatient physician with a 1-week history of a scaly, red, itchy rash on his right forearm. He denies the use of new soaps, skin products, or environmental exposures. He reports taking long hot showers and has a history of eczema.

The primary diagnostic concern here is eczema, which is *not* a serious disease and is highly responsive to topical corticosteroid therapy. Topical corticosteroids have minimal side effects when used for a short duration. A skin biopsy can confirm the diagnosis of eczema, but is invasive, somewhat expensive, and the patient is not likely to want it.

This patient's clinical presentation is classic for eczema and does not immediately suggest an alternative diagnosis. Based on subjective assessment of his signs and symptoms, the estimated pre-test probability of eczema is quite high, perhaps 90%.

A high pre-test probability of disease is *not* an ideal starting point for initiating further diagnostic testing, because we are already fairly confident that the disease is present. An alternative plan to diagnostic skin biopsy in this case is empirical treatment with topical corticosteroids and follow-up skin examination in 2 weeks. Given the high pre-test probability of eczema and known effectiveness of treatment, this plan will likely treat the disease successfully. If the patient were to return without improvement from topical corticosteroids, we would be concerned that our initial diagnostic suspicion of eczema was incorrect. Nonresponse to topical corticosteroid treatment would yield a new clinical environment that may be more conducive to testing, specifically a more intermediate pre-test probability and a more concerned patient and physician.

Example 13.3 An 8-year-old boy presents to his pediatrician with a 3-day history of a sore throat, runny nose, and fever. He has some tender anterior cervical lymph nodes and marked erythema of the oropharynx, but no exudates are present. His father is highly concerned about the possibility of strep throat.

The primary diagnostic concern in this case is streptococcal pharyngitis, which can be a serious disease if left untreated. Potential complications, albeit uncommon, include retropharyngeal abscess, rheumatic fever, and glomerulonephritis. The rapid strep test has good validity characteristics for diagnosing streptococcal pharyngitis; sensitivity and specificity are 90% and 95%, respectively.

Antibiotic treatment is curative and has few side effects; however, antibiotic overuse has potentially serious consequences for the community. Excessive antibiotic use has produced more virulent bacterial strains that increase the risks of major complications

and death among affected individuals. The treating physician must balance the potential benefits of antibiotic therapy to the individual patient with the possibility of significant harm to the community.

The pre-test probability of strep pharyngitis is estimated from aspects of this child's clinical presentation. Streptococcal infection is often accompanied by a sore throat, fever, and cervical lymphadenopathy, which are present in this case. However, strep pharyngitis is a relatively unusual cause of upper respiratory infections compared to common viruses, is not typically associated with a runny nose, and is usually characterized by exudative lesions, which are absent in this case. Based on these considerations, the estimated pre-test probability of strep pharyngitis is somewhat low in this child, probably around 10%.

The decision to proceed with diagnostic testing in this case is less clear compared to the first two examples. While the pre-test probability of disease is somewhat low, the disease can be serious if missed, and the diagnostic test is safe and easy to administer. Does the rapid strep test help guide the decision to initiate antibiotic therapy in this case? Consider the data in Table 13.3:

Table 13.3 Positive and negative predictive values of the rapid strep test

	Streptococcal pharyngitis		
Rapid strep test	Yes	No	
Positive	100×0.90 = 90	45	135
Negative	10	900×0.95 = 855	865
	100	900	1,000 total

If the rapid strep test is positive, there is a 67% probability that this child has streptococcal pharyngitis (positive predictive value = 90/135 = 67%). Given the potential seriousness of the disease and availability of a relatively safe and effective treatment, a 67% probability of disease would warrant treatment of this child.

If the rapid strep test is negative, there is a 1% probability that this child has streptococcal pharyngitis (negative predictive value = 855/135 = 99%). Given a negative test, it is reasonable not to treat with antibiotics and have the child return for follow-up. It is reasonable to order the rapid strep test in this case because it can influence clinical management.

13.2 Likelihood Ratios

A complementary approach to solving diagnostic testing problems is to use *likelihood ratios*. Likelihood ratios combine elements of test sensitivity and test specificity in a manner that allows them to be easily combined with the pre-test probability, or disease prevalence, to obtain the post-test probability of disease. Positive and negative likelihood ratios are defined below.

$$\text{Likelihood ratio positive} = \text{sensitivity}/(1 - \text{specificity})$$

$$\text{Likelihood ratio negative} = (1 - \text{specificity})\ /\text{specificity}$$

Notice that the likelihood ratios include only the sensitivity and specificity characteristics of a test. An estimate of the pre-test probability of disease, or the prevalence of disease, is needed to use likelihood ratios to calculate the post-test probability of disease.

Likelihood ratios are typically plotted using a nomogram that permits quick and easy conversion of pre-test probabilities into post-test probabilities for any given test (Fig. 13.1).

To use the likelihood ratio nomogram, start with the estimated pre-test probability of disease in the left-hand column. If the diagnostic test is positive, use a straight line or ruler to connect the pre-test probability of disease with the positive likelihood ratio for the test, which corresponds to values greater than 1.0 in the middle column. The resultant point on the post-test probability column represents the probability of disease given a positive test result.

If the diagnostic test is negative, connect the pre-test probability of disease with the negative likelihood ratio for the test, which corresponds to values less than 1.0 in the middle column. The resultant point on the post-test probability column represents the probability of disease given a negative test result.

For example, the rapid strep test has a likelihood ratio positive of 18 and a likelihood ratio negative of 0.1. Likelihood ratios can be found in the medical literature, or can be calculated directly from known sensitivity and specificity data using the equations provided above. The nomogram in Fig. 13.2 depicts the results of a positive rapid strep test for the 8-year-old boy with a 10% pre-test probability of strep pharyngitis.

The nomogram in Fig. 13.3 demonstrates the results of a negative rapid strep test in the same individual.

Notice that these results are identical to those obtained using the more cumbersome 2×2 table approach; likelihood ratio nomograms simply provide a more efficient method to combine pre-test probability with sensitivity and specificity characteristics of a particular test to obtain the post-test probability.

Analogous to the 2×2 table approach, likelihood ratio nomograms can be used to analyze consecutive test results within the same individual, assuming that the tests are independent. In Example 13.1, a 50-year-old woman presented with signs and symptoms suggesting a 50% pre-test probability of DVT. Given a negative likelihood ratio of 0.04 for ultrasound diagnosis of DVT, the nomogram in Fig. 13.4 depicts the post-test probability of disease following a first negative ultrasound test.

If the first ultrasound test is negative, the post-test probability is about 5%. Given the serious potential complications of DVT, a second ultrasound test is ordered one week later. Provided that the second test is independent of the first, the nomogram in Fig. 13.5 describes the post-test probability of DVT following a second negative ultrasound test.

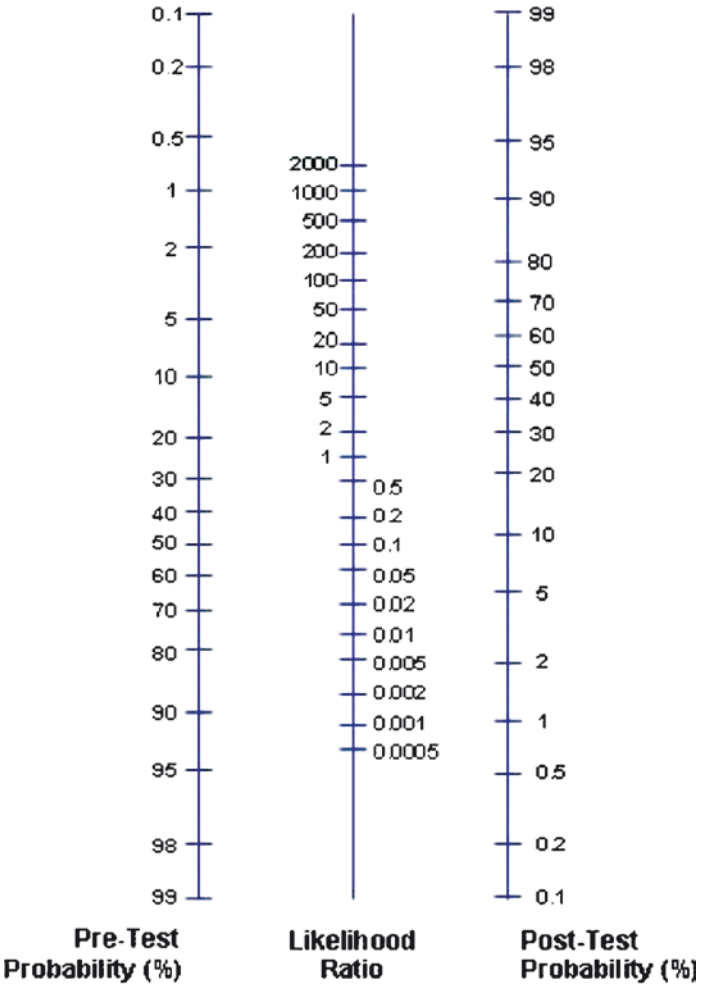


Fig. 13.1 Likelihood ratio nomogram

The likelihood ratio nomogram can be used to determine post-test probabilities for a number of common conditions given results of diagnostic testing. A partial list appears in Table 13.4.

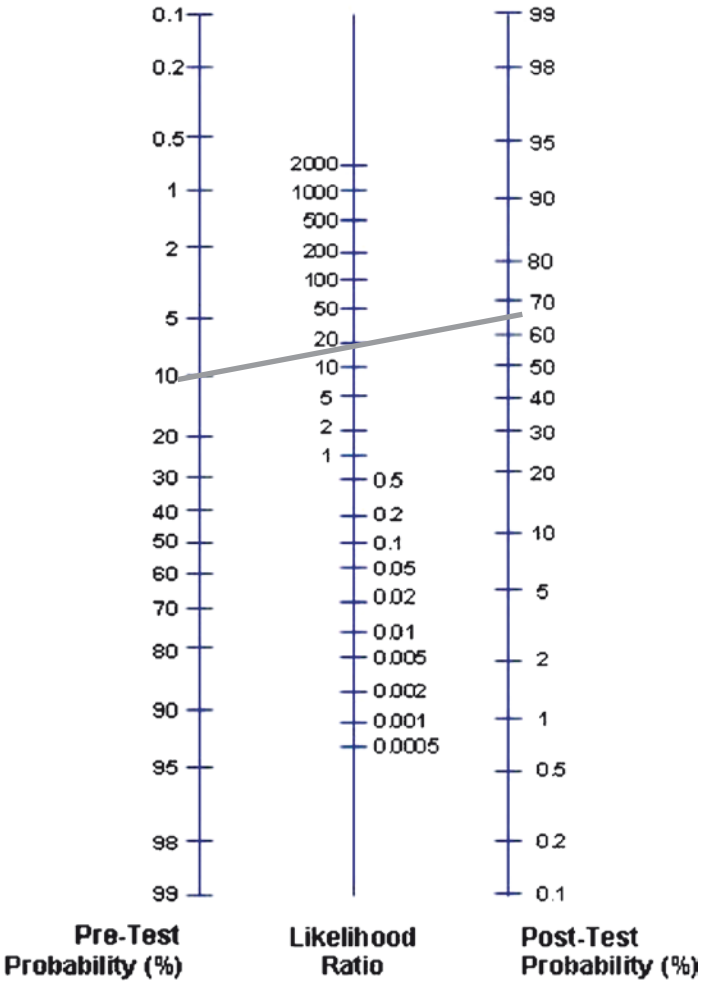


Fig. 13.2 Likelihood ratio nomogram for positive rapid strep test. Given a positive rapid strep test, the post-test probability of disease is about 67%

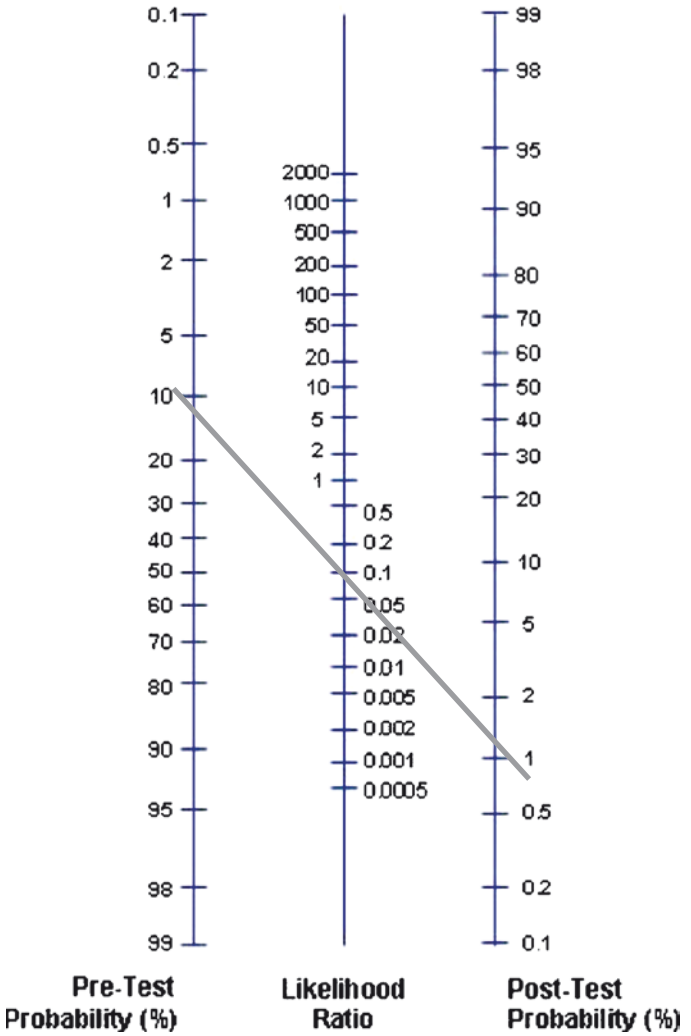


Fig. 13.3 Likelihood ratio nomogram for positive rapid strep test. Given a negative test result, the post-test probability of disease is about 1%

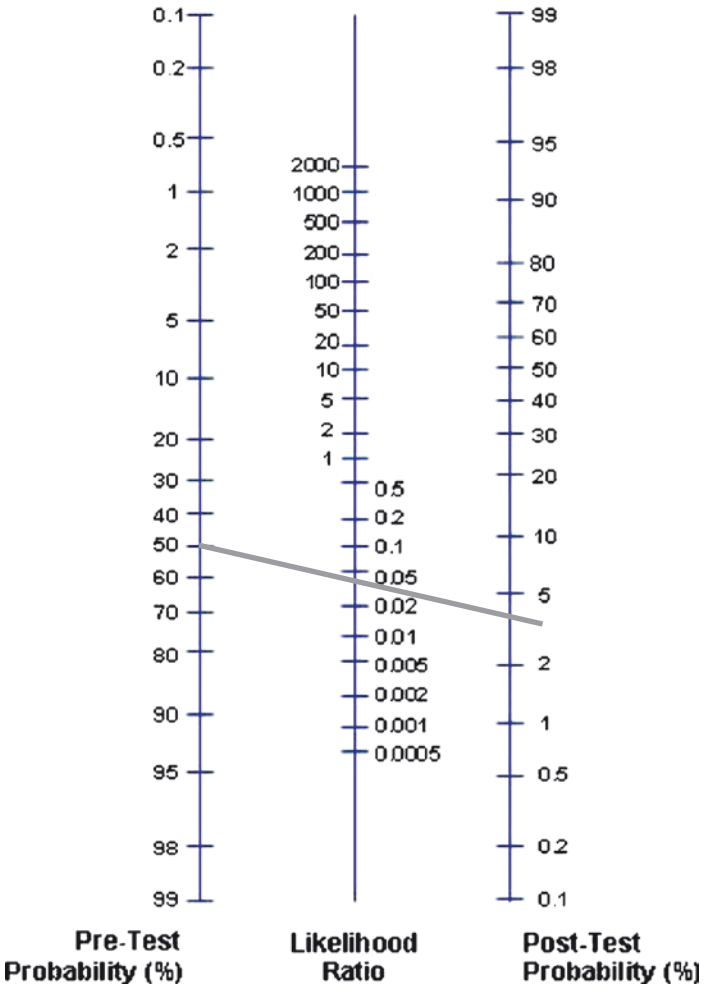


Fig. 13.4 Likelihood ratio nomogram for the first negative ultrasound test

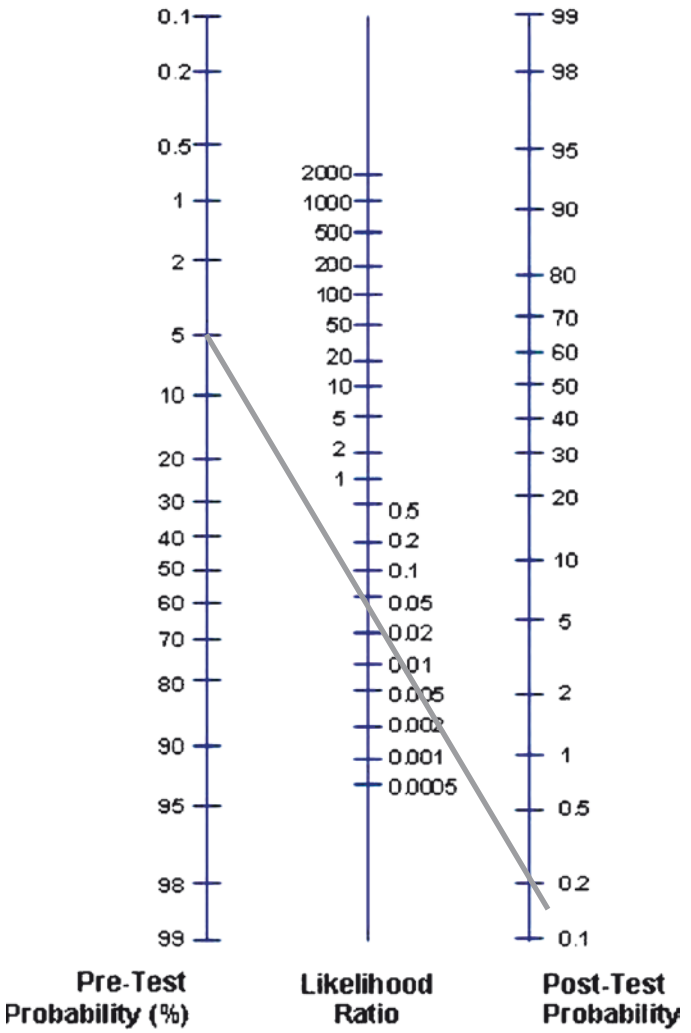


Fig. 13.5 Likelihood ratio nomogram for the second negative ultrasound test. Given a second negative ultrasound test the post-test probability of DVT is only about 0.1%

Table 13.4 Likelihood ratios for common clinical conditions

Condition	Diagnostic test	LR+	LR−
Abdominal abscess	Abdominal ultrasound	19.2	0.04
Acute cerebral hemorrhage	Head computed tomography	23.8	0.05
Acute cholecystitis	Abdominal ultrasound	23.8	0.05
Acute myocardial infarction	Cardiac enzymes	32.3	0.03
Aortic stenosis	Echocardiogram	2.6	0.14
Brain tumor	Head computed tomography	31.7	0.05
Breast cancer	Mammogram	8.7	0.14
Breast cancer	Clinical breast exam	3.8	0.69
Clostridium difficile colitis	Clostridium difficile toxin assay	19.6	0.02
Coronary artery disease	Exercise ECG – 1 mm depression	3.5	0.45
Coronary artery disease – women	Exercise echocardiogram	4.29	0.18
Coronary artery disease – women	Exercise thallium test	2.87	0.36
Coronary artery disease – women	Exercise treadmill test	2.25	0.55
Carotid atherosclerosis	Duplex ultrasound	9	0.11
Common duct stone	Abdominal ultrasound	3.8	0.76
COPD (in middle-aged patients)	Forced expiratory volume <80%	2.7	0.81
Endocarditis	Erythrocyte sedimentation rate >20	23	0.07
Endocarditis	Echocardiogram	9.3	0.66
Iron deficiency anemia	Serum transferrin <16.6% saturation	3.2	0.06
Hematuria	Urine dipstick	6.4	0.06
Lung cancer	Chest X-ray	15	0.42
Left ventricular hypertrophy	Echocardiogram	18.4	0.08
Myocardial infarction	Electrocardiogram, single	28.5	0.44
Myocardial infarction	Electrocardiogram, serial	68	0.32
Osteomyelitis	Plain film- bone	5.6	0.55
Acute pancreatitis	Serum amylase	47.5	0.05
Acute pancreatitis	Serum lipase	87	0.13
Renal artery stenosis	Renal scan	4.1	0.28
Systemic lupus	Antinuclear antibodies	4.5	0.125
Systemic lupus	Anti-DNA antibodies >1:80	73	0.27
Chronic subdural hematoma	Head computed tomography	15.5	0.07
Temporal arteritis	Erythrocyte sedimentation rate >20	24.8	0.01
Ureteral obstruction	Abdominal ultrasound	9.8	0.02

LR+ indicates likelihood ratio positive; LR− indicates likelihood ratio negative.
Data obtained from http://www.med.unc.edu/medicine/edursrc/lrdis.htm.

Chapter 14
Summary Measures in Statisitics

Abigail Shoben

Learning Objectives

1. Three types of variables are commonly used in clinical research studies:
   (a) Continuous – can take on an infinite number of possible values
   (b) Binary – can take on only two possible values
   (c) Categorical – can take on only a few possible values
2. A histogram plots the observed values of a variable on the *X*-axis versus the relative frequency of these values on the *Y*-axis.
3. The arithmetic mean describes the middle of the data for symmetric distributions, including data from normal-appearing distributions.
4. The median refers to the value within a distribution for which exactly half of the data fall above this value and half fall below it.
5. Disagreement between mean and median values suggests that a distribution is *not* normally distributed.
6. A continuous variable may be divided into 3, 4, or 5 equally sized groups called tertiles, quartiles, and quintiles, respectively.
7. The interquartile range is defined as the 25th and 75th quantiles of a distribution.
8. Some techniques to describe the joint distribution between two variables include tabulation across categories, scatter plots, correlation, and quantile-continuous plots.
9. Correlation coefficients are interpreted as:
   (a) +1 indicates perfect positive agreement between two variables
   (b) 0 indicates no agreement between two variables
   (c) –1 indicates perfect negative agreement between two variables

14.1 Types of Variables

Most variables used in clinical research studies can be described as belonging to one of three categories: continuous, categorical, or binary.

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_14, © Springer Science+Business Media, LLC 2009

Continuous variables can take on an infinite number of possible values theoretically arising from anywhere along a (perhaps truncated) number line. Examples of continuous variables include body temperature, the serum sodium concentration, and left ventricular mass. Many continuous variables in clinical research can assume only positive values, for example weight and blood pressure, because negative values for these variables are scientifically impossible.

Binary variables can take on only two possible values, for example sex. Binary variables are often used as *indicator variables*, which take the value of 1 if a specific characteristic or disease is present, and 0 if it is not. For example, aspirin use in a clinical research study can be represented by an indicator variable that will be equal to 1 if a subject is using aspirin and 0 if they are not. Binary variables are also known as *dichotomous* variables.

Categorical variables can take on only a few possible values, for example race, or stage of chronic kidney disease. Some categorical variables can be further classified as *ordered* when the possible responses correspond to a scientific scale that has a clear hierarchy. For example, the severity of chronic kidney disease is graded on a scale from 1 (mild) to 5 (most severe). Other categorical variables are classified as *nominal* if there is no ordered distinction among the possible responses. Examples include marital status (never married, married, widowed, divorced), and race (for example, Caucasian, African American, Asian, Other).

Continuous variables can be transformed into categorical or binary variables in clinical research studies. For example, systolic blood pressure, a naturally continuous variable, can be transformed into an ordered categorical variable with responses grouped by clinically accepted categories of "normal" (<120 mmHg), "pre-hypertension" (120–140 mmHg), and "hypertension" (>140 mmHg).

14.2 Univariate Statistics

14.2.1 Histograms

Clinical research studies typically contain too many subjects to practically list all of the observed values for any particular variable. For example, it would be impractical for most blood pressure studies to list the individual blood pressures for every person in the study. Instead, some compact description of blood pressures is desired. A *summary measure* is a compact description of the data that conveys information about the distribution of a variable quickly. Summary measures that pertain to a single variable are called *univariate statistics*.

One summary measure for a single variable is called a *histogram*, which plots observed values of a variable on the *X*-axis versus the relative frequency of these values on the *Y*-axis. For example, the systolic blood pressures of more

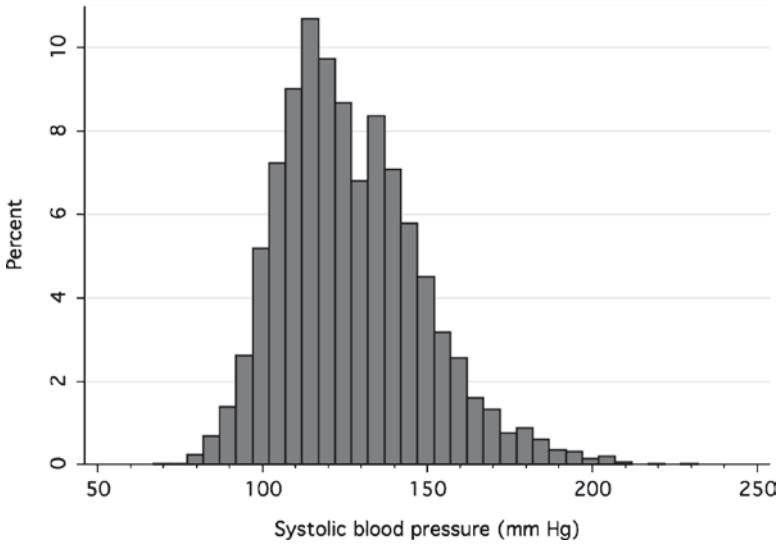


Fig. 14.1 Histogram of systolic blood pressure values. Each bar on the *X*-axis is 5 mmHg wide

than 6,000 research participants are described by the histogram in Fig. 14.1, which conveys a sense of how systolic blood pressures are distributed in the entire data set.

The *Y*-value indicates the frequency (%) that each systolic blood pressure value is found in the distribution. For example, about 4% of systolic blood pressures in this study are between 150 and 155 mmHg. The histogram demonstrates that "typical" values for systolic pressure in this study are usually between 100 and 150 mmHg with some extreme values as high as 200 mmHg.

A second histogram (Fig. 14.2) describes serum levels of C-reactive protein (CRP), an inflammatory biomarker, in this same 6,000-person cohort.

The CRP histogram has a very different shape than that for systolic blood pressure. In general, most study participants have very low CRP levels, and a select few have progressively higher levels. More than 30% of participants appear to have a CRP level that is close to 0 mg/l.

Based on the appearance of the histogram, systolic blood pressures would be described as having a "normal" distribution. A normal distribution means that the data appear to be shaped roughly like a bell-shaped curve. Close inspection of the blood pressure histogram reveals a nonsymmetric bump on the right side that distracts from an otherwise clean, bell-shaped appearance. A more descriptive categorization of the systolic blood pressure data would be "normal appearing with some rightward skew." In contrast the CRP histogram reveals a variable that is clearly *not* normally distributed, and is highly skewed.

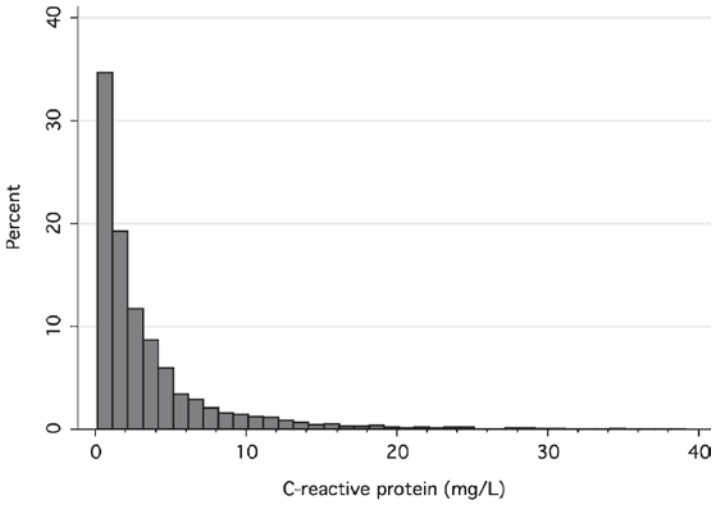


Fig. 14.2 Histogram of C-reactive protein values

14.2.2 Measures of Location and Spread

In addition to describing the general shape of a distribution, the histogram also provides graphical information regarding "typical" values for a particular variable, and the range of possible values for that variable. Formally, estimates of the "typical" value in a distribution are termed *measures of location*, and estimates of how far apart the data are from this typical value are termed *measures of spread*.

The most common measure of location is the arithmetic *mean*. Statistically, the mean is the expected value of a variable, and in many circumstances it will be a reasonable description of the middle of the data. The mean is calculated by summing all of the observed values of a variable and then dividing by the total number of observations.

$$\text{Mean} = \frac{\sum x_i}{N}$$

The most common measure of spread is the *standard deviation*. Distributions with a high standard deviation will be very spread out, whereas distributions with a low standard deviation will be tightly grouped. The standard deviation is calculated as the square root of the variance.

$$\text{Variance} = \frac{\sum (\mu - x_i)^2}{N}$$

$$\text{Standard deviation} = \sqrt{(\text{variance})}$$

Calculation of the variance requires going through each value in a distribution, calculating the squared distance between that value and the mean, then dividing this sum by the total number of observations. Thankfully, this is accomplished by a keystroke on a computer.

The mean value of a study variable is usually presented along with the standard deviation. For example the mean systolic blood pressure among the 6,000 research participants is 126.6 mmHg, and the standard deviation is 21.5. If a variable is normally distributed, then 67% of the data will fall within one standard deviation of the mean value, and 95% of the data will fall within two standard deviations of the mean. Since the histogram for systolic blood pressure reveals a fairly normal appearing distribution, we can expect approximately 95% of the systolic blood pressures in this study to fall within two standard deviations of 126.6 mmHg, or 126.6 ± (21.5 × 2), or (83.6, 169.6) mmHg, as shown in Fig. 14.3.

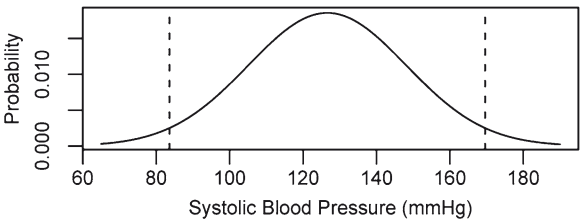


Fig. 14.3 Normal distribution; 95% of data within two standard deviations of the mean

The arithmetic mean is fairly sensitive to extreme values of a distribution. For example, the mean income for a particular town may not reflect the "typical" income for that town if the majority of residents earn between $20,000 and $120,000, but there is also one billionaire. The mean CRP value in the 6,000-person research study was 3.8 mg/l. However, inspection of the histogram in Fig. 14.2 reveals that 3.8 mg/l is not really a typical CRP value for this population, but is highly influenced by the few individuals with very high CRP values. Better options for describing typical values for such highly skewed distributions are the *median* and the *geometric mean*, which are less sensitive to extreme values.

The geometric mean is calculated by taking the arithmetic mean of log-transformed data, and then converting back to the original scale by exponentiation (taking the antilog). In symbols:

$$\text{Geometric mean} = \exp\left(\frac{\sum \text{In}(x_i)}{N}\right)$$

By taking the mean of the log-transformed data, the geometric mean is less influenced by values far away from most of the others, unlike the arithmetic mean. The geometric mean of the CRP data is 1.92 mg/l. In terms of sensitivity to outlying values, the geometric mean is between the arithmetic mean, which is very sensitive to extreme values, and the median, which is completely unchanged by outlying values.

14.2.3 Quantiles

Another common technique to describe the distribution of a variable is *quantiles*, also called *percentiles*. Quantiles describe specific values within a distribution that divide the data into groups. For example, 90% of participants in the 6,000-person research study have a CRP level less than 9.0 mg/l. Therefore, the 90th percentile for the CRP distribution is 9.0 mg/l. The 50th quantile, also called the *median*, refers to the value within a distribution for which exactly half of the data fall above this value and half fall below it. The median CRP level is 1.9 mg/l, meaning that half of the participants have a CRP level less than 1.9 mg/l and half have a level greater than 1.9 mg/l.

In this example, the mean and median CRP values are dissimilar. The CRP histogram suggests that the median value of 1.9 mg/l is a better estimate of a "typical" CRP level for this particular population compared to the mean value of 3.8 mg/l. Disagreement between mean and median values within a distribution suggests that the data are *not* normally distributed.

Continuous variables are commonly converted into categorical variables using quantiles. For example, continuous CRP levels in the study could be divided into three, four, or five equally sized groups, known as tertiles, quartiles, and quintiles, respectively. Categorical variables have the advantage of being easier to describe in tables, as seen in Table 14.1.

Table 14.1 Quartiles of serum CRP levels in a population-based study

	CRP level (mg/l)			
	Quartile 1	Quartile 2	Quartile 3	Quartile 4
	0.1–0.7 mg/l (N = 1,500)	0.8–1.8 mg/l (N = 1,500)	1.9–4.2 mg/l (N = 1,500)	4.2–27 mg/l (N = 1,500)
Age	61 (11)	63 (11)	63 (10)	61 (10)
Systolic blood pressure (mmHg)	123 (22)	126 (21)	129 (22)	129 (21)
Waist circumference (cm)	91 (12)	96 (12)	101 (14)	105 (15)

All values are mean (standard deviation)

Quantiles can be displayed graphically using *box plots*. Box plots contain three main pieces of information, (1) a shaded region representing the 25th and 75th quantiles, which is also called the *interquartile range*, (2) a horizontal bar representing the median, and (3) "whiskers" that typically extend to 1.5 times the 25th and 75th quantiles or to the minimum and maximum observation, whichever is less extreme. Additional extreme observations are displayed as open circles beyond the whiskers. Box plots for two distributions, A and B, are shown in Fig. 14.4.

These two distributions have the same mean and standard deviation, however distribution B has more outlier observations.

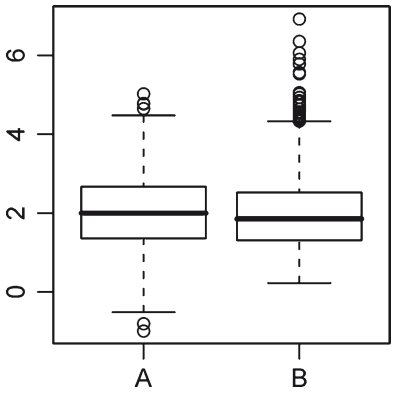


Fig. 14.4 Box plots for two distributions with the same mean and standard deviation

14.2.4 Univariate Statistics for Binary Data

For binary data, the arithmetic mean conveys everything. For example, if aspirin use is represented by a binary variable that takes on a value of 1 if a person uses aspirin and a value of 0 if they do not, then a mean value of 0.25 for the aspirin use variable would indicate that 25% of subjects use aspirin.

The standard deviation of binary data is calculated directly from known mathematical properties. For a binary variable with a mean value of p,

$$\text{Standard deviation} = \sqrt{[p \times (1-p)]}$$

So, the standard deviation of the aspirin use variable would be $\sqrt{\left[0.25 \times \left(1-0.25\right)\right]} = 0.43 \cdot$

With categorical data, the mean is rarely scientifically meaningful, even if it is an ordered categorical variable. The primary interest is in the percentage of subjects in each classification, which is typically described in a table.

14.3 Bivariate Statistics

14.3.1 Tabulation Across Categories

Bivariate descriptive statistics are used to describe the joint relationship between two variables of interest. One important application of bivariate statistics in clinical research studies is to identify potential confounding factors by examining the joint distribution of the exposure variable with other variables in the study. Consider an observational study that examines whether aspirin use lowers the risk of myocardial infarction. Aspirin use might be a marker of other health-related factors that also influence myocardial infarction risk, thereby confounding this association. Bivariate

statistics could be used to describe the joint distribution of the exposure variable (aspirin use) with potential confounding variables in the study. Since aspirin use is a binary variable, bivariate statistics could simply be the tabulation of means and standard deviations for continuous study variables according to values of aspirin use.

For the example of systolic blood pressure:

	Aspirin use	
	Yes	No
Systolic blood pressure mmHg	155.3 (26.1)	152.9 (24.7)

(Mean values with standard deviation in parentheses.)

Tabulation of systolic blood pressure by aspirin use reveals a roughly similar distribution of blood pressure among participants who take or do not take aspirin, eliminating systolic blood pressure as a potential confounder of the association of aspirin use with myocardial infarction in this study.

In clinical research articles, binary variables are usually presented as the mean (number of subjects) rather than mean (standard deviation), because the standard deviation of a binary variable is simply calculated from the mean using $\sqrt{[p \times (1-p)]}$

	Aspirin use	
	Yes	No
Male (%)	40.5 (2650)	36.6 (2210)

(Mean values with number of participants in parentheses.)

A full set of bivariate statistics for aspirin use with each important covariate in the study would be typical for a table of baseline characteristics, which is usually the first table in a clinical research article.

14.3.2 Correlation

Direct comparison of two continuous variables can be performed graphically using *scatter plots*. The three scatter plots in Fig. 14.5 below describe two continuous variables with a weak positive association (plot A), no association (plot B), and a strong negative association (plot C).

Although scatter plots provide subjective assessment of the joint distribution between two variables, they do not provide an objective measure of the strength of the relationship. One objective summary measure that is used to describe the joint distribution of two continuous variables is called *correlation*, which measures the tendency of larger values of one variable to match up with larger measurements of another variable. A correlation coefficient of +1 indicates perfect positive agreement

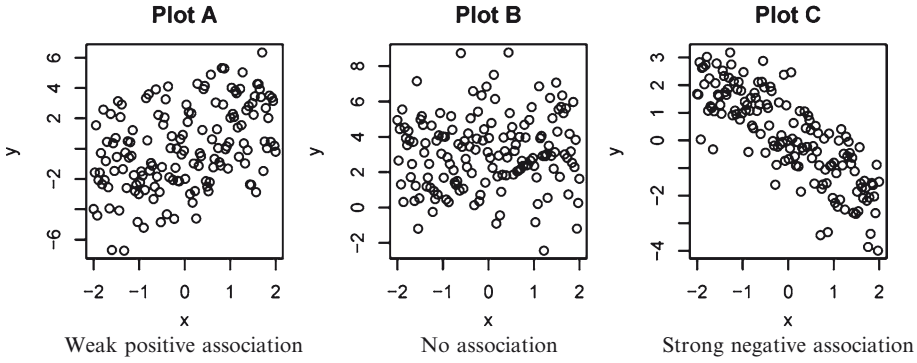


Fig. 14.5 Scatter plots depicting relationships between two continuous variables

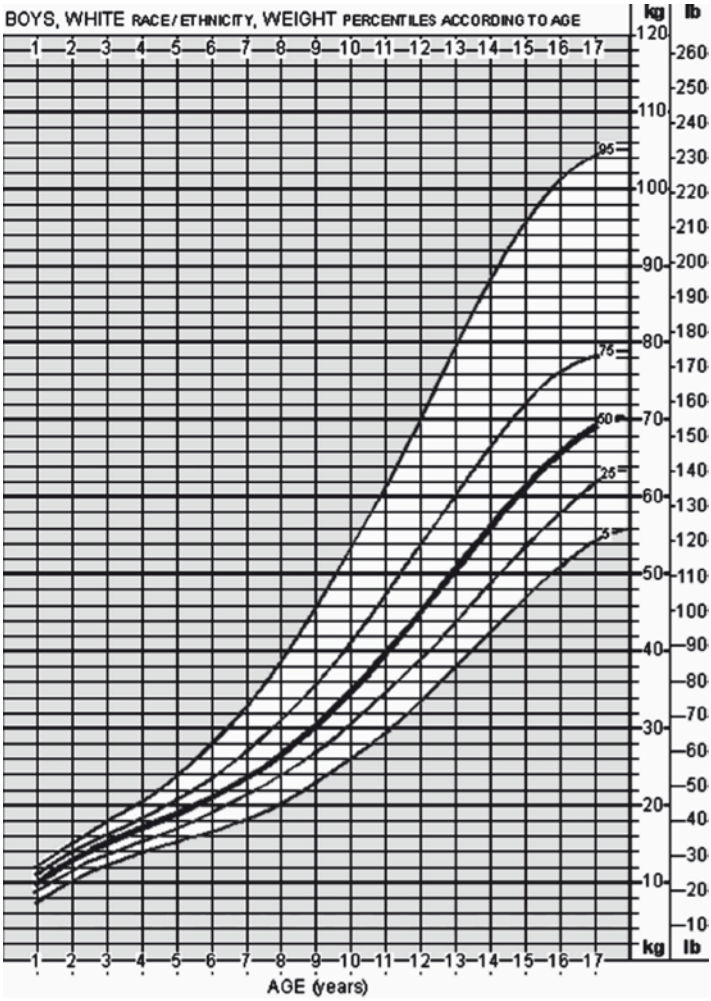


Fig. 14.6 Example of quantile–continuous plot: pediatric growth chart for boys

between two variables, such that higher values of one variable are always linked with higher values of the second variable. A correlation coefficient of +1 indicates perfect positive agreement between two variables, such that higher values of one variable are always linked with higher values of the second variable. A correlation coefficient of −1 indicates perfect negative agreement, such that higher values of one variable are always linked with *lower* values of the second variable, and a correlation coefficient of 0 indicates no agreement. The scatter plots in Fig. 14.5 have correlation coefficients of +0.43, +0.06, and −0.81, respectively.

14.3.3 Quantile–Continuous Variable Plots

Finally, a continuous variable may be plotted as a function of quantiles of a second variable. A classic example is a pediatric growth chart, shown in Fig. 14.6 for boys.

The growth chart displays different quantiles for weight (5th, 25th, 50th, 75th, and 95th quantiles) plotted on the *Y*-axis as a function of age plotted on the *X*-axis. This chart demonstrates that the median weight of a 10-year old boy is about 75 pounds, meaning that half of 10 year-old boys would be expected to have a weight below 75 pounds and half would be expected to have a weight above this value. The interquartile range for weight of 10-year-old boys is 65–95 pounds, describing the middle 50% of the data. A weight below 58 pounds is expected in less than 5% of 10-year-old boys.

Chapter 15
Introduction to Statistical Inference

Learning Objectives

1. A sample is a subset of individuals derived from a given population.
2. Statistical inference relates findings from a sample to those in the population.
3. Generalizability refers to the scientific/practical relevance of the underlying population.
4. Factors that increase the likelihood that sample findings reflect those of the population:
   a. Large sample size
   b. Small population variation
5. 95% confidence intervals:
   (a) If a study is repeated indefinitely and a 95% confidence interval placed around each sample mean, then 95% of the intervals will contain the true population mean.
   (b) Can be calculated without knowing the true population mean.
   (c) Will be narrower when the sample size is large and the sample variation is small.
6. *P*-values:
   (a) Given a null hypothesis *regarding the population*, the *p*-value is the probability of observing the sample result, or a more extreme result, due to sampling variation.
   (b) Findings from a particular sample are used to make an inference about the true results in the population.

15.1 Definition of a Population, Sample, and random Sample

A *population* refers to all people in the world or universe that fit some particular description. A *sample* is some subset of a given population. By definition, a sample is always completely contained within a given population.

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_15, © Springer Science+Business Media, LLC 2009

For example, consider a study that examines whether aspirin use protects against the development of heart disease among middle-aged American men. The *population* in question would be *all* middle-aged men in the United States. The study subjects (unfortunately referred to as the "study population" even though they are not a population) are just one of many possible samples derived from the population of interest.

Consider another study looking at whether increasing the dosage of dialysis can improve clinical outcomes among patients with kidney failure recruited from eight dialysis centers around the world. In this case, the *population* in question would be all people in the world who are receiving chronic dialysis. The study subjects are just a tiny sample of the population in question.

From the definition of a population, we can see that clinical studies rarely, if ever, study entire populations. However research studies would like to *infer* results from a particular study sample to the whole population. If aspirin use is found to be effective at reducing the risk of heart disease in a sample of 100 middle-aged American men, then the authors would like to tell you that you can expect the same effect of aspirin among all middle-aged American men in your practice. If study results applied only to the people in the study, they would be of little use. So, inferential statistics is about relating findings from a sample to those from a larger population.

15.2 Statistical Inference

When samples are drawn at random from a population (called a *random sample*) they may not accurately reflect the qualities of the population simply due to chance. For example, assume that the average score for the Epidemiology final exam is 90% for the whole medical school class. We could draw a random sample of 10 students, average those test scores, and get 82%. We could then repeat this process for another 10 students and get 88%, or 97%. Only by sampling the entire population (the whole medical school class in this example) would we be guaranteed to find the true mean score of 90%.

Mathematical concepts provide a sense of how accurately sample results will reflect true results from the underlying population. Two factors that determine how accurately a sample represents the population are *sample size* and *variance*. Sample size is intuitive – as larger samples are selected, the samples will more closely mirror the characteristics of the population in question.

To understand variance, reconsider the final exam scores again, in which the average score is 90%. If everyone in the class scores between 85% and 95% then the mean value for most samples will probably come out somewhere near 90%. On the other hand, if the range of exam scores is 40–120% (with extra credit), there will be a greater possibility of selecting an unusual sample, such as one with a mean score of 110%, or 65%. So, lower variation in a population increases the likelihood that a particular sample will accurately reflect that population.

15.3 Generalizability

Generalizability describes whether an underlying *population* is clinically or scientifically relevant. Generalizability is a subjective term, and is *not* related to inferential statistics. For example, if investigators conducting a clinical trial decide to exclude participants who are greater than 75 years old, those who have diabetes, and those who plan to leave the study area within the next 6 months, then the role of inferential statistics is to relate the results obtained in the trial to those in the entire population of individuals who are less than 75 years old, non-diabetic, and not planning to move. Common sense and clinical/scientific knowledge, and *not statistics*, are used to decide whether the underlying population is scientifically important and/or applicable to clinical practice.

P-values and *95% confidence intervals* are tools of statistical inference. The function of *p*-values and 95% confidence intervals is to describe how well sample findings relate to an underlying population, but not whether that population makes any clinical, scientific, or practical sense.

15.4 Confidence Intervals

If a particular experiment is conducted an infinite number of times, with each experiment having a fixed sample size, then some interval can be placed around each sample mean such that 95% of all the intervals contain the true population mean.

It's not an easy concept. Here is one example:

Consider that the average weight of American adults is 160 pounds. We will start by selecting a random sample of 100 American adults and then calculate the average (mean) weight for that sample. Then, we will repeat this process indefinitely, so that we end up with an infinite number of samples, each of size 100, and each with its own mean weight value. Shown below are the first 20 such repetitions of the study:

Study 1:	Select 100 subjects, calculate mean weight	165 pounds
Study 2:	Select 100 subjects, calculate mean weight	157 pounds
Study 3:	Select 100 subjects, calculate mean weight	149 pounds
Study 4:	Select 100 subjects, calculate mean weight	166 pounds
Study 5:	Select 100 subjects, calculate mean weight	171 pounds
Study 6:	select 100 subjects, calculate mean weight:	148 pounds
Study 7:	Select 100 subjects, calculate mean weight	152 pounds
Study 8:	Select 100 subjects, calculate mean weight	158 pounds
Study 9:	Select 100 subjects, calculate mean weight	155 pounds
Study 10:	Select 100 subjects, calculate mean weight	163 pounds
Study 11:	Select 100 subjects, calculate mean weight	160 pounds
Study 12:	Select 100 subjects, calculate mean weight	157 pounds
Study 13:	Select 100 subjects, calculate mean weight	163 pounds
Study 14:	Select 100 subjects, calculate mean weight	161 pounds
Study 15:	Select 100 subjects, calculate mean weight	149 pounds

Study 16:	Select 100 subjects, calculate mean weight	166 pounds
Study 17:	Select 100 subjects, calculate mean weight	160 pounds
Study 18:	Select 100 subjects, calculate mean weight	152 pounds
Study 19:	Select 100 subjects, calculate mean weight	159 pounds
Study 20:	Select 100 subjects, calculate mean weight	151 pounds

The idea behind confidence intervals is to find some interval that can be placed around each sample mean, such that *95% of the intervals will contain the true population mean*, which in this case is 160 pounds. For our 20 samples above, a 95% confidence interval would be such an interval that includes the population mean 19/20 (95%) times. This interval turns out to be +/− 11 pounds. Applying this interval to each sample mean yields the following data:

Study 1:	Select 100 subjects, calculate mean weight	165 pounds (154, 176)
Study 2:	Select 100 subjects, calculate mean weight	157 pounds (146, 168)
Study 3:	Select 100 subjects, calculate mean weight	149 pounds (138, 160)
Study 4:	Select 100 subjects, calculate mean weight	166 pounds (155, 177)
Study 5:	Select 100 subjects, calculate mean weight	171 pounds (160, 182)
Study 6:	Select 100 subjects, calculate mean weight	148 pounds **(137, 159)**
Study 7:	Select 100 subjects, calculate mean weight	152 pounds (141, 163)
Study 8:	Select 100 subjects, calculate mean weight	158 pounds (147, 169)
Study 9:	Select 100 subjects, calculate mean weight	155 pounds (144, 166)
Study 10:	Select 100 subjects, calculate mean weight	163 pounds (152, 174)
Study 11:	Select 100 subjects, calculate mean weight	160 pounds (149, 171)
Study 12:	Select 100 subjects, calculate mean weight	157 pounds (146, 168)
Study 13:	Select 100 subjects, calculate mean weight	163 pounds (152, 174)
Study 14:	Select 100 subjects, calculate mean weight	161 pounds (150, 172)
Study 15:	Select 100 subjects, calculate mean weight	149 pounds (138, 160)
Study 16:	Select 100 subjects, calculate mean weight	166 pounds (155, 177)
Study 17:	Select 100 subjects, calculate mean weight	160 pounds (149, 171)
Study 18:	Select 100 subjects, calculate mean weight	152 pounds (141, 163)
Study 19:	Select 100 subjects, calculate mean weight	159 pounds (148, 170)
Study 20:	Select 100 subjects, calculate mean weight	151 pounds (140, 162)

The interval of +/− 11 pounds results in 19/20, or 95% of the intervals containing the true population mean of 160 pounds, and 1/20 of the intervals, shown in bold, not containing this mean. These intervals can also be appreciated graphically in Fig. 15.1.

Details regarding how to actually calculate 95% confidence intervals are beyond the scope of this book. However, it is important to note that 95% confidence intervals can be calculated *without any knowledge of the population mean*, using only the known mathematical properties of sampling. In fact, the width of the confidence interval is calculated from just the sample size and the sample variation. A narrower confidence interval results from a *larger sample size* and a *smaller sample variation*. This makes intuitive sense; a larger sample size and

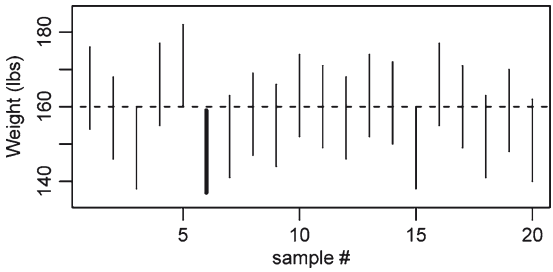


Fig. 15.1 95% confidence intervals for 20 samples of size 100

smaller degree of sample variation will result in having a better idea about the true mean value in the population.

In practice, it is important to be able to *interpret* confidence intervals in the research setting. Returning to the weight example, we will draw a new sample of 100 American adults, calculate their mean weight, and then use a computer to calculate the 95% confidence interval:

New sample: mean weight = 158 pounds, 95% confidence interval (147 pounds, 169 pounds)

In real life, we never know the actual population mean because we cannot practically sample the entire population. So, in this example we do not know that the true mean weight of all American adults is 160 pounds. We *do* know from the definition of a confidence interval that if our particular study, which consists of sampling 100 people and calculating their mean weight, were repeated indefinitely, and a 95% confidence interval placed around the mean weight from each sample, then 95% of these confidence intervals would contain the true mean weight in the population.

However, we have no way of knowing whether our particular 95% confidence interval (147 pounds, 169 pounds) happens to be one of the 95% "correct" confidence intervals that actually contains the true population mean. So it is not technically correct to interpret our confidence interval directly as, "there is a 95% chance that the population mean lies between 147 and 169 pounds." All we know is that if the experiment was repeated indefinitely, and a 95% confidence interval placed around each sample mean, then 95% of the intervals will contain the population mean. Since 95% of the confidence intervals will contain the population mean and 5% will not, we can be "95% confident" that our particular confidence interval is one of those that does contain the population mean, and therefore, we can state that we are "95% confident" that the population mean lies between 147 and 160 pounds.

For an analogy, imagine that you have been handed a lottery ticket. Prior to the drawing, the probability of winning the lottery with your ticket is 1 in 10 million. However, *after the drawing*, it is technically incorrect to consider the probability that you have a winning ticket; either you have a winning ticket, or you do not. The same is true for confidence intervals. The mean value for the population is some fixed quantity, it just happens to be unknown to us. The confidence interval for a particular sample either contains the true population mean, or it does not.

15.5 *P*-values

P-values relate the probability of observing a specific sample result, given some pre-specified null hypothesis regarding the population. The definition of a *p*-value is:

Given a null hypothesis regarding the population, the p-value is the probability of observing a particular sample result, or a more extreme result, due to sampling variation alone.

For example, investigators study whether aspirin use can lower the risk of heart disease in middle-aged men. They identify a sample of 100 middle-aged men and find that aspirin use is associated with a 20% lower risk of developing heart disease; relative risk = 0.8, *p*-value = 0.03.

The interpretation of this *p*-value would be: if there is truly no association of aspirin use with heart disease among the entire population of middle-aged men, then the chance of finding a relative risk of 0.8 or a more extreme relative risk, in some random 100-person sample is 3%. Stated another way, given no association between aspirin use and heart disease in the population of middle aged-men (a true relative risk of 1.0), then the chance of picking some unusual sample of 100 men and finding a 20% treatment effect, or an even greater treatment effect, due to sample-to-sample variation alone is only 3%.

Since there is only a 3% chance of obtaining this sample relative risk of 0.8, or a more extreme relative risk, in the setting of no true population association, it is reasonable to conclude that there probably *is* a true association of aspirin use with heart disease in the population. Results from this particular sample of 100 subjects were used to make an *inference regarding the population*.

To make life more complicated, *p*-values in research studies are almost always "two sided," meaning that the phrase "more extreme values" in the *p*-value definition refers to values that are greater than your result or less than the opposite result.

For the association of aspirin use with heart disease, the interpretation of the *p*-value was: given no association of aspirin use with heart disease among the population of middle-aged men, the probability of finding a relative risk of 0.8, or a more extreme relative risk in a random sample of size 100 is 3%. The meaning of a "more extreme relative risk" in this example refers to relative risks that are ≤0.8 *and* relative risks that are ≥(1.0/0.8), or ≥1.25, as shown in Fig. 15 2 below.

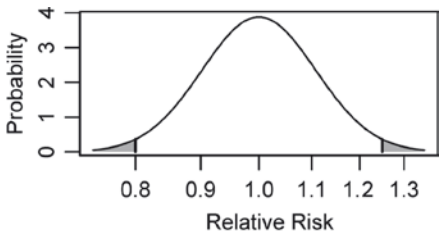


Fig. 15.2 Two-sided probability of finding a relative risk of 0.8 or a more extreme value

15.6 Confidence Intervals and *p*-values in Clinical Research

Consider the association of aspirin use with the development of heart disease as it might be presented in a journal article.

	Relative risk	P-value	95% confidence interval
Aspirin use	0.8	0.03	(0.65, 0.98)

The association has three components: the relative risk, the *p*-value, and the 95% confidence interval. Each has a distinct interpretation.

Interpretation of the relative risk: Among our particular study sample of 100 middle-aged men, aspirin use is associated with an estimated 20% lower risk of developing heart disease. The relative risk value relates the *strength* of the association between aspirin use and heart disease in this sample.

Interpretation of the *p*-value: Given no true association of aspirin use with heart disease in the population of middle-aged men, the chance of finding this sample relative risk of 0.8, or a more extreme relative risk, is 3%. In other words, finding a relative risk of 0.8 in a sample of 100 middle-aged American men suggests that there *is* a true association of aspirin use with heart disease in the entire population of middle-aged American men. The *p*-value provides *statistical evidence regarding a specific hypothesis about the population*.

Interpretation of the 95% confidence interval: If the study were conducted an infinite number of times, and a 95% confidence interval placed around each sample relative risk, then 95% of the confidence intervals constructed will contain the true relative risk in the population of middle-aged American men. In other words, these sample data are consistent with a true population relative risk between 0.65 and 0.98.

Confidence intervals and *p*-values provide complementary, but different information regarding statistical inference. In this example, the 95% confidence interval provides a *range of possible values* for the association of aspirin use with heart disease in the population.

A confidence interval that overlaps the null hypothesis leaves us less certain of a true association in the population. For example, if the association of aspirin use with heart disease was: relative risk = 0.8; 95% confidence interval (0.3, 1.3), we would be less certain that aspirin use was truly associated with heart disease risk among the population of middle-aged American men.

Confidence intervals also offer a convenient way to *compare associations*. For example, consider the association of aspirin use with heart disease stratified by diabetes status.

	Relative risk (95% confidence interval)
Diabetes	0.7 (0.5, 0.9)
No-diabetes	0.9 (0.7, 1.1)

These relative risks demonstrate that aspirin use is associated with a 30% lower risk of heart disease among diabetic subjects, but only a 10% lower risk of heart disease among non-diabetic subjects. However, the 95% confidence intervals for these relative risks overlap. These sample findings do *not* provide statistical evidence that the association of aspirin use with heart disease in the population is truly different between diabetic and non-diabetic individuals.

Chapter 16
Hypothesis Testing

Learning Objectives

1. Hypothesis testing evaluates the likelihood that results from a particular sample reflect those of the population from which it is drawn.
2. The null hypothesis is the entire universe of possibility that excludes the study hypothesis.
3. The distribution of sampling means is the distribution of mean values from an infinite number of samples of a specific size.
4. The distribution of sampling means has three important properties:
   (a) Normal distribution for large sample sizes
   (b) Mean equal to the population mean
   (c) Variation inversely related to sample size and directly related to population variation
5. The *p*-value for comparing the means of two samples is calculated using the mean and standard deviation of each sample.
6. The *p*-value is the probability of obtaining the sample result, or more extreme result, if the null hypothesis were true. A low *p*-value (< 0.05) therefore implies that the null hypothesis is probably false and that the *study hypothesis about the population* is probably true.

Hypothesis testing addresses the *statistical validity* of experimental results. Research studies are typically carried out among samples of individuals who are drawn from some larger population. A potential problem is that findings from a particular sample may differ substantially from the actual results in the population, due to inherent variation between samples. Typically, we do not know whether characteristics of a particular sample accurately reflect those of the population, because information regarding the full population is rarely available. Hypothesis testing utilizes mathema-tical knowledge regarding the natural variability expected from sampling to describe the likelihood that a particular sample result is representative of the underlying population from which it is drawn. The procedural framework for hypothesis testing is as follows:

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_16, © Springer Science+Business Media, LLC 2009

(1) Construct a study hypothesis *regarding the entire population*.
(2) Construct a null hypothesis, which is the diametric opposite of the study hypothesis.
(3) Assess the probability of obtaining your sample result *if the null hypothesis was true.*
(4) Reject the null hypothesis if this probability is very low.

Consider a study examining whether coffee consumption is associated with high blood pressure. One possible approach would be to measure blood pressures in 500 coffee drinkers and 500 non-coffee drinkers who live in Seattle and are willing to participate in a research study. The 500 coffee drinkers who participate in the study represent one of an infinite number of possible samples of 500 Seattle coffee drinkers that could be selected from the *population* of Seattle coffee drinkers. One might argue that study subjects are further characterized by a desire to participate in medical studies and the opportunity to observe study flyers posted around the University campus. Ok - we can more precisely define the *study population* to be all coffee drinkers who live in Seattle, are willing to participate in medical studies, and visit the University campus frequently enough to notice study flyers. Hypothesis testing specifically refers to the mathematical considerations that evaluate the likelihood that the sample of 500 Seattle coffee drinkers is representative of the underlying study population. Whether the study population is clinically, scientifically, or practically relevant is a subjective matter.

16.1 Study Hypothesis and Null Hypothesis

A reasonable study hypothesis for this example is that coffee drinkers have, on average, higher blood pressures than non-coffee drinkers. For the purposes of this experiment, we will focus on differences between *mean* blood pressures in the two groups. The mean is just one of many possible summary measures of blood pressure that could be studied. For example, it might be more clinically relevant to compare the *proportions* of coffee drinkers and non-coffee drinkers who have hypertension, defined by a systolic blood pressure greater than or equal to 140 mmHg, or a diastolic pressure greater than or equal to 90 mmHg.

We will now construct the following study hypothesis, written as H_a, *regarding the population*:

Study hypothesis: Mean systolic blood pressure of all Seattle coffee drinkers **is different than** mean systolic blood pressure of all Seattle non-coffee drinkers.

From this study hypothesis we then define an opposite null hypothesis, written as H_0:

Null hypothesis: Mean systolic blood pressure of all Seattle coffee drinkers **is equal to** mean systolic blood pressure of all Seattle non-coffee drinkers.

For this example, we will keep an open mind and construct the study hypothesis to consider the possibility that coffee drinking has *some* effect on blood pressure, either positive or negative. The null hypothesis is then constructed as the entire universe of possibility that excludes the study hypothesis - either Seattle coffee drinkers have a different mean systolic blood pressure than Seattle non-coffee drinkers (study hypothesis), or they do not (null hypothesis). There are no other possibilities, by design. We create the null hypothesis this way because the mathematical machinery of hypothesis testing can only provide evidence that a particular hypothesis is false. We purposely construct a null hypothesis in order to disprove it, using hypothesis testing, and thereby provide evidence that the alternative hypothesis is true.

We now turn to the task of providing mathematical evidence to disprove the null hypothesis. The evidence comes from properties of the *distribution of sampling means*.

16.2 Distribution of Sampling Means

For the purposes of this discussion, imagine that the average (mean) systolic blood pressure in the entire population of Seattle coffee drinkers is 145 mmHg. If we recruit 500 Seattle coffee drinkers to our study, we are unlikely to find a mean blood pressure of exactly 145 mmHg in this particular sample. In fact, there are an infinite number of possible samples of size 500 that could be drawn from the Seattle coffee drinking population, each with its own mean blood pressure, as shown in Fig. 16.1.

The left-hand portion of the Fig. 16.1 is a histogram of the relative frequency of systolic blood pressures in the population of Seattle coffee drinkers. We can randomly sample 500 individuals from this population, calculate the mean systolic blood pressure within the sample, plot this mean value on the right-hand graph, and then repeat the procedure indefinitely. The resulting right-hand portion of the Fig. 16.1 describes the distribution of mean values obtained from all possible samples of size 500. The distribution of all possible mean values obtained from samples of a given size is called *the distribution of sampling means*. A different distribution of sampling means exists for every possible sample size. The distribution of sampling means has important mathematical properties that are used to disprove the null hypothesis.

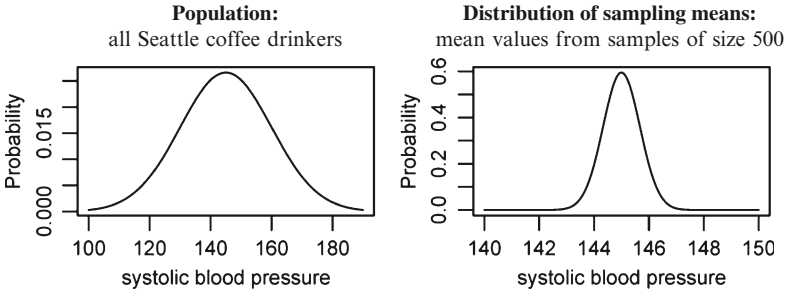


Fig. 16.1 Distribution of sampling means for samples of size 500

16.3 Properties of the Distribution of Sampling Means

16.3.1 Normal (Bell-Shaped) Distribution for Reasonably Large Sample Sizes

A powerful statistical theory called the central limit theorem tells us that the distribution of sampling means has essentially a "normal," or bell-shaped appearance, regardless of the shape of the population distribution from which it is drawn, provided that the sample size is reasonably large. "Large" is an arbitrary term but sample sizes of around 20–30 begin to give a normally shaped distribution of sampling means. For the coffee drinking example, the large sample size of 500 coffee drinkers assures that the distribution of sampling means will be bell-shaped.

How is it possible to know what the distribution of sampling means looks like without knowing the distribution of the underlying population? The averaging process (taking the mean for each sample) smoothes out outlying values, constraining the shape of the distribution of sample means. To see how this works, imagine that the distribution of blood pressure among the population of all Seattle coffee drinkers is not bell shaped, but is skewed, with most people having normal blood pressure, and a small group of people having high blood pressure (Fig. 16.2).

To create a distribution of sampling means (of size 500) from this population, we would draw an infinite number of samples of size 500, calculate the mean blood pressure within each sample, and then plot each mean value. The process of calculating the mean for each 500-person sample will smooth the underlying nonnormal shape of the population distribution, yielding a bell-shaped distribution of sampling means, as shown in Fig. 16.3.

In contrast, imagine that the sample size was equal to one, such that each experiment consisted of drawing only a single individual from the population, and then plotting their mean (only) blood pressure. In this case, infinite selection of samples of size 1 will produce a distribution of sampling means that is identical to the population distribution, because no smoothing, or averaging, is being applied to the samples. The central limit theorem would not apply in this case, because the sample size is too small, leaving us with no specific information about the appearance of the distribution of sampling means (other than its appearance will be identical to

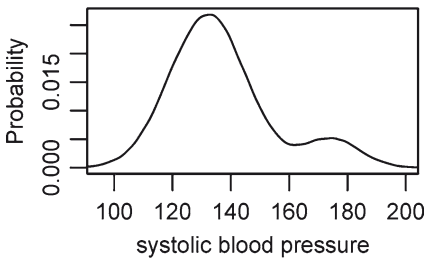


Fig. 16.2 Hypothetical population with non-normal blood pressure distribution

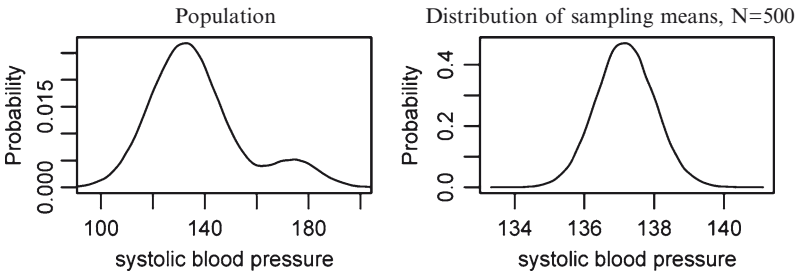


Fig. 16.3 Application of the central limit theorem

the population distribution, which we may never know). The fact that the distribution of sampling means assumes a known shape, or distribution, for relatively large sample sizes is essential for its use in probability and statistics.

16.3.2 Mean Equal to the Population Mean

A second important property of the distribution of sampling means is that its mean value is equal to the mean value from the population from which it is drawn. For example, if the mean systolic blood pressure of the population of all Seattle coffee drinkers is 145 mmHg, then the mean systolic blood pressure of the distribution of sampling means of size 500, or any other sample sizes, is also 145 mmHg. The property of equal means between the population and the distribution of sampling means holds regardless of the sample size. Note that this does *not* imply that the sampling mean for any one particular sample is equal to the mean in the population - just that the *means* of the two distributions are the same.

16.3.3 Spread of the Distribution Related to Population Variation and Sample Size

The distribution of sampling means is characterized by having some mean value (center of the distribution), and some variance (width of the distribution). Distributions with more variance will appear wider, or more spread out, than distributions with less variance. The variance of a distribution is typically described using the term *standard deviation*, which is just the square root of the variance. The standard deviation for the distribution of sampling means has a specific name, the *standard error of the mean.* The standard error of the mean is influenced by two factors: (1) the variance of the population and (2) the sample size. Mathematically,

$$\text{standard error of the mean} = \sigma / \sqrt{n}$$

where σ is the standard deviation of the population, and n is the sample size.

16.3.3.1 Variance Directly Related to Population Variance

To understand how the standard error of the mean relates to the standard deviation of the population, consider two hypothetical populations, one with systolic blood pressures ranging from 80 to 240 mmHg, the other with systolic blood pressures ranging from 110 to 150 mmHg. If random samples of a fixed size were selected from each population, then we would expect samples from the first population to have more variability, as shown in Fig. 16.4.

16.3.3.2 Variance Inversely Related to Sample Size

The width of the distribution of sampling means is *inversely* proportional to the sample size, meaning that larger sample sizes drawn from a predefined population will result in a narrower distribution of sampling means. Reconsider the example of selecting a single Seattle coffee drinker from the population, such that the sample size is one. Each experiment consists of recording the blood pressure of a single, randomly selected person from the population. If this experiment were repeated indefinitely, the resultant distribution of sampling means would appear identical to the population distribution, meaning that the variation in blood pressure for the distribution of sampling means would be equal to the variation in the population. On the other hand, imagine that we selected random samples of 1,000,000 Seattle

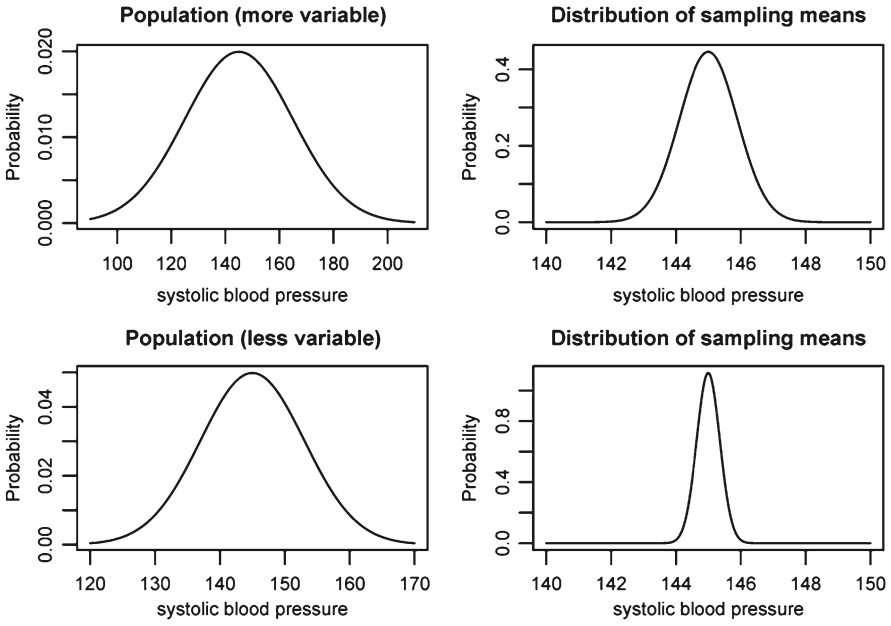


Fig. 16.4 Variation in sampling means directly related to variation in the population

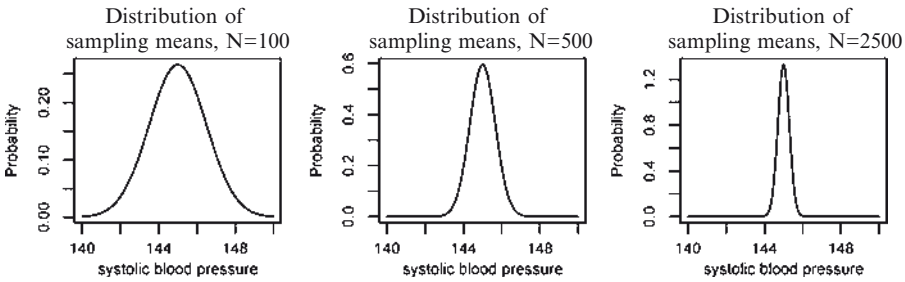


Fig. 16.5 Variation in sampling means inversely related to sample size

coffee drinkers for our study, such that each sample consisted of the entire Seattle coffee drinking population. In this case, each experiment would return the same mean blood pressure of 145 mmHg, which is exactly the mean blood pressure in the population. If this experiment were repeated indefinitely, the distribution of sampling means would be a single value, 145 mmHg, which is as thin as a distribution can get. These extreme examples demonstrate that calculation of a mean value for increasingly large sample sizes will smooth outlying values, and will lower the variation of the distribution of sampling means, creating a distribution that appears increasingly more narrow (less varied) than the population distribution from which it is drawn. The relationship between sample size and width of the distribution of sampling means is depicted in Fig. 16.5.

16.3.4 Distribution of Sampling Means: Summary

To summarize, the distribution of sampling means has three important properties:

1. Bell-shaped appearance (distribution) if the sample size is reasonably large
2. Mean equal to the population mean
3. Variation directly related to the population variation, and inversely related to sample size

16.4 Conducting the Experiment

Armed with properties of the distribution of sampling means, we now return to our original experiment, testing whether systolic blood pressure is significantly different between our sample of 500 coffee drinkers and our sample of 500 non-coffee drinkers. If the population of Seattle coffee drinkers has a mean blood pressure of x_1 and a standard deviation of σ_1, then properties of the distribution of sampling means tell us that the distribution of sampling means in this case will have a normal distribution, a mean equal to the population mean of x_1, and a standard error of the mean equal to $\sigma_1 / \sqrt{500}$.

Similarly, if the population of non-coffee drinkers has a mean blood pressure of x_2, and a standard deviation of σ_2, then the distribution of sampling means for size 500 will have a normal distribution, a mean equal to x_2 and a standard error of the mean equal to $\sigma_2 / \sqrt{500}$.

To answer the question of whether blood pressure is statistically different between coffee drinkers and non-coffee drinkers, we need to obtain one new distribution of sampling means: the distribution of sampling means for the *difference* in mean blood pressure between coffee drinkers and non-coffee drinkers. The following procedure will create this distribution.

1. Sample 500 coffee drinkers and 500 non-coffee drinkers from their respective populations.
2. Calculate the mean blood pressure for each sample.
3. Calculate the difference in mean blood pressure between the two samples.
4. Plot the difference on the distribution of sampling means for the difference in blood pressure.
5. Repeat this process indefinitely.

Or, since we know that the distribution of sampling means for blood pressure of coffee drinkers and the non-coffee drinkers is normally distributed, we can infer that the distribution of sampling means for the difference in blood pressure is also normally distributed.

It is fairly straightforward to show that if the coffee drinking and non-coffee drinking populations have mean blood pressures of x_1 and x_2, respectively, then the distribution of sampling means for the difference in blood pressure, comparing coffee drinkers to non-coffee drinkers, will be $x_1 - x_2$. Calculating the standard deviation for the distribution of sampling means for the difference in blood pressure is less straightforward. However, this standard deviation is derived from a combination of the standard deviations from the coffee drinkers and non-coffee drinkers.

The distribution of sampling means for the difference in blood pressure is used to directly test the null hypothesis. Our null hypothesis states that mean blood pressures for the *population* of coffee drinkers and non-coffee drinkers are equal, or that $x_1 = x_2$. Because the distribution of sampling means has the same mean value as the underlying population, the distribution of sampling means for coffee drinkers and non-coffee drinkers will also be equal if the null hypothesis is true, implying that the distribution of sample means for the difference in blood pressure would have a mean value of 0.

Let's go ahead and do the experiment:

Recruit 500 Seattle coffee drinkers, calculate mean blood pressure: 147 mmHg

Recruit 500 Seattle non-coffee drinkers, calculate mean blood pressure: 142 mmHg

Calculate difference in mean blood pressure: 5 mmHg

Question: Is blood pressure significantly different, comparing coffee drinkers to non-coffee drinkers?

In other words, given that we found a difference of 5 mmHg in our particular samples of size 500, how likely is it that blood pressure is actually higher among the population of Seattle coffee drinkers compared to the population of

Seattle non-coffee drinkers? Stated another way, can our sample results provide information regarding the likelihood of a true difference in blood pressure among the population? This is a standard question of statistical inference. Given that we found a 5 mmHg difference in blood pressure among a particular sample of 500 coffee drinkers, and a particular sample of 500 non-coffee drinkers, how likely is it that that the mean blood pressure among the *population* of coffee drinkers is equal to the mean blood pressure among the *population* of non-coffee drinkers?

The answer to this question comes directly from the distribution of sampling means for the difference in blood pressure. We have already concluded that if the null hypothesis were true, this distribution would have a mean value equal to 0. So, we can ask, if the null hypothesis were true, what is the probability of finding a difference of 5 mmHg, or a more extreme difference, from any particular experiment? This probability is represented by the area under the curve to the right of 5, and to the left of −5, as shown in Fig. 16.6.

The distribution of sampling means for the difference in blood pressure shows that if the null hypothesis were true, there are many possible experimental results that evaluate the difference in mean blood pressure between 500 coffee drinkers and 500 non-coffee drinkers. If there were truly no difference in mean blood pressure among the population of coffee drinkers and non-coffee drinkers, then most experiments using samples of size 500 would likely find results that are close to 0. However, it is possible to select unusual samples of 500 coffee drinkers and 500 non-coffee drinkers, such that the difference in blood pressure between the samples is 5 mmHg, or even greater. The question is how likely is it for that to happen?

To calculate this probability we need to know the area under the curve for a distribution with mean = 0 and standard deviation that derives from the standard deviation of coffee drinkers and non-coffee drinkers. Because the standard deviation is estimated from the sample data, this distribution is no longer a normal distribution, but a closely related one called the *t*-distribution. Calculating probabilities from *t*-distributions can be performed by computer, or by using *t*-distribution tables. For this example, we can enter the following data into the computer:

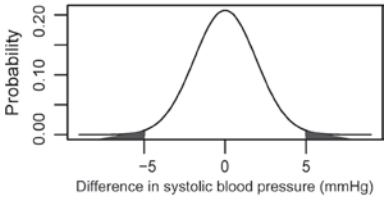


Fig. 16.6 Distribution of sampling means for the difference in blood pressure under the null hypothesis

Sample 1	Size 500	Mean 147	Standard deviation 28.5
Sample 2	Size 500	Mean 142	Standard deviation 32.1

The computer will return the following information:
Sample 1: mean = 147, standard error of the mean = 1.27
Sample 1: mean = 142, standard error of the mean = 1.44
Two sample t-test, p = 0.009

The specific data needed to calculate the p-value are the mean and the standard error of the mean for each sample. Recall that the standard error of the mean is calculated as the standard deviation divided by the square root of the sample size, or $\sqrt{500}$ for this problem.

The p-value represents the area under the curve of Figure 16.6 to the right of +5 plus the area under the curve to the left of –5. Multiplying the p-value by 100% returns the percent chance of finding our sample result, or a more extreme result, given the null hypothesis of equal blood pressures between coffee drinkers and non-coffee drinkers in the Seattle population. In this case, if the null hypothesis were true, then the chance of finding a difference of 5 mm Hg, or a more extreme difference, in our samples is 0.9%. So, although it is possible to find a difference of 5 mm Hg or a more extreme difference in samples of size 500 under the null hypothesis, the chance of this happening is very unlikely and we are therefore forced to reject the null hypothesis. By our strategy of creating diametrically opposite null and alternate hypotheses, rejection of the null hypothesis leaves us no other alternative than to conclude that the study hypothesis is likely to be true. In this way, we have used results from our particular experimental sample data to make an inference about the results in the population.

Note that areas from both sides of the curve were added to obtain the *p*-value. This procedure is known as a *two-sided test*, and is most commonly used. The interpretation relates to finding a blood pressure difference of 5 mmHg, or a more extreme difference, which could mean finding a >5 mmHg higher *or lower* blood pressure, comparing coffee drinkers to non-drinkers, under the null hypothesis that blood pressure is equal between the two groups.

Chapter 17
Interpreting Hypothesis Tests

Abigail Shoben

Learning Objectives

1. The *t*-test is used to compare mean values between two different groups.
2. The Chi-square test is used to compare proportions between two different groups.
3. The ANOVA test is used to compare mean values across multiple groups.
4. A type I error occurs when a hypothesis test declares a result to be statistically significant when in fact no true effect or association exists in the population.
5. Replication of study findings is an effective method for addressing type I errors.
6. The Bonferroni correction addresses the problem of multiple comparisons by setting a more stringent *p*-value threshold for statistical significance.
7. A type II error occurs when a hypothesis test declares a result to be statistically insignificant when in fact a true effect or association exists in the population.
8. Power is the pre-specified probability that a particular study will *not* make a type II error.
9. Power increases with
   (a) A larger sample size
   (b) Decreased variability of measurements within individuals
   (c) A greater pre-specified effect or association

17.1 Common Tests of Hypothesis in Clinical Research

17.1.1 T-Tests

The coffee drinking and blood pressure example from Chap. 16 represents a specific case of the general procedure of comparing mean values between two groups. *The t-test is a statistical test that is used to compare mean values between two different groups. T*-tests are widely used in clinical research studies. For example, investigators may wish to compare the mean body mass index of patients treated with a new weight loss medication to the mean body mass index of patients who receive

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_17,

weight loss counseling. For a second example, investigators may wish to compare mean values from a depression index scale among groups of patients who receive a new antidepressant medication versus a traditional serotonin reuptake inhibitor.

The *t*-test returns a *p*-value that represents the *probability of finding the observed sample results, or more extreme results, if the null hypothesis regarding the population is true*. A statistically significant *p*-value, typically <0.05, represents a low probability that the null hypothesis is true, implying that the mean values of the two groups are likely to be different in the population.

17.1.2 Chi-Square Tests

A second common comparison in clinical research studies is the comparison of proportions. *The chi-square test is a statistical test that is used to compare proportions between two different groups.* For the coffee drinking and blood pressure example in Chap. 16, the chi-square test could be used to compare the proportion of hypertension among coffee drinkers and non-coffee drinkers, defining hypertension as a systolic blood pressure ≥140 mmHg, diastolic blood pressure ≥90 mmHg, or the use of an antihypertensive medication. The null hypothesis for this chi-square test would be that the proportion of hypertension is equal in the populations of Seattle coffee drinkers and non-coffee drinkers.

For a second example, researchers document a 5% incidence of middle ear infection among children vaccinated with the pneumococcal vaccine and a 10% incidence of middle ear infection among children who were not vaccinated. The chi-square test could be used to compare these proportions. The *p*-value from this chi-square test is interpreted as the probability of finding this 5% difference in proportions observed in the study sample, or a more extreme difference in proportions, given no true difference among the population. Hypothesis testing using the chi-square test is conceptually similar to the *t*-test. Taken together, the chi-square test and the *t*-test cover many situations in clinical research.

17.1.3 ANOVA Tests

Sometimes research studies involve more than two comparison groups. For example, investigators might randomize study participants to receive one of four different antihypertensive medications. The ANOVA test is used to compare the mean values across multiple groups, and therefore represents a generalization of the *t*-test, which is limited to comparing only two groups. Like the *t*-test, the ANOVA test is used to evaluate the *mean values of continuous variables* between different groups.

The null hypothesis for the ANOVA test is that *all* of the group means are equal to each other in the population. The *p*-value for the ANOVA test indicates

Table 17.1 Comparison of mean blood pressures among four treatment groups

	Number of subjects	Mean blood pressure at study end (mmHg)
Treatment A	115	140
Treatment B	97	142
Treatment C	155	128
Treatment D	105	152
P-value (ANOVA)		0.001

the probability of observing means as different or more different than those observed in the study sample, given that all of the mean values are the same in the population. A statistically significant ANOVA *p*-value indicates that the means are somehow different from each other; however a significant *p*-value does not specify *which* specific mean or means may be different.

Consider an example of comparing mean blood pressures after treatment with one of four different antihypertensive medications, as illustrated in Table 17.1.

Here, the *p*-value for the ANOVA test indicates that *one or more* of the treated mean blood pressures are statistically different. Stated another way, if end-of-study mean blood pressures were truly equal among entire populations of subjects given treatments A, B, C, and D, then the chance of observing these sample results, or more extreme results, is 0.001 (0.1%). The ANOVA *p*-value tells us that at least one of the mean values is statistically different among the groups.

The observed sample data suggest that treatment C might be superior to the other treatments. These results could be followed by three individual *t*-tests comparing the mean blood pressure in treatment group C to the mean blood pressure in treatment groups A, B, and D, although such tests should correct for multiple comparisons, which are explained below.

17.2 An Imperfect System

17.2.1 Type I Errors

It is important to note that statistical significance may be observed even if the null hypothesis is true, simply due to chance. If the statistical significance level is set to the near-universal threshold value of 0.05, then there is a 5% chance that an experiment could detect a statistically significant result in the study sample even if the result were truly negative in the population. Conceptually, if 100 randomized trials were performed comparing the effect of two identical medications, five of these 100 trials would be expected to observe a "statistically significant" difference between the two medication groups even though they have identical effects.

A *type I error* occurs when a hypothesis test declares a result to be statistically significant even though the null hypothesis is true (there is no true effect or association in the population). Type I errors are a major reason that study results should be *replicated in other studies*; if there is a 5% chance of a type I error in one study, then the chance of two independent studies both finding a statistically significant result when there is no true difference in the population is only 0.25%.

Type I errors can arise when performing multiple hypothesis tests on the same data, for example a study exploring a list of potential hypertension risk factors. Because each individual hypothesis test has a type I error rate of 5% under the statistical significance threshold of 0.05, every 20 hypothesis tests would be expected to yield one significant test due to chance alone, even if none of the evaluated risk factors are truly significant in the population.

A simple correction to this *problem of multiple comparisons* is called the *Bonferroni correction*. This procedure limits the experiment-wide type I error rate to 5% by setting a more stringent *p*-value threshold for declaring a study result to be "significant." For example, if an experiment tests 12 risk factors for hypertension, the *p*-value threshold for declaring each risk factor to be statistically significant would not be 0.05, but instead would be 0.05/12 = 0.0042. A classic example of the multiple comparisons problem is whole genome association studies, which may perform millions of hypothesis tests regarding individual single nucleotide polymorphisms. *P*-value thresholds for declaring statistical significance in these studies may be set to 10^{-9} or even lower. More sophisticated methods that are used to account for multiple hypothesis testing are beyond the scope of this book.

17.2.2 Type II Errors

A *type II error* occurs when a hypothesis test declares a result to be statistically insignificant even though there is a true difference in the population. Type II errors, in which potentially important positive findings are missed due to chance, are generally considered to be less toxic than type I errors, which report falsely positive findings. Type II errors often occur *due to inadequate study power*, which is discussed below.

17.2.3 Power

Power is the probability that a particular study will *not* make a type II error. Power represents the ability of a statistical test to detect some specified difference or effect.

Example 17.1. Investigators design a randomized clinical trial to compare the blood pressure lowering effects of two novel antihypertensive agents. Before the trial begins, the investigators decide that a 10 mmHg difference in blood pressure between treatment groups is "clinically important." Power calculations reveal that the study has 80% power to detect this pre-specified 10 mmHg difference in blood pressure.

The interpretation of 80% study power in this example is that there is a 20% chance that this study will return a statistically nonsignificant result even if a true 10 mm Hg treatment effect, or greater treatment effect, exists in the population.

What factors might cause a study to miss potentially important effects or associations? Study power is affected by four major factors: sample size (N), variability of measurements within individuals (σ), the magnitude of the true effect (often the difference in means between two groups), and the significance level of the test (α), usually fixed at 0.05. These factors influence perception of the p-value for a study result, so let's examine each of them more carefully.

17.2.3.1 Sample Size

Statistical power increases with increasing sample size, such that a study with 5,000 subjects is considerably more likely to detect a statistically significant result compared to a study with 500 subjects. An implication of this relationship is that *large clinical studies have the ability to detect very small effects that may be statistically significant, but not clinically important*. For example, a large randomized trial may detect a statistically significant 0.03 mmHg difference in systolic blood pressure comparing two different antihypertensive medications (p-value = 0.01). When evaluating the importance of these findings, it is important to look at the *magnitude of effect* and the *confidence intervals*, not just the statistical significance (p-values). The 95% confidence interval for such a blood pressure study might be (+0.01, +0.05 mmHg), suggesting that these data would not be unusual if the true difference in blood pressure among the population were between 0.01 and 0.05 mmHg, a clinically insignificant difference. On the other hand, a smaller study might find that a new hypertension drug lowers systolic blood pressure by 15 mmHg compared to conventional agents (p-value = 0.15, 95% confidence interval +2, −32 mmHg). The confidence interval and p-value do not exclude the possibility that the drug is actually ineffective in the population, but does suggest that the drug may have clinically important effects and that the study may have been underpowered, motivating further studies of this medication.

It should also be noted that large study populations are more likely to contain people with unusual characteristics in the population. For example, if a drug causes a rare toxic side effect in 0.2% of the population, a sample size of 100 is unlikely to contain such a person, but a sample size of 5,000 will, on average, contain 10 such people. The study of rare side effects is another strength of observational studies of medication use (pharmacoepidemiology), which have the opportunity to evaluate very large numbers of treated patients.

17.2.3.2 Variability

Statistical power increases with decreasing variability of the outcome measurement. In other words, a more tightly spaced distribution of the outcome variable will yield greater statistical power compared to a more widely spaced distribution. Consider two hypothetical distributions of the difference in mean blood pressure from a clinical trial that compares the effect of different antihypertensive medications, shown in Fig. 17.1.

Both distributions demonstrate an approximate 7 mmHg mean difference in blood pressure between the two medications. However, there is considerable spread, or variability, of this difference in the left-hand distribution. In contrast, the effect of the blood pressure medications is more tightly spaced in the right-hand distribution, which has greater power to demonstrate that the 7 mmHg mean difference is statistically significant.

In general, studies that evaluate highly variable outcome measurement, such as coronary artery calcium or dietary fat intake, will have less study power, and therefore require more subjects to demonstrate statistical significance, compared to studies that evaluate more consistent outcome measures.

17.2.3.3 Effect Size

Statistical power increases fairly dramatically with increasing effect size. This means that it is much easier to detect a significant result when the magnitude of an effect or association is large. For example, a small study may contain adequate power to demonstrate a statistically significant treatment effect for a breakthrough cancer drug that reduces cancer mortality by 50%. In contrast, very large sample sizes are needed to power studies to detect subtle treatment effects, for example, comparing the effect of two drugs within the same class. The relationship between effect size and sample size is demonstrated in the Table 17.2, which describes three hypothetical risk factors for cancer that are strong, moderate, and relatively weak.

A statistically significant association of risk factor one with cancer can be detected in a study of only 170 subjects, whereas a statistically significant association of risk factor three with cancer will require 1,548 subjects. Given a fixed

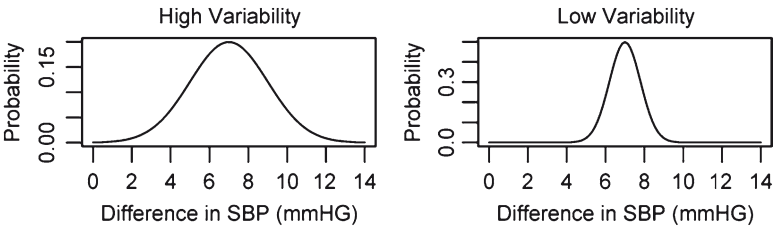


Fig. 17.1 Mean difference in end-of-study blood pressures: medication 1 – medication 2

Table 17.2 Power calculations demonstrating effect size and number of subjects

	Relative risk of cancer	Number of subjects needed to show that the risk factor is significantly associated with cancer
Risk factor one	2.0	170
Risk factor two	1.5	388
Risk factor three	1.2	1,548

number of subjects, studies of relatively strong risk factor one will have considerably more power to detect a statistically significant association than studies of weaker risk factors two and three.

17.2.3.4 Significance Level (α)

Significance level, the *p*-value threshold for defining a statistically significant result, is almost always fixed at 0.05, however it is useful to note that study power will increase if the significance level is set to a value greater than 0.05. Conversely, power will decrease if the significance level is set to a value less than 0.05, for example when using the Bonferroni correction for multiple comparisons. This property makes sense in that it will be easier to detect a "statistically significant" result when the threshold for significance is set to 0.2, as compared to 0.001.

Chapter 18
Linear Regression

Learning Objectives

1. A regression line is identified by the smallest sum of squared distances from the data.
2. A *residual* is the difference between the data predicted from a model and the actual data.
3. Potential pitfalls in fitting a linear regression model are influential data points and nonlinear associations.
4. In a multiple linear regression model each coefficient represents the independent association of the covariate with the outcome variable, holding all other variables constant.
5. The null hypothesis for a coefficient in a simple linear regression model is that the coefficient is equal to 0.
6. Confounding can be detected by a substantial change in the coefficient of interest after including the potential confounding variable in the multiple regression model.

Regression is a mathematical method that is used to describe the relationship, or association between two or more factors. *Multiple regression* is a widely used tool in clinical/epidemiological research to adjust for confounding, and has important advantages over other methods, such as restriction, stratification, and matching.

18.1 Describing the Association Between Two Variables

We begin by considering a new research study that seeks to examines the association between vitamin D and inflammation. For background information, vitamin D demonstrates anti-inflammatory properties in cell culture and animal models. To investigate whether vitamin D might be related to inflammation *in humans*, a research team recruits 50 patients from an internal medicine clinic and assays serum levels of 25-hydroxyvitamin D, a marker of vitamin D stores, and interleukin 6 (IL-6), a circulating pro-inflammatory hormone. Raw data from the first 10 patients in the study are presented in Table 18.1.

The data in Table 18.1 provide a sense that higher 25-hydroxy vitamin D levels tend to track with lower circulating IL-6 levels. How can we better describe this

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_18, © Springer Science+Business Media, LLC 2009

Table 18.1 Serum vitamin D and interleukin 6 measurements from 10 subjects

Patient number	25-hydroxy vitamin D level (ng/ml)	Interleukin 6 level (pg/ml)
1	5.0	2.7
2	83.2	0.3
3	49.1	0.7
4	68.0	1
5	33.4	0.4
6	8.0	3.2
7	48.8	1.4
8	76.4	1.9
9	47.5	2.3
10	36.1	3.2

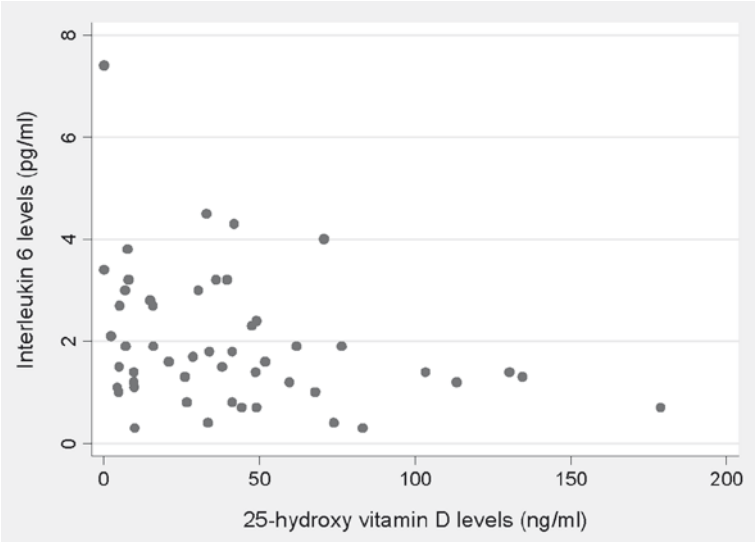


Fig. 18.1 Cross-sectional scatter plot of vitamin D and interleukin 6 levels

relationship? A possible next step is to graph the data for all 50 study subjects, as shown in Fig. 18.1.

The graph seems to confirm our suspicion that higher 25-hydroxyvitamin D levels track with lower circulating IL-6 levels; consistent with our hypothesis that vitamin D might possess antiinflammatory properties. However, we have only the scatter plot for evidence. We have not formally *tested the hypothesis* that lower 25-hydroxyvitamin D levels are associated with higher IL-6 levels. Moreover, the scatter plot does not provide quantitative information regarding the *strength of the association* between vitamin D and IL-6, i.e. about how about much different are IL-6 levels for each unit difference in 25-hydroxyvitamin D?

We have previously used the t-test to compare mean values between two different groups. The *t*-test could be applied here if we are willing to divide the naturally

continuous 25-hydroxyvitamin levels into two groups. Some experts have recommended that vitamin D deficiency be defined by a 25-hydroxyvitamin D level <15 ng/ml. Based on this definition, we could examine mean IL-6 levels by vitamin D deficiency status.

	Mean IL-6 level in pg/ml (standard deviation)
Vitamin D < 15 ng/ml (n = 18)	2.4 (1.6)
Vitamin D ≥ 15 ng/ml (n = 32)	1.7 (1.1)

Applying the t-test to these data yields a p-value of 0.11. The interpretation of this p-value is, "if IL-6 levels are equal among vitamin D replete and vitamin D deficient people in the population, then the chance of observing these sample results or more extreme results is 11%."

The use of the t-test in this situation involves compromises. The data must be divided into two groups using some cut point. We selected one particular definition of vitamin D deficiency (<15 ng/ml), but other definitions, such as <20 ng/ml, are equally valid. Using different cut points to define vitamin D deficiency will result in different results from the t-test. Further, dividing the data into 2 groups prevents evaluation of vitamin D and IL-6 levels across the full measured spectrum of these markers. An alternate method for studying the relationship between vitamin D and IL-6 levels is linear regression. To demonstrate linear regression, we return to the plot of vitamin D and IL-6 levels, this time fitting a 'regression line' through the data points as shown in Fig. 18.2.

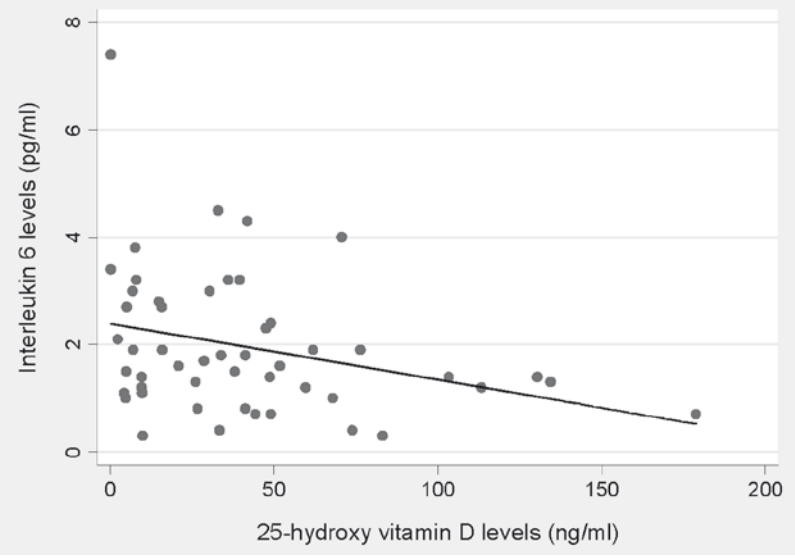


Fig. 18.2 Scatter plot of vitamin D and interleukin 6 levels, with regression line

The regression line is selected as the line that *most closely fits all of the data points*. Imagine an infinite number of possible lines that could be drawn through these data in Fig. 18.2. For each possible line, the distances between each data point and the line are summed. The regression line is defined as the one line, of all possible lines, that has the smallest sum of squared distances from the data points. Stated another way, the regression line *lies closest to all of the data points* where "close" is measured in squared distance.

18.2 Univariate Linear Regression

18.2.1 The Linear Regression Equation

All lines are characterized by having some *slope*, and some *intercept* (the *Y* value when $X = 0$), as described by the equation:

$$Y = mX + b,$$

where *m* is the slope and *b* is the intercept.

It follows that our regression line, relating vitamin D and IL-6 levels, has the form:

$$\text{IL-6} = \text{m} * (\text{vitamin D level}) + \text{b}$$

Standard parlance for regression analysis is to use the term "β_1" to represent the slope, and the term "β_0" to represent the intercept. So, our regression equation can be rewritten as:

$$\text{IL-6 level} = \beta_0 + \beta_1 * (\text{25-hydroxyvitamin D level})$$

Before examining the solution to this equation, some definitions are needed.

1. **Model**: The whole equation, i.e., IL-6 level = $\beta_0 + \beta_1 *$ (25-hydroxyvitamin D level)
2. **Coefficients**: The beta terms, β_0 and β_1. Their values are not known yet.
3. **Dependent variable**: The factor being predicted. In this case the dependent variable is IL-6 levels. The dependent variable resides on the left side of a regression equation.
4. **Independent variables or covariates**: The factor(s) in the model used to predict the dependent variable. In this case, the only independent variable is 25-hydroxyvitamin D levels. The independent variables reside on the right side of a regression equation.

Examining the graph in Fig. 18.2 yields an estimated a slope of -0.01 and an estimated intercept of 2.4. Therefore, the equation relating IL-6 and vitamin D levels in our study is:

$$\text{IL-6 level} = 2.4 - 0.01 * (\text{25-hydroxyvitamin D level})$$

This linear regression model can be interpreted as follows:

For each 1 ng/ml higher serum 25-hydroxyvitamin D level, serum IL-6 levels are, on average, 0.01 pg/ml lower.

This interpretation provides a quantitative estimate of the linear association of 25-hydroxy vitamin D levels with IL-6 levels across the full range of measured values for these measurements.

18.2.2 Residuals and the Sum of Squares

After fitting the regression equation, each study participant will have a predicted IL-6 level according to the model, and their own, actual measured IL-6 level. The term *residual* is defined as the difference between the actual study data and the data that is predicted from the regression equation. Graphically, residuals are described by the vertical distance between a particular data point and the regression line, as depicted in Fig. 18.3.

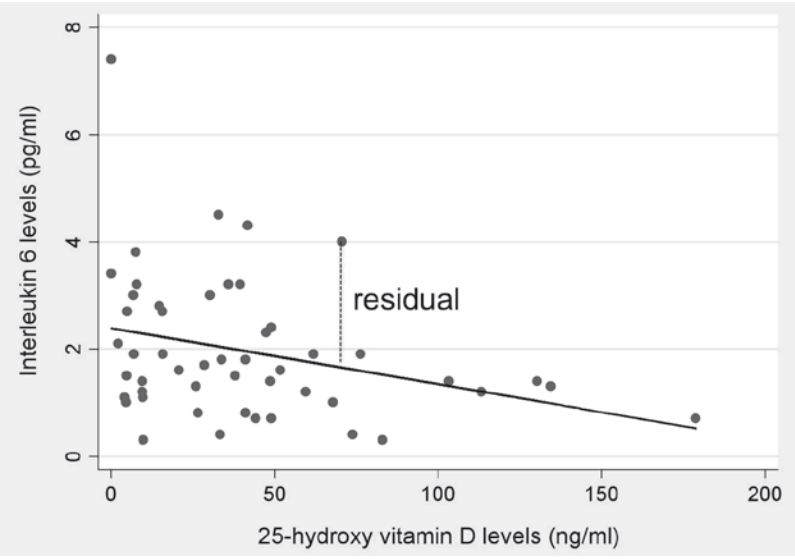


Fig. 18.3 Residual value for a single data point in a linear regression model

For example, study subject #1 has a measured 25-hydroxy vitamin D level of 5 ng/ml. According to the regression equation, his predicted IL-6 level should be:

$$\text{IL6} = 2.4 - 0.01 * (\text{25-hydroxyvitamin D}) = 2.4 - 0.01 * (5) = 2.35\text{pg/ml}$$

However, subject #1's actual measured IL-6 level is 2.7 pg/ml. This subject's residual (defined as observed – predicted) = 2.7 – 2.35 = 0.35 pg/ml. In aggregate, residuals provide a sense of how accurately the regression line fits the study data. The regression line with the smallest sum of residuals fits the data as tightly as possible. In practice, residuals are squared, so the best-fit regression line is identified as the one

line, of all possible lines, that has the *lowest sum of squared residuals for all data points*. The mathematical methods (calculus) used to find the equation with the lowest sum of squared residuals are beyond the scope of this book. Statistical software packages will quickly find the solution for a regression line from a set of raw data.

18.2.3 Absolute Versus Relative Fit

The regression line is an artificial construct that does not go through most (if any) of the actual data points. The sum of squares procedure will identify a line that best fits the study data; however, there is no assurance that the best fitting line will be good *in any absolute sense*. A number of statistical tests are designed to evaluate how well a fitted regression line fits the study data. One simple approach is to compare the fitted regression line with a plain, horizontal line through the mean of the data. In this case, the mean IL-6 level is 1.96 pg/ml, so we will compare our filled regression line to a horizontal line drawn through the data at 1.96 pg/ml (Fig. 18.4).

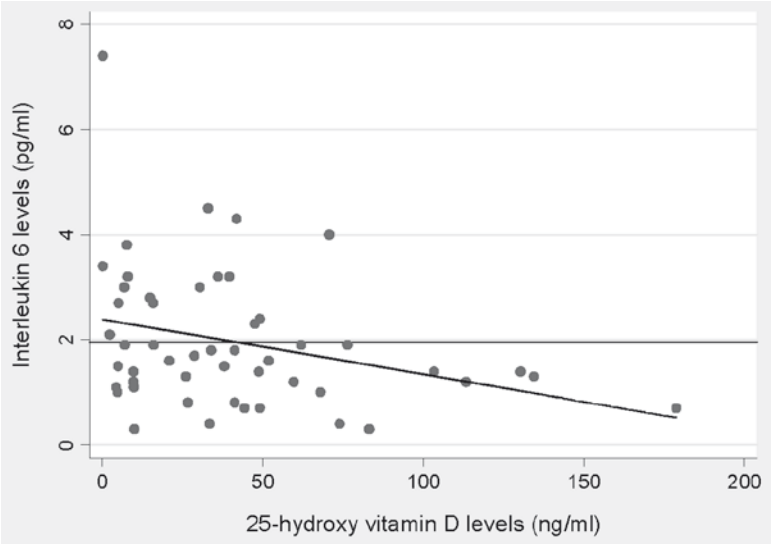


Fig. 18.4 Comparison of regression line to horizontal line through the mean

We can calculate the sum of squared residuals for each line, and then compare these sums using the concept of the null hypothesis construct introduced earlier. In this case, the null hypothesis is: there is no linear trend between vitamin D and IL-6 levels, or that the best fitting line has a slope equal to 0. The alternative hypothesis is that there *is* a linear association between vitamin D and IL-6 levels, or that the best fitting line has a slope that is nonzero. The null hypothesis is assessed by statistically comparing the sum of squares for the horizontal line (the null hypothesis)

to the sum of squares for the calculated regression line (the alternative hypothesis). If the regression line has a significantly smaller sum of squares, then the fitted regression line more accurately explains the observed study data and the null hypothesis of no association is rejected.

18.3 Interpreting Results from Univariate Regression Equations

18.3.1 Interpreting Continuous Covariates

Using the sum of squares method, we found the best fitting regression line for vitamin D and IL-6 levels to be:

$$\text{IL6} = 2.4 - 0.01*(\text{25-hydroxyvitamin D})$$

The interpretation of the coefficient for vitamin (-0.01) is, "for each 1 ng/ml greater 25-hydroxyvitamin D level, IL-6 levels are, on average 0.01 pg/ml lower." The sign (positive or negative) before a specific coefficient specifies whether an association is positive (higher levels of the covariate associated with higher levels of the dependent variable), or negative (higher levels of the covariate associated with lower levels of the dependent variable). In this case the coefficient for vitamin D is negative indicating an inverse association.

The linear regression equation can be used to compare expected differences in IL-6 levels across different levels of vitamin D. For example, given two study subjects who have 25-hydroxy vitamin D levels of 20 and 10 ng/ml, we would expect IL-6 levels to differ, a average, by 0.01 * (20 – 10) = 0.1 pg/ml. Specifically, we would expect the subject with a vitamin D level of 20 ng/ml to have an IL-6 level that is 0.1 pg/ml *lower*, on average, than the subject with a vitamin D level of 10 ng/ml. Similarly, for two study subjects with vitamin D levels of 110 and 100 ng/ml, we would also expect IL-6 levels to differ by 0.01 * (110 – 100) = 0.1 pg/ml because we are assuming a constant linear difference in IL-6 levels according to the slope of −0.01.

18.3.2 Interpreting Binary Covariates

The serum 25-hydroxyvitamin D level is a continuous variable, because it could theoretically take on an infinite number of values. How does linear regression change when modeling a binary covariate, such as sex? To use binary variables in regression models, one category of the variable is assigned the value of 1, and the other is assigned the value of 0. For example, to relate IL-6 levels to sex in a linear regression model we will assign men a value of 1 and women a value of 0 as shown in Fig. 18.5.

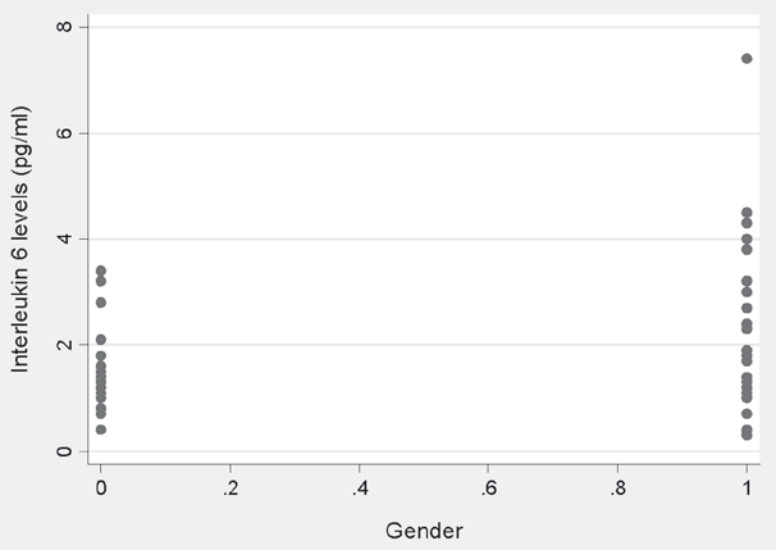


Fig. 18.5 Association between IL-6 levels and sex

Using sum of squares method, the best fitting regression equation is:

$$\text{IL-6} = 1.5 + 0.67 * (\text{sex});$$

where sex = 1 for men and 0 for women

How can we interpret the results from this model? For each 1-unit increase in the variable that identifies sex, IL-6 levels are 0.67 pg/ml higher, on average. We have defined the sex variable in such a way that a one-unit increase means comparing men to women. Another way to interpret these regression results is that IL-6 levels are, on average, 0.67 pg/ml higher in men, compared to women. We can directly examine the study data to check whether this finding is in fact true.

	Mean IL-6 level (pg/ml)
Men	2.17
Women	1.50
Difference in means	0.67

So, the regression equation exactly predicts the mean difference in IL-6 levels by sex. Note that because there were only two categories, we could have also performed this analysis using a *t*-test. The test of whether the slope is significant (whether 0.67 is statistically different from 0) is identical to the *t*-test for a difference in mean IL-6 levels between men and women. So for binary variables, linear regression and a *t*-test are equivalent analytic techniques. For categorical variables, linear regression analysis matches ANOVA analysis exactly as well, so one can think of both ANOVA and *t*-tests as special cases of linear regression.

18.4 Special Considerations

18.4.1 Influential Points

As discussed, the fitted regression line minimizes the sum of the squared vertical distances between all of the observed data points and the regression line. This process implies that not all of the data points will contribute equally when it comes to their ability to change the slope of the regression line. Consider the two scatter plots in Fig. 18.6, in which the question of interest is a possible association between body mass index (BMI) and the average number of weekly trips to the grocery store.

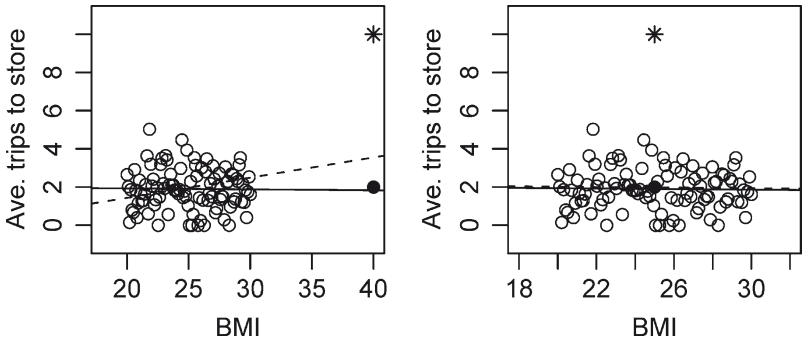


Fig. 18.6 Association of body mass index with trips to the grocery store

Both plots have the same data points, except for one extra point, shown as either a dark circle or an asterisk. In the left-hand plot, the additional data point, whether a dark circle or an asterisk, has great potential to influence the fitted regression line because its x-value lies far away from the majority of the other data points (this person has an extremely high BMI). If the y-value (number of trips to grocery store) for this data point is similar to the other data points, represented by the dark circle, then this outlying data point will have little influence on the fitted regression line, shown as the solid line. However, if this one individual has both a very high BMI *and* an unusually high number of trips to the grocery store, represented by the asterisk, then the slope of the fitted regression line will change dramatically, shown as the dashed line. In this case, the data point is said to have high *influence* on the regression line, because removing this single point from the dataset would change the slope of the regression line by a substantial amount.

In contrast, in the right-hand data of Fig. 18.6, the outlying point of interest is not likely to be influential, because its x-value (BMI) lies close to the other x-values in the dataset. Regardless of whether this person's number of trips to the grocery store is similar to the rest of the data (dark circle) or is unusually large (asterisk), the slope of the regression line will not change very much (dashed vs. solid lines). In the right-hand data plot, the asterisk point is simply called an "outlier" because it has a large residual compared to the other data points, but is not influential.

Single data points with high influence are concerning because statistical inference and point estimates can be driven by only a handful of data points, or in the above example, by a single data point. Generally in such cases, researchers will report the statistical inference using all of the data points, and then report the inference using the same methods but omitting a handful of data points that have high influence. Finding similar results using both methods implies that study findings are robust; dissimilar results suggest that more data are needed to answer the question of interest. For the left-hand data plot in Fig. 18.6, we would like to include more individuals with very high BMIs before drawing conclusions about the association of BMI with trips to the grocery store.

18.4.2 Nonlinear Associations

The fitted regression line represents the best possible *line* that can be fit to the observed data. The regression line will not accurately describe the data if they are not related as a line. For example, what if the IL-6 versus vitamin D scatter plot assumed the appearance of Fig. 18.7?

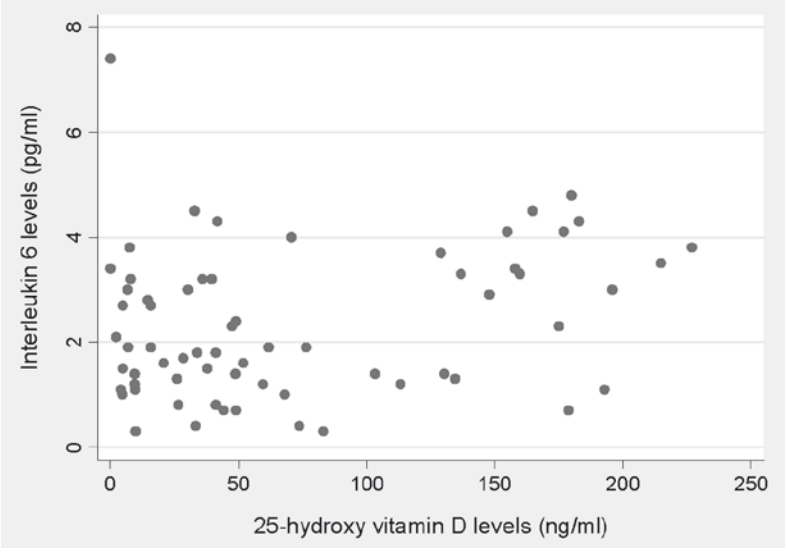


Fig. 18.7 Hypothetical nonlinear IL-6 versus vitamin D plot

In this example, IL-6 levels appear to have a more complex relationship with vitamin D levels than previously observed. These data suggest that IL-6 levels first tend to decrease as vitamin D levels increase, until vitamin D levels reach about 100

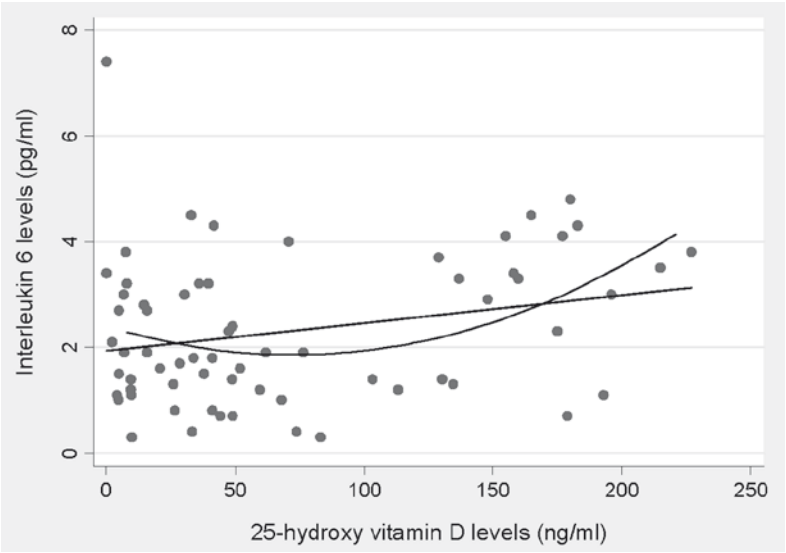


Fig. 18.8 Nonlinear relationship between IL6 and vitamin D levels

ng/ml. Then IL-6 levels appear to increase as vitamin D levels increase. Linear regression could still be used to describe these data. The fitted regression line would represent the best possible *line* that could be fit to the data; however, these data are not well described by *any line* in this case. A U-shaped curve appears to be a better fit, as described graphically in Fig. 18.8.

A clinical example of a nonlinear relationship between two factors is blood pressure and cerebral infarct size in the setting of acute stroke. Control of blood pressure is critical for stabilizing cerebral infarct size in patients with stroke, because inappropriately low blood pressures can extend cerebral infarction by exacerbating ischemia, and inappropriately high blood pressures can increase the risk of intracerebral bleeding. Consider a hypothetical study that examines the association of blood pressure with cerebral infarct size using data from patients in a neurologic intensive care unit. Using linear regression, the investigators report that, "each 10-mmHg higher systolic blood pressure is associated with a 5-mm^3 greater cerebral infarct size." At first glance this result suggests that systolic blood pressures should be kept as low as possible in patients with an acute stroke. However, more careful inspection of the study data in Fig. 18.9 reveals that greater cerebral infarct size is associated with *lower* blood pressures when the systolic blood pressure is below about 120 mm Hg and with *higher* blood pressures when the systolic blood pressure is greater than about 140 mm Hg. The ideal systolic blood pressure, in terms of the smallest infarct size, is between 120-140 mm Hg. Although a best-fit line can always be forced through the study data, it is important to consider whether the data actually demonstrate a linear association.

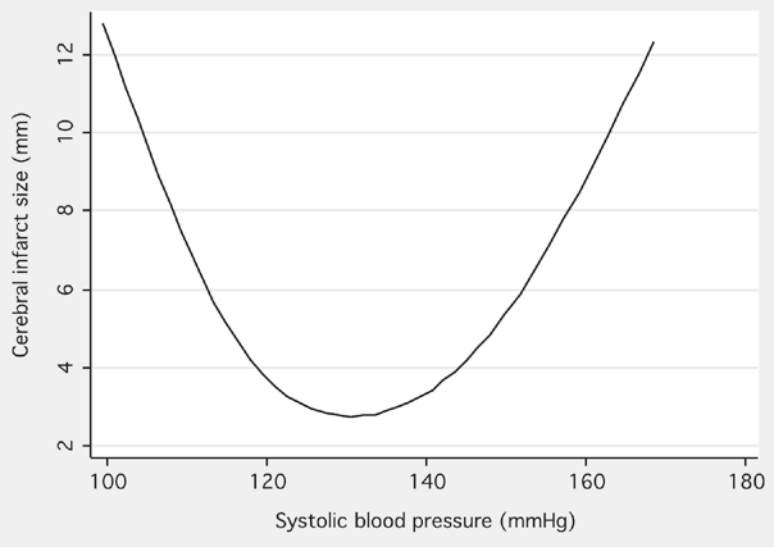


Fig. 18.9 Association of systolic blood pressure with cerebral infarct size in acute stroke

18.4.3 Extrapolating the Regression Equation Beyond the Study Data

Returning to the vitamin D IL-6 example, the equation IL6 = 2.4 − 0.01 * (25-hydroxy-vitamin D) was derived using data from patients whose vitamin D levels varied from about 5 to 200 ng/ml. There is no guarantee that this equation would be valid for vitamin D levels outside of this range. Extrapolating regression results beyond the range of observed study data frequently leads to erroneous findings and should be avoided. A classic example of this problem is the relationship between serum creatinine levels and kidney function. Studies in people with elevated serum creatinine levels, who have kidney disease, demonstrated a clear relationship between higher serum creatinine levels and decreased kidney function. However, the relationship between serum creatinine levels and kidney function is weak, or non-existent in people whose kidney function is normal. Inappropriately applying creatinine-based equations to predict kidney function in normal individuals from the general population can lead to faulty estimates.

18.5 Multiple Linear Regression

18.5.1 Definition of the Multivariate Model

Suppose that we are interested in evaluating how both serum vitamin D levels *and* sex are related to IL-6 levels? We originally defined the regression equation for one covariate to be:

$$Y = \beta_0 + \beta_1 * (\text{predictor})$$

We can now extend this regression equation for multiple predictor variables:

$$Y = \beta_0 + \beta_1 * (\text{predictor1}) + \beta_2 * (\text{predictor2}) + \beta_3 * (\text{predictor3}) \ldots$$

The following equation can be used to predict IL-6 levels from both vitamin D levels and sex:

$$\text{IL-6} = \beta_0 + \beta_1 * (\text{vitamin D}) + \beta_2 * (\text{sex}); \text{ where sex} = 1 \text{ for men and } 0 \text{ for women}$$

The resulting equation no longer represents a line, but rather a three-dimensional surface, because each IL-6 level predicted by the equation will be associated with both a specific vitamin D level and a specific sex, male or female. Multiple regression analysis is typically performed by computer. Specific values for IL-6, vitamin D, and sex for each study subject are entered into the computer, which obtains the best fitting regression equation. In other words, the computer selects the best possible values for the coefficients β_0, β_1, and β_2, such that the sum of squared residuals is as small as possible. The result is the best fitting linear equation to study the data:

$$\text{IL-6} = 1.97 - 0.01 * (\text{25-hydroxyvitamin D}) + 0.70 * (\text{sex})$$

The residuals from a multiple regression model are calculated in the same way as for a univariate regression model. A predicted IL-6 value can be calculated from the regression equation for each study subject based on his or her vitamin D level and sex. For example, a female subject who has a 25-hydroxy vitamin D level of 10 ng/ml would have the following predicted IL-6 value:

$$\begin{aligned} \text{IL-6} &= 1.97 - 0.01 * (\text{25-hydroxyvitamin D}) + 0.70 * (\text{sex}) \\ &= 1.97 - 0.01 * (10) + 0.7 * (0) = 1.87 \end{aligned}$$

The difference between this subject's actual measured IL-6 level and their predicted IL-6 value of 1.87 pg/ml represents the residual for this particular individual.

Analogous to the procedure for univariate linear regression, once a lest squares multiple regression model is fit, a series of statistical tests can then be employed in order to (1) evaluate how well the derived equation fits the study data, (2) check for possible influential values, and (3) investigate whether associations are linear. For the purposes of this example, we will assume that the fitted model is a reasonably good fit to the data and will proceed with interpretation.

18.5.2 Interpreting Results from the Multiple Regression Model

18.5.2.1 Obtaining Estimated Values for a Particular Set of Data

The fitted multiple regression model provides an estimated IL-6 level for any combination of vitamin D level and sex. For a hypothetical *male* subject with vitamin D level of 10 ng/ml:

$$\begin{aligned}\text{Mean IL6} &= 1.97 - 0.01 * (\text{vitamin D}) + 0.70 * (\text{sex}) \\ &= 1.97 - 0.01 * (10) + 0.70 \times (1) = 2.57 \text{ pg/ml}\end{aligned}$$

For a hypothetical *female* subject with a vitamin D level of 10 ng/ml:

$$\begin{aligned}\text{Mean IL6} &= 1.97 - 0.01 * (\text{vitamin D}) + 0.70 * (\text{sex}) \\ &= 1.97 - 0.01 * (10) + 0.70 * (0) = 1.87 \text{ pg/ml}\end{aligned}$$

18.5.2.2 Obtaining Relative Differences in the Outcome Variable by a Covariate

A clinical question regarding these data might be, "how much higher are IL-6 levels, on average, comparing men to women, independent of vitamin D status?" In other words, "what is the association between sex and IL-6 levels, holding vitamin D constant?"

According to the model, the average IL-6 level for *any* man in the study is

$$\text{Mean IL-6} = 1.97 - 0.01 * (\text{vitamin D}) + 0.70 * (1)$$

According to the model, the average IL-6 level for *any* two man in the study is

$$\text{Mean IL-6} = 1.97 - 0.01 * (\text{vitamin D}) + 0.70 * (0)$$

Therefore, the *difference* in IL-6 levels, comparing men to women, holding vitamin D levels constant is

$$\begin{aligned}&1.97 - 0.01 * (\text{vitamin D}) + 0.70 * (1) - 1.97 - 0.01 * (\text{vitamin D}) + 0.70 * (0) \\ &= 0.7 \text{ pg/ml}\end{aligned}$$

All of the terms in these equations will cancel out, leaving a difference of 0.70 pg/ml.

The interpretation of this result is that men have IL-6 levels that are, on average, 0.70 pg/ml higher than those of women, holding all of the other terms in the model constant.

Returning the original equation:

$$\text{IL-6} = 1.97 - 0.01 * (\text{vitamin D}) + 0.70 * (\text{sex})$$

The above example illustrates that the *individual coefficients from a multiple regression model represent the independent association between a covariate and the dependent variable, holding all other variables in the model constant.* This is an essential finding that is used to interpret regression equations that have more than one predictor variable. This example also demonstrates the use of multiple regression as a method to control for confounding.

For a second example, we will again predict IL-6 levels using vitamin D levels, and sex, but will now add age, smoking, and LDL cholesterol levels to the model. Again, sex will be coded as 0 for women, 1 for men, and smoking will be coded as 0 for a non-smoker, and 1 for a smoker. The general equation for this model is:

$$\begin{aligned}\text{Mean IL6} = \beta_0 &+ \beta_1 * (\text{vitamin D}) + \beta_2 * (\text{sex}) \\ &+ \beta_3 * (\text{age}) + \beta_4 * (\text{smoke}) + \beta_5 * (\text{LDL})\end{aligned}$$

Using a computer to minimize the sum of squares results returns a set of beta coefficients ($\beta_0 - \beta_5$) for a multiple regression model that most closely fits the study data:

$$\text{Mean IL6} = 1.1 - 0.01*(\text{vitamin D}) + 0.7*(\text{sex}) + 0.02*(\text{age}) + 0.4*(\text{smoke}) - 0.01*(\text{LDL})$$

This fitted equation can be used to describe estimated mean IL-6 levels for any combination of vitamin D level, LDL cholesterol level, sex, age, and smoking status by plugging specific values of these covariates into the model. Further, the independent association between each covariate and mean IL-6 levels is specified by their specific coefficients in the model. For example,

- For each one-year greater age, serum IL-6 levels are, on average, 0.02 pg/ml higher, holding vitamin D, sex, smoking, and LDL cholesterol levels constant.
- For each one mg/dl greater LDL cholesterol level, IL-6 levels are, on average, 0.01 pg/ml lower, holding vitamin D levels, sex, age, and smoking constant.
- Smokers have, on average, 0.4 pg/ml higher IL-6 levels compared to non-smokers, holding all of the other factors in the model constant.

18.5.2.3 Multiple Regression Results in Clinical Research Articles

Clinical research articles typically do not typically show the multiple regression models that were used to derive the study results. Instead, coefficients from these models are presented in table form. For example, a study reports the following predictors of kidney function, expressed as the glomerular filtration rate (GFR) in ml/min, among patients treated at a single clinic (Table 18. 2).

Table 18.2 Independent association of four covariates with kidney function

	Kidney function (GFR in ml/min)	95% confidence interval (GFR in ml/min)	*P*-value
Age (per year increase)	− 1.2	(−0.8, −1.6)	0.02
Non-steroidal use	− 12.5	(−10.0, −15.0)	0.01
Diabetes	− 10	(−7.0, −13.0)	0.01
Hypertension	− 3	(1.0, −7.0)	0.04

What model was used to obtain these results?

$$\text{GFR (ml/min)} = \beta_0 + \beta_1*(\text{age}) + \beta_2*(\text{non - steroidal use}) + \beta_3*(\text{diabetes}) + \beta_4*(\text{hypertension})$$

What coefficients were obtained for the fitted model?

$$\text{GFR (ml/min)} = 130 - 1.2*(\text{age}) - 12.5*(\text{non - steroidal use}) - 10*(\text{diabetes}) - 3*(\text{hypertension});$$

Where non-steroidal use, diabetes, and hypertension are coded as 1 if present, 0 if absent. There is no way to determine that the intercept term is 130 from table 18.2.

In this example, the coefficient for diabetes (β_3) is -10. What is the interpretation of this coefficient? Given two people who are the same age, have the same non-steroidal use status (yes versus no), and the same hypertension status (yes versus no), diabetes is independently associated with an estimated 10 ml/min lower GFR. Stated another way, "diabetes is associated with an estimated 10 ml/min lower GFR, independent of age, non-steroidal use, and hypertension," or, "diabetes is associated with an average 10 ml/min lower GFR, *after adjustment* for age, non-steroidal use, and hypertension.

18.5.2.4 *P*-values and Confidence Intervals for Regression Coefficients

For linear regression coefficients*, the null hypothesis is that the coefficient is equal to 0,* implying no association between covariate and the outcome in the population. For example, a coefficient for diabetes that was equal to 0 in the above example would indicate no association between diabetes and kidney function. The *p*-value of 0.01 for the diabetes coefficient in table 18.2 is interpreted as, "given no independent association diabetes with kidney function *in the population*, the chance of observing an independent association of −10 ml/min, or an even stronger association in this sample is 1%." The inference test result implies that diabetes *is* likely to be associated with differences in kidney function *in the population*.

The interpretation of the 95% confidence interval for diabetes is, "if the experiment were repeated an infinite number of times, and a 95% confidence interval placed around each experimental result, then 95% of the confidence intervals would contain the true adjusted association of diabetes with kidney function in the population." We don't know if our particular 95% confidence interval happens to be one of the 'good' confidence intervals, but since 95% of these intervals are 'good,' in that they contain the true population result within the interval, we can be "95% confident" that the adjusted association of diabetes with kidney function in the population is likely to be somewhere between −7.0 and −13.0 ml/min.

18.6 Confounding and Effect Modification in Regression Models

Although the scientific concepts of confounding and effect modification were covered in previous chapters, we pause here to note how they can be viewed within a regression framework.

18.6.1 Confounding

Suppose we were interested in describing the association between height and serum IL-6 levels? We sample 200 participants and fit the following simple linear regression equation to the data:

$$\text{Mean IL - 6} = -0.40 + 0.41 * (\text{height})$$

This unadjusted finding reveals that greater height is associated with higher IL-6 levels (coefficient for height is positive). However, a colleague points out that sex could be a potential confounder of the height-IL-6 association, because sex is likely to be associated with both height and IL-6 levels. Indeed, when sex is added to the multiple regression model, the following equation is obtained:

$$\text{Mean IL - 6} = 5.84 - 0.056 * (\text{height}) + 0.72 * (\text{sex});$$
$$\text{Where sex} = 1 \text{ for Men and 0 for Women}$$

The coefficient for height has changed dramatically, from +0.41 to −0.056, after adjustment for sex. Before adjustment, greater height is associated with *higher* IL-6 levels. Recall from Chapter 10 that one method used to identify a confounding factor is to examine whether an association of interest changes "substantially" after adjusting for the potential confounding factor. The mathematical equivalent to this process is to examine whether the coefficient of interest changes substantially after adjusting for a potential confounding factor, where adjustment is accomplished by including that confounding factor in the multiple regression model. What constitutes a substantial change for a regression coefficient? There is no universal agreement on this matter – some experts think 5%, others think 10% or more; therefore a subjective definition of "substantial change" is used. In this case, the coefficient changed from strongly positive to weakly negative, reversing the interpretation of the results and therefore indicating a substantial change. These results are shown graphically in the scatter plots in Fig. 18.10. The association of greater height with lower IL-6 levels is present in both men and women, further illustrating the confounding role of sex.

18.6.2 Effect Modification

Now suppose that the association of height with IL-6 levels was different for men and women. Our standard multiple regression model would not be able to detect this difference, because the model outputs only a single term, β_1, that represents the overall association of height with IL-6 levels for all people in the study (men and women combined).

$$\text{Mean IL-6} = \beta_0 + \beta_1 * (\text{height}) + \beta_2 * (\text{sex}); \text{ where sex} = 1 \text{ for men and 0 for women}$$

A different model is needed to explore whether the association of height with IL-6 levels might be different among subgroups within the study population, specifically men and women. In order to explore this potential interaction, we fit a more complex regression equation:

$$\text{Mean IL - 6} = -7.90 + 0.16 * (\text{height}) + 17.2 * (\text{sex}) - 0.25 * (\text{height} * \text{sex})$$

By design, this model has different interpretations for men and women.

For women:

$$\text{Mean IL -6} = -7.90 + 0.16 * (\text{height}) + 17.2 * (0) - 0.25 * (\text{height} * 0)$$
$$= -7.90 + 0.16 * (\text{height})$$

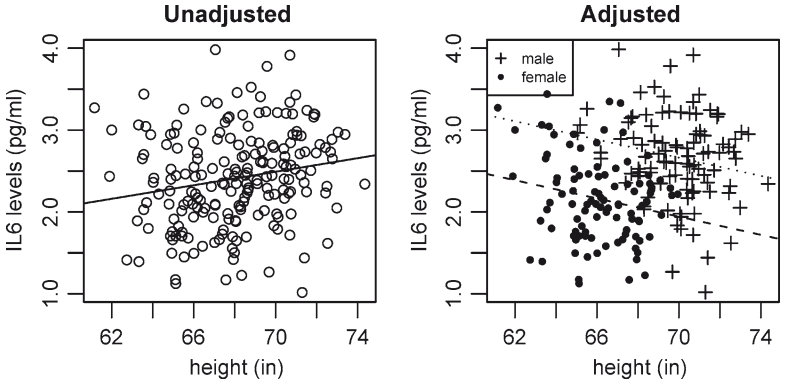


Fig. 18.10 Association of height with IL-6 level before and after adjustment for sex

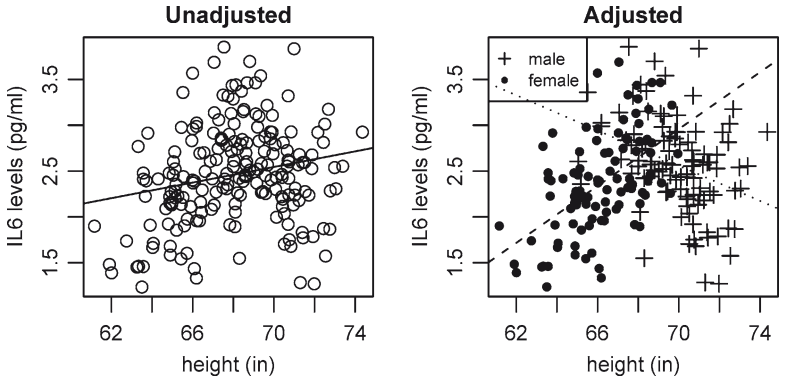


Fig. 18.11 Association of height with IL-6 level; effect modification by sex

Interpretation: for each one-unit increase in height, mean IL-6 levels are, on average, 0.16 pg/ml higher in women.

For men:

$$\begin{aligned}\text{Mean IL -6} &= -7.90 + 0.16 * (\text{height}) + 17.2 * (1) - 0.25 * (\text{height} * 1)\\ &= -7.90 + 0.16 * (\text{height}) - 0.25 * (\text{height}) + 17.2\\ &= -7.90 - 0.09 * (\text{height}) + 17.2\end{aligned}$$

Interpretation: for each one-unit increase in height, mean IL-6 levels are, on average, 0.09 pg/ml lower in men.

Adding the height * sex term to the model allows for distinction of separate height effects by sex. In this contrived example, the association of height with IL-6 levels

is modified by sex, meaning that the strength of association differs between men and women. The coefficient for the interaction term (height * sex) is the difference between the slopes relating height and IL-6 levels. The difference in slopes is demonstrated in Fig. 18.11.

Chapter 19
Non-Linear Regression

Learning Objectives

1. Regression with log-link is useful for studying the relative change in an outcome variable.
2. In a log-link regression model, the antilog of each coefficient represents the independent association of that covariate with the relative change in the outcome variable, holding all other variables constant.
3. Logistic regression is useful for studying associations for a binary outcome variable.
4. In a logistic regression model, the antilog of each coefficient represents the odds ratio of that covariate with the outcome variable, holding all other variables in the model constant.
5. In log-link and logistic regression models, the null hypothesis for a covariate is that the antilog of the coefficient for that covariate equals 1.0.

19.1 Regression for Ratios

In the previous chapter, we studied linear regression models, which assume the general form:

$$\text{Mean outcome} = \beta_0 + \beta_1 * (\text{predictor1}) * \beta_2 \times (\text{predictor2}) + \beta_3 * (\text{predictor3}) \ldots$$

By definition, linear regression models specify a linear relationship between the outcome variable and the predictor variables in a model. In linear regression, each one-unit difference in a predictor variable is linked with some constant difference in the outcome variable. For example, the linear regression model in the previous chapter linked each 1 ng/ml higher 25-hydroxyvitamin D level with a 0.01 pg/ml lower IL-6 level. In many instances, the assumption of a linear relationship between two factors is reasonable. However, there are certain circumstances in which nonlinear relationships might be expected.

For example, the HIV viral load, a measure of HIV disease severity, grows exponentially with time in untreated patients, such that each week the viral load

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_19, © Springer Science+Business Media, LLC 2009

may be 10% larger than it was the previous week. This growth pattern motivates evaluation of potential HIV risk factors in relation to the *relative change* (or percent change) in the HIV viral load. A type of regression model that can be used to study relative changes in an outcome variable is called *regression with log-link*, or *Poisson regression*. The log-link model assumes the following general form:

$$\log(\text{mean outcome}) = \beta_0 + \beta_1 \times (\text{predictor1}) + \beta_2 \times (\text{predictor2}) + \beta_3 \times (\text{predictor3})$$

This model is identical to the linear regression model, except for the addition of the log term on the left-hand side of the equation. Consider the following log-link regression model to investigate age, HIV subtype, and hepatitis C infection as potential risk factors for the severity of HIV infection, quantified by the HIV viral load:

$$\log(\text{Mean HIV viral load}) = \beta_0 + \beta_1 \times (\text{age}) + \beta_2 \times (\text{HIV subtype}) + \beta_3 \times (\text{hepatitisC})$$

where HIV subtype is coded as 1 for viral subtype D and 0 for all other subtypes; hepatitis C is coded as 1 if hepatitis C is present and 0 if absent.

Using a computer to solve the above model for $\beta_0 - \beta_3$ yields the following fitted equation:

$$\log(\text{mean HIV viral load}) = 9.4 + 0.08 * (\text{age}) + 0.5 * (\text{HIV subtype}) + 0.3 * (\text{hepatitis C})$$

We now focus on a specific question of interest, "what is the independent association of hepatitis C infection with the HIV viral load, holding age and HIV subtype constant?"

From the model, the log (mean HIV viral load) for *any hepatitis C positive person* in the study is

$$\log(\text{mean HIV viral load}) = 9.4 + 0.08 * (\text{age}) + 0.5 * (\text{HIV subtype}) + 0.3 * (1)$$

The log (mean HIV viral load) for *any hepatitis C negative person* in the study is

$$\log(\text{mean HIV viral load}) = 9.4 + 0.08 * (\text{age}) + 0.5 * (\text{HIV subtype}) + 0.3 * (0)$$

Therefore, the *difference* in the log (mean HIV viral load), comparing any hepatitis C positive person to any hepatitis C negative person in the study is

$$\log(\text{mean HIV viral load})_{\text{hepatitis C positive}} - \log(\text{mean HIV viral load})_{\text{hepatitis C negative}} = \{9.4 + 0.08 * (\text{age}) + 0.5 * (\text{HIV subtype}) + 0.3 * (1)\} - \{9.4\} + 0.08 * (\text{age}) + 0.5 * (\text{HIV subtype}) + 0.3 * (0)\} = 0.3$$

Analogous to the linear regression model, the coefficient for any predictor variable in a log-link model represents the difference in the log (mean outcome) associated with each one-unit change in the predictor variable.

To clarify interpretation of results from the log-link model, a mathematical property of logarithms is needed. For any two numerical values, a and b, it can be shown that: $\log a - \log b = \log (a/b)$. Therefore, results from our log-link model above can be rewritten as:

$$\log\left\{\frac{\text{mean HIV viral load}_{\text{hepatitis C positive}}}{\text{mean HIV viral load}_{\text{hepatitis C negative}}}\right\} = 0.3$$

Now we can take the antilog, or exponent, of both sides of the equation:

$$\exp\left\{\log\frac{\text{mean HIV viral load}_{\text{hepatitis C positive}}}{\text{mean HIV viral load}_{\text{hepatitis C negative}}}\right\} = \exp\{0.3\}$$

Since the natural log (ln) is typically used, the antilog is base e, so exp (ln (x)) = x. Therefore,

$$\frac{\text{mean HIV viral load}_{\text{hepatitis C positive}}}{\text{mean HIV viral load}_{\text{hepatitis C negative}}} = 1.35$$

The interpretation of this result is that the HIV viral load is, on average, *35% higher* for subjects who have hepatitis C compared to those who do not have hepatitis C, holding age and HIV status constant. The interpretation of any coefficient from a log-link model is *the proportionate (relative) change in the outcome variable associated with each one-unit change in the predictor variable*.

Recall that for linear regression models, the null hypothesis for a particular factor in the model is that the coefficient for that factor equals 0, indicating no association with the mean value of the outcome. For log-link regression, the null hypothesis for a particular factor is that the *antilog of the coefficient equals 1.0*, indicating a ratio of 1.0, or no association between that factor and the relative change in the outcome variable.

19.2 Logistic Regression

In both linear and log-link regression models, the dependent (outcome) variable on the left-hand side of the equation is continuous, meaning that it can take on an infinite number of possible values. Continuous outcomes are widespread in clinical/epidemiological research; examples include blood pressure, levels of serologic markers, myocardial ejection fraction, and the number of missed work days due to headache.

Other types of clinical research questions focus on evaluating the *risk of a binary outcome variable*, which can take on only two possible values. Examples include myocardial infarction, successful alcohol rehabilitation, and hospitalization for pneumonia. A general regression equation can be constructed to evaluate the probability of a binary outcome variable:

$$\text{Probability outcome} = \beta_0 + \beta_1\ \text{x(predictor1)} + \beta_2\,\text{x (predictor2)} + \beta_3 \times \text{(predictor 3)}$$

However, a problem with this model is that probabilities must always be between 0 and 100%, constraining possible values for the beta coefficients. A common

method used to overcome this problem is to model the *odds of the outcome* rather than the probability. Odds are mathematically related to probabilities but are not restricted by boundaries of 0 and 100% simplifying their use in regression models. The specific definition of odds is

$$\text{Odds} = p/(1-p)$$

where p represents the probability of a given outcome

While mathematically accommodating, odds are not intuitive. Interpretation of odds is facilitated by the property that *odds closely approximate probability when the outcome of interest is rare*. A rare disease signifies that its probability, *p*, is small, resulting in a denominator for odds that is close to 1.0.

$$\text{Odds} = p/(1\ p) = p \quad \text{if } p \text{ is small}$$

For example, if a rare event occurs 2% of the time, then the odds of this event are close to 2%:

$$\text{Odds} = p/(1-p) = 0.02/(1-0.02) = 0.02/0.98 = 0.0204$$

However, this is not true for common events. The odds of an event that occurs 20% of the time is

$$\text{Odds} = p/(1-p) = 0.20/(1-0.20) = 0.2/0.8 = 0.25$$

Replacing probability with odds in the regression model results in a new model called *logistic regression*, which is commonly used to study binary outcomes in clinical research. Logistic regression models assume the following form:

$$\log(\text{odds outcome}) = \beta_0 + \beta_1 \times (\text{predictor1}) + \beta_2 \times (\text{predictor2}) + \beta_3 \times (\text{predictor3})$$

This model shares similarities to the log-link model above, so similar interpretations apply.

Consider a fitted logistic regression model for a study that evaluates predictors of pneumonia:

$$\log \text{odds pneumonia} = \beta_0 + 0.1 \times (\text{age}) + 0.3 \times (\text{sex}) + 0.9 \times (\text{asthma})$$

What is the association of asthma with pneumonia, adjusting for age and gender?

For any person in the study with asthma:

$$\log \text{odds pneumonia} = \beta_0 + 0.1 \times (\text{age}) + 0.3 \times (\text{sex}) + 0.9 \times (1)$$

For any person in the study without asthma:

$$\log \text{odds pneumonia} = \beta_0 + 0.1 \times (\text{age}) + 0.3 \times (\text{sex}) + 0.9 \times (0)$$

Therefore,

$$\log\left\{\frac{\text{odds pneumonia}_{\text{asthma}}}{\text{odds pneumonia}_{\text{no asthma}}}\right\} = 0.9$$

Taking the antilog of both sides of the equation (using natural logs) yields:

$$\frac{\text{odds pneumonia}_{\text{asthma}}}{\text{odds pneumonia}_{\text{no asthma}}} = 2.5$$

This result can be interpreted as: "asthma is associated with a 2.5-fold greater relative odds of pneumonia after adjustment for age and sex," or, "a person with asthma has, on average, a 2.5-fold greater odds of pneumonia compared to a person without asthma, after adjustment for age and sex." In logistic regression, the antilog of each coefficient in the regression model represents the independent association between that factor and the *odds ratio* of the outcome variable.

If pneumonia is relatively uncommon in this population, then we can replace the term *odds* in these interpretations with the term *risk* or *probability*, yielding a sentence that is much more interpretable. On the other hand, if pneumonia *is* common, then interpretation of the logistic regression results is less clear. In general, odds ratios tend to exaggerate associations of the relative risk, with the extent of exaggeration being proportionate to the prevalence of the condition. Consider the discrepancy between odds ratio and relative risk when the risk of pneumonia is 30%.
Odds ratio of pneumonia, comparing asthma to no asthma = 2.5
Relative risk of pneumonia, comparing asthma to no asthma = 1.9

Similar to log-link regression, the null hypothesis for a particular factor in a logistic regression model is that the *antilog of the coefficient equals 1.0*, indicating an odds ratio of 1.0, or no association between that factor and the outcome variable.

19.3 Application of Logistic Regression Models

The results in Table 19.1 were obtained from a study of risk factors for peptic ulcer disease.

Table 19.1 Risk factors for peptic ulcer disease

	Adjusted odds ratio	95% confidence interval	*P*-value
Age (per decade)	1.15	(1.10, 1.30)	0.001
Female	0.85	(0.68, 1.05)	0.143
Smoking	1.40	(1.20, 1.75)	0.040
Spicy food consumption	2.25	(0.90, 3.90)	0.060
Income (per $100,000 increase)	1.03	(1.02, 1.04)	0.001
H. pylori	1.70	(1.50, 2.80)	0.020

Implicit in the table is that female is being compared to male, smoking compared to non-smoking, spicy food consumption compared to no spicy food consumption, and H. Pylori positive status compared to H. Pylori negative status.

What type of regression model was used in this analysis?

Logistic regression was used, because the outcome being predicted is binary (peptic ulcer disease), and because adjusted odds ratios are presented in the table.

What is the functional form of the regression model?

$$\text{log odds peptic ulcer disease} = \beta_0 + \beta_1 * \text{age} + \beta_2 * \text{gender} + \beta_3 * \text{smoke} + \beta_4 * \text{spicy food} + \beta_5 * \text{income} + \beta_6 * \text{H. pylori}$$

What is the coefficient for *H. pylori*?
Each adjusted odds ratio in Table 19.1 was obtained by taking the exponent of that coefficient from the logistic regression model. Therefore, taking natural log of each adjusted odds ratio in the table yields the model coefficient. The coefficient for H. Pylori is:

$$\beta_6 = \ln(1.70) = 0.53$$

What is the interpretation of the *H. pylori* result in the table?

"Positive *H. pylori* status is associated with a 1.7-fold greater odds of peptic ulcer disease, independent of age, smoking, spicy food consumption, and income," or stated another way, "The odds ratio of peptic ulcer disease, comparing *H. pylori* positive to *H. pylori* negative subjects, is 1.7, after adjustment for age, sex, smoking, spicy food consumption, and income.

What is the interpretation of the *p*-value for *H. pylori* in the table?

If there is no association of *H. pylori* status with peptic ulcer disease in the population, then the chance of observing an adjusted odds ratio of 1.7, or a more extreme adjusted odds ratio in this sample is 2%. This p-value implies that there is an association of H. Pylori with peptic ulcer disease in the population.

Chapter 20
Survival Analysis

Learning Objectives

1. Incidence measures do not fully describe the development of events over time.
2. The survival function, $S(t)$ represents the probability of being alive at a particular time, t.
3. For a graphical presentation of the survival function:
   a. Survival for any particular follow-up time is estimated by a vertical line to $S(t)$.
   b. Median survival is estimated by a horizontal line from $S(t) = 0.5$.
4. The logrank test evaluates whether whole survival curves are statistically different from each other.
5. Survival analysis is typically used to describe the *first* occurrence of a particular outcome.
6. Censoring occurs when a subject leaves a study before incurring the outcome of interest.
7. The Kaplan–Meier method is used to estimate $S(t)$ in the presence of censoring.
8. Kaplan–Meier plots are typically unadjusted.
9. The Cox model can adjust for confounding and account for censoring.
10. The Cox model yields a hazard ratio, which very closely parallels the relative risk.
11. Hazard ratios are meaningful for studies in which the relative risk remains constant throughout the study period; studies with changing risks over time should present separate hazard ratios for the relevant time periods of interest.

20.1 Limitations of Incidence Measures for Evaluating Risk

We begin by examining a new clinical question: what is the optimal treatment strategy for patients who have a small abdominal aortic aneurysm (AAA)? The United Kingdom Small Aneurysm Trial compared elective surgery with medical surveillance for small AAAs.[44] A total of 1,090 patients with aneurysms smaller than 5.5 cm were randomly assigned to receive either elective surgery or surveillance. After 8 years of follow-up, there were 242 deaths among 563 patients assigned to surgery, and 254 deaths among 527 patients assigned to medical surveillance.

B. Kestenbaum, *Epidemiology and Biostatistics: An Introduction to Clinical Research*,
DOI 10.1007/978-0-387-88433-2_20, © Springer Science+Business Media, LLC 2009

20.1.1 Incidence Measures: Oversimplification of Study Results Over time

Previously, the terms *incidence proportion* and *incidence density* were used to describe risk during follow-up. For the above example,

	Incidence proportion
Surgery group	242 deaths/563 people = 43.0%
Surveillance group	254 deaths/527 people = 48.2%

To obtain incidence rates, risk-time information is needed. Given risk-times of 3,660 person-years in the surgery group, and 3,820 person-years in the surveillance group:

	Incidence rate
Surgery group	242 deaths/3,660 person-years = 6.6 deaths per 100 person-years
Surveillance group	254 deaths/3,820 person-years = 6.6 deaths per 100 person-years

These incidence data provide a compact summary of the clinical experience of 1,090 people with AAA in the study; however, they do not address several important clinical questions. For example, a patient with a newly diagnosed 5.0 cm AAA asks the following questions:

(1) "If I decide not to have surgery, what is the chance that I will die in the next 5 years?"
(2) "If I decide to have surgery, about how long will I live?"
(3) "I'm pretty healthy and think I can survive the surgery. If I make it through the operation, will I do better than if I just watched and waited?"

These and similar questions are not addressed by incidence measures. Survival analysis provides a more detailed description of what actually happens in a study over time.

20.1.2 Incidence Measures: Crude Handling of Participant Dropout

Incidence measures may also yield biased relative risks in the setting of certain patterns of participant dropout. Consider a hypothetical problem in which subjects assigned to the AAA surgical group decide to drop out of the study after having successful surgery. Premature dropout in the surgical group might occur because study participants no longer perceive any direct benefit to themselves from contin-

Table 20. 1 Premature dropout specific to the surgical group

	Follow-up time	Reason for disenrollment
Patient 1	24 days	Requests to leave the study
Patient 2	8 months	Requests to leave the study
Patient 3	14 months	Lost to follow-up
Patient 4	6 months	Requests to leave the study
Patient 5	17 days	Requests to leave the study
Patient 6	4 months	Lost to follow-up
Patient 7	11 months	Lost to follow-up

ued participation. Table 20.1 describes data from seven sample patients in the surgical group who drop out prematurely.

Given the high-risk nature of AAA surgery, it is reasonable to expect that there will be some post-operative mortality, followed by a lower long-term risk of death due to potential benefits of the surgery. Preferential dropout of participants after they have successful surgery will delete the late post-operative risk time, when beneficial effects of surgery may be most pronounced. Removing this "quality risk time" from the surgical group will result a higher observed death rate, and will distort the comparison with the medicine group.

20.2 Survival Data

Another approach for describing the AAA study data is to examine mortality risk continually throughout follow-up. This approach can be accomplished with the *survivor function*, which we will denote as S(t). The survivor function is a function fit to the study data that returns the probability of being alive at a particular time, t. For example, say the equation for $S(t)$ is: $S(t) = 1/(t + 1)$ for the surgical group, where t represents years of follow-up. Then, the chance of surviving for 1 year in the surgical group would be $1/(1 + 1) = 0.5$, or 50%. The chance of surviving for 2 years would be $1/(1 + 2) = 0.33$, or 33%. We will explore how to obtain $S(t)$ shortly. For now, imagine that it has been handed to you for interpretation.

Since $S(t)$ returns the chance of being alive at any time during the study, it is convenient to look at $S(t)$ graphically as a function of follow-up time. Figure 20.1 depicts hypothetical survival data for the surgical treatment group.

Compared to the incidence density data (6.6 deaths per 100 person-years), $S(t)$ provides much more detailed information about patient survival over time. To estimate survival for any particular follow-up time, simply draw a *vertical line* from the time point of interest to the survivor function. For example, survival at 1 year is about 75%; survival at 2 years is about 60%. These data answer clinical questions such as, "If I have surgery, what is the chance that I will be alive after two years?"

The *median survival* is defined as the time point during follow-up in which 50% of the study population has died. Median survival is estimated by drawing a

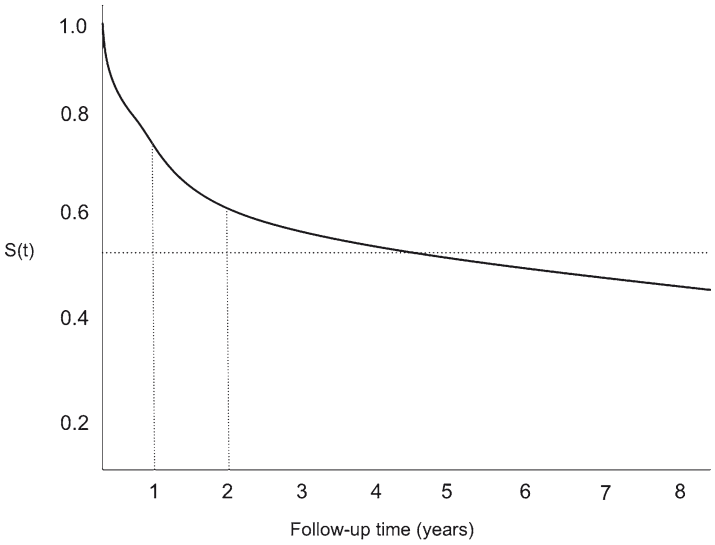
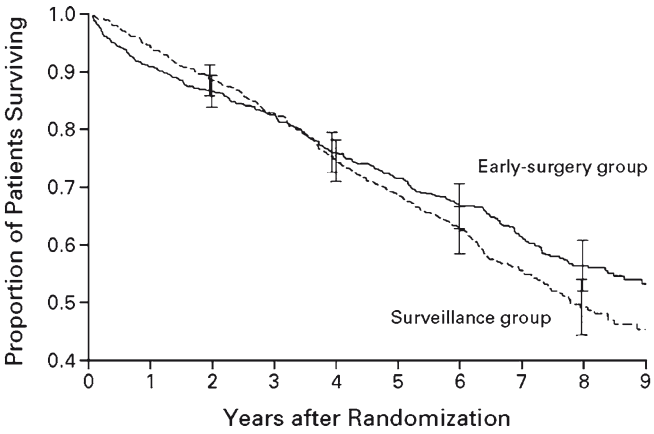


Fig. 20.1 Hypothetical plot of the survivor function for the surgical treatment group



No. at Risk

Surveillance group	527	497	468	437	394	363	316	173	97	41
Early-surgery group	563	513	489	465	429	402	371	253	154	66

Fig. 20.2 Survival in United Kingdom Small Aneurysm Trial

horizontal line to the survivor function when $S(t) = 0.5$. In this case, median survival is about 4.5 years. Notice that median survival can be estimated only from studies in which at least 50% of the study population incurs the outcome of interest. Because $S(t)$ is typically estimated using a method developed by Kaplan and Meier, survivor function plots are often called *Kaplan–Meier plots*.

Survival data for the nonsurgical group may be overlaid with that of the surgical group to compare relative survival differences over time as shown in Fig. 20. 2.

In this case, survival is higher in the surveillance group (dashed line) until about three years, after which time survival is higher in the surgical group (solid line).

20.3 Statistical Testing of Survival Data

Examining the superimposed survival curves for the AAA study, two relevant questions might be, "are the two survival curves significantly different from each other?" and "is survival at some particular time point significantly different between the two treatment groups?"

The first question is usually addressed by a statistical test called the *logrank test*, although other tests have been developed for this purpose. The logrank test examines *entire survival curves* for each treatment group. The *p*-value from the logrank test represents the probability of obtaining the observed survival curves, or more extremely different survival curves, if in fact survival curves are identical in the population. In the case of the AAA study, the *p*-value for the logrank test was found to be 0.05. The interpretation of this *p*-value is: if survival curves for surgery and surveillance groups were identical in the population, then the chance of observing these or more extremely different survival curves in this sample is 5%.

However, the logrank test is not particularly useful for these AAA survival data because the two survival curves cross. Results from the logrank test tell us only that the two survival curves are statistically different from each other. The more complete story is that survival is first better in the surveillance group, and then is better in the surgery group. The logrank test is most useful for studies in which survival curves remain separated throughout follow-up. For these studies, a significant logrank test result implies that survival in one group is statistically better or worse than survival in the other group.

Examining long-term survival from the AAA study, survival at 8 years is 0.57 in the surgical group, and 0.50 in the surveillance group. How do we test whether that 7% survival difference is statistically significant? Recall from chapter 17 that the chi-square test was used to test proportions from two different treatment groups. We can test whether 57% survival is statistically different from 50% survival at 8 years using methods analogous to (though different from) the chi-square test.

Comparison of 8-year survival between the surgical and surveillance arms in the AAA study yields a *p*-value of 0.05. The interpretation of this *p*-value is, "if there is truly no difference in eight-year survival, comparing surgery to surveillance, among the entire population of AAA patients, then the chance of observing this 7% survival difference at eight years, or a more extreme survival difference, in this particular sample is 5%." These results of inference testing imply that 8-year survival *is* likely to be different among the population of AAA patients.

Limitations of statistical testing of the survivor function are analogous to concerns about statistical testing discussed previously. First, multiple comparisons problems can occur if many survival times are tested. Beware of studies that statisti-

cally test 1-year survival, 2-year survival, 3-year survival, and so on, because a "significant" *p*-value is bound to turn up somewhere. The Bonferroni correction, or other correction for multiple comparisons, can be used to address this problem by setting a more stringent threshold to define a significant *p*-value. Second, patient attrition during follow-up markedly diminishes statistical power to detect long-term survival differences. For the AAA study, only 97 participants remained in the surveillance group, and only 154 participants remained in the surgical group after 8 years.

20.4 Definitions of Events and Censoring

Survival analysis is used to examine the risk of a *binary outcome variable*, also called an *event*, or *failure*. For the AAA study, the outcome of interest was chosen to be mortality (yes *versus* no); however, the outcome could have also been chosen to be ruptured AAA (yes *versus* no), hospitalization for any cause (yes *versus* no), or acute stroke (yes *versus* no).

In most instances, survival analysis is used to describe the *first occurrence* of a particular outcome of interest during follow-up. Subjects who incur the outcome of interest are considered to have completed the study at that time. Once the outcome of interest occurs, study subjects are typically no longer followed, and stop contributing risk time. More advanced survival analysis techniques are available for studying multiple events per study subject; however, these methods are beyond the scope of this book.

The survivor function, $S(t)$, is specifically defined as the cumulative probability of survival without incurring the outcome of interest for time, t. For a hypothetical study examining the risk of a first hospitalization for heart failure, $S(t)$ would represent the cumulative probability of survival without hospitalization for heart failure. $S(t)$ can therefore be interpreted as the *probability of event-free survival.*

The most accurate estimate of $S(t)$ is obtained when the study population is followed until every subject incurs the outcome of interest. However, in clinical studies, subjects may leave the study before incurring the outcome because of (1) dropout, (2) loss to follow-up, or (3) the study ends at some pre-determined time point. Further, for studies in which the outcome of interest is not death, subjects will also leave the study when they die. *Censoring is defined as when a subject leaves the study for any reason before incurring the outcome of interest.* Censoring results in having some information about a person's survival time, but not knowing their survival time exactly.

In survival analysis, subjects are followed from some specified start date until the first occurrence of the study outcome or their data are censored, whichever comes first. Clinical research studies that use survival data should clearly describe exactly when study subjects started and stopped accruing risk time and the specific reasons for censoring. Here is an example of precise language used to describe survival data for the hypothetical heart failure.

"The outcome of interest was a first hospitalization for heart failure. Subjects were followed from study enrollment in 1998 until the first occurrence of a heart failure hospitalization or the data were censored due to death, lost to follow-up, or the study ended on July 31, 2001."

20.5 Kaplan–Meier Estimation

20.5.1 Kaplan–Meier Estimation of S(t)

The survivor graphs for the AAA example assumed that we knew $S(t)$. In practice $S(t)$ is *estimated* using the study data. The most direct way to estimate $S(t)$ from the AAA study data would be to check on the survival of the study population as often as possible. For example, we could update survival status every day for first 10 subjects in the AAA study (Table 20. 2).

$S(t)$ represents the probability of survival after the specified follow-up time has elapsed. For example, $S(6 \text{ days}) = 0.6$, indicating that there is a 60% chance of survival after 6 days of follow-up. It is important to note that *S(t) changes only when an event occurs*. Intervals of follow-up time without events do not add any new information to $S(t)$. Therefore, we update $S(t)$ only when new failures occur and present the survival data in a more compact form, as shown in Table 20. 3.

These data demonstrate that $S(t)$ is not a smooth, continuous curve, but rather is a step function, with the "steps" occurring at the failure times. Figure 20.3 plots $S(t)$ for the above survival data.

Table 20. 2 Survival status for 10 subjects in the small aneurysm trial

Follow-up time (days)	Number of people at risk	Events (deaths)	$S(t)$
1	10	0	1.0
2	10	1	0.9
3	9	1	0.8
4	8	0	0.8
5	8	0	0.8
6	8	2	0.6
7	6	0	0.6
8	6	2	0.4

Table 20. 3 Compact presentation of survival for 10 subjects in the small aneurysm trial

Follow-up time (days)	Number of people at risk	Events (deaths)	$S(t)$
2	10	1	0.9
3	9	1	0.8
6	8	2	0.6
8	6	2	0.4

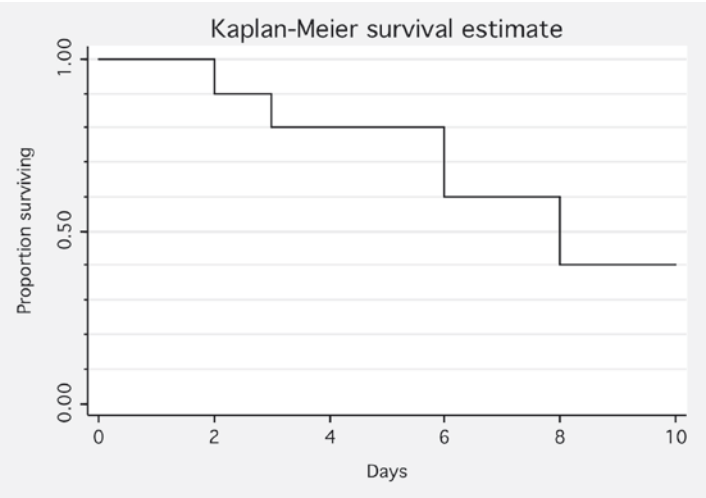


Fig. 20.3 Kaplan–Meier survival estimate for 10 subjects in the small aneurysm trial

20.5.2 Kaplan–Meier Estimation of S(t) with Censored Data

The above example does not consider people who dropout of the study. In real-life situations, $S(t)$ must be estimated in the presence of censoring. Table 20. 4 adds censoring to the original survival data.

Dropout at any time interval will decrease the number of subjects at risk at the start of the next time interval. Importantly, $S(t)$ can change only when a failure occurs; therefore, $S(t)$ for day 5 will remain 0.8, despite the two dropouts that occur on that day. Calculation of $S(t)$ for day 6 is not as straightforward. We can't really say that survival at day 6 is 0.4, because that would assume that the two subjects who dropped out of the study would have died had they remained in the study. We also can't say that survival at day 6 is 0.6, because that would assume that the two dropouts would still be alive had they remained in the study. In reality, we do not know exactly what would have happened to these two subjects, because we no longer are following them.

The solution to this problem is to consider $S(t)$ as the product of two probabilities*: the probability of survival until time t and the probability of survival past time t.* For example, the probability of surviving until study day 6 is 0.8. The probability of surviving past study day 6, given survival until study day 6, is 4/6, because 6 subjects are alive and at risk beginning on day 6, and only 4 of these subjects survive past day 6. Therefore $S(t)$ for day 6 can be calculated as $0.8 \times (4/6) = 0.53$. This probability is then carried forward until the next failure time, because of the important property that $S(t)$ remains constant between failures. The next failure occurs on study day 8. The probability of surviving past day 8, given survival until study day 8, is 1/3, because there are 3 subjects alive at the beginning of day 8, and

Table 20. 4 Survival status for 10 subjects in the small aneurysm trial with censoring

Follow-up time (days)	Subjects at risk	Events (deaths)	Dropouts	$S(t)$
1	10	0	0	1.0
2	10	1	0	0.9
3	9	1	0	0.8
4	8	0	0	0.8
5	8	0	2	0.8
6	6	2	1	?
7	3	0	0	?
8	3	2	0	?

Table 20. 5 Calculation of $S(t)$ in the presence of censoring

Follow-up time (days)	Subjects at risk	Events (deaths)	Dropouts	$S(t)$
1	10	0	0	1.0
2	10	1	0	0.9
3	9	1	0	0.8
4	8	0	0	0.8
5	8	0	2	0.8
6	6	2	1	$0.8 \times (4/6) = 0.53$
7	3	0	0	0.53
8	3	2	0	$0.53 \times (1/3) = 0.18$

1 of them survives past day 8. Therefore, the cumulative probability of surviving past day 8 is $0.53 \times (1/3) = 0.18$ (Table 20. 5).

This method to estimate the survival function in the presence of censoring is called the *Kaplan–Meier method*. Because this method estimates the survivor function using the product of multiple probabilities, it is also called the product-limit method. The Kaplan–Meier method allows subjects to contribute information to $S(t)$ as long as they remain in the study, and then stop contributing information when they are censored. For the two study subjects who dropped out on day 5, the Kaplan-Meier method uses their data for the first 5 days and then removes them from the risk set, reducing the number of subjects who are at risk for the next calculation of $S(t)$.

This method assumes that censored individuals are similar to those who remain in the study, an assumption known as "non-informative" censoring. If this assumption is unreasonable, for example if study subjects are censored because they are admitted to a hospital for another cause, then estimates of $S(t)$ may be biased.

Important specific points to remember about the Kaplan–Meier estimate of $S(t)$ are

(1) $S(t)$ changes only when events occur. $S(t)$ does not change when subjects drop out of a study.
(2) $S(t)$ is calculated as the probability of surviving past time t, given survival until time t.

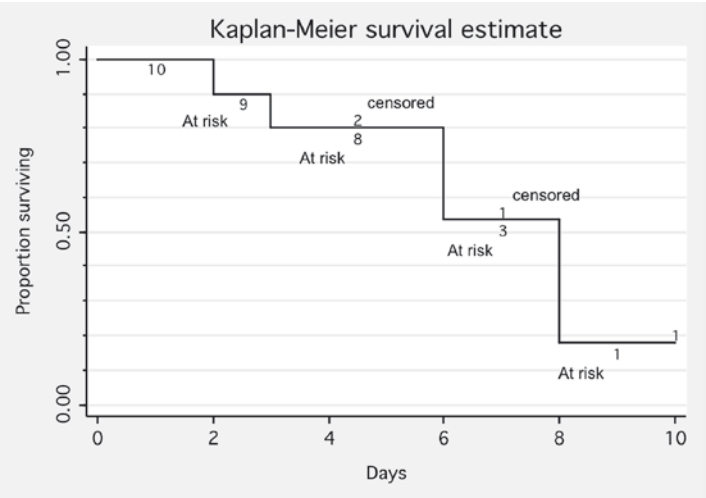


Fig. 20.4 Kaplan–Meier survival estimate with censoring

(3) Censored patients contribute information to $S(t)$ only while they remain in the study.
(4) Censoring refers to removing subjects from the risk set for any reason other than incurring the study outcome

The Kaplan–Meier plot for the censored data is shown in Fig. 20. 4.

20.6 Cox's Proportional Hazards Model

20.6.1 Description of the Proportional Hazards Model

Plots of the Kaplan–Meier estimated survivor function reflect observed survival differences for different groups of subjects with respect to follow-up time. It is important to note that Kaplan–Meier plots represent crude or unadjusted study data. These plots are not typically adjusted for potential differences that might exist between treatment groups, such as differences in age, health status, and co-morbidity.

Unadjusted survival plots are easy to interpret for large randomized clinical trials, when participant characteristics are generally balanced between the treatment groups. However, survival plots may be less meaningful in observational studies. What if the AAA study was not a randomized trial, but instead was an observational comparison of surgical versus medical treatment for AAA? In this situation, AAA patients who undergo surgery may be younger and healthier than those who do not,

such that potential associations of surgery with improved survival could be confounded. The Kaplan-Meier plot for these observational study data would show the observed unadjusted data, which may be confounded. How can we deal with potential confounding in observational studies of survival?

One approach to observational AAA data is to construct and test separate survival curves within categories of age and health status. This approach would hold age and health status roughly constant within each category. Analogous to the method of stratification, this approach yields multiple survival curves for comparison, and may become unwieldy if many confounding variables are considered. What is needed is a multiple regression model that deals with survival data.

What about logistic regression? Recall that logistic regression models the odds ratio of a binary outcome variable as a function of predictor covariates:

$$\text{Log odds mortality} = \beta_0 + \beta_1 \times (\text{treatment}) + \beta_2 \times (\text{age}) + \beta_3 \times (\text{health status}) \ldots$$

Here, *treatment* would be coded as 1 for surgery, 0 for surveillance. After fitting the model, β_1 would represent the relative odds of mortality, comparing surgery to surveillance, holding all of the other terms in the model constant. The problem is that follow-up time is *not* included in the logistic model. If there are differences in follow-up time between the surgical and surveillance groups, erroneous findings may result. We could add follow-up time to the model:

$$\begin{aligned}\text{Log odds mortality} = {} & \beta_0 + \beta_1 \times (\text{treatment}) + \beta_2 \times (\text{age}) + \beta_3 \times (\text{health status}) \\ & + \beta_4 \times (\text{time}) \ldots\end{aligned}$$

In this model, β_1 represents the odds ratio of death, comparing surgery to surveillance, holding all of the variables in the model constant *including follow-up time*. However, this is not the correct solution to the problem. The exact nature of the relationship between the log odds of mortality and follow-up time would have to be precisely known to truly hold follow-up time constant (the above example assumes a linear relationship). Further, we have seen that certain patterns of dropout during follow-up can lead to bias if the approach to follow-up time is to simply count risk time for each treatment group.

Cox first published a solution to the problem in 1973, and he was later knighted for it. The proportional hazards model remains the most widely used model for survival data in clinical research, because it allows for simultaneous adjustment of multiple confounding factors, and accurately handles differences in follow-up time and censoring between treatment groups. The following is an oversimplified description of how the proportional hazards model works.

As we observed in the Kaplan–Meier estimation method, information about survival is obtained only when failures occur. The Cox model essentially performs logistic-like regression *repeatedly at every failure time* in the study. This procedure assures that follow-up time will be identical for all study subjects at the time of each comparison. A simplified diagram of the proportion hazards model approach to the AAA study appears in Fig. 20. 5 below.

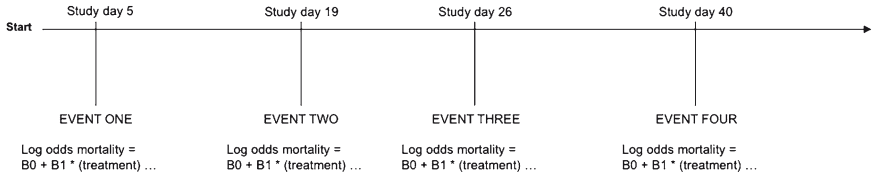


Fig. 20.5 Simplified approach to Cox's proportional hazards model

For this example, we will examine the first 4 study deaths, occurring at days 5, 19, 26, and 40. For the first event, occurring on day 5, a fully adjusted logistic-like regression model is fit to predict the independent association between treatment type (surgery versus surveillance) and mortality, holding all other variables in the model constant, *for the one death that occurs on study day 5*. This procedure will yield a wildly variable estimate for all of the variables in the model, because the estimates will be based on only a single death. The identical model is then fit again for the second death, occurring at study day 19, and the results from this model are averaged with those from day 5. Models are then fit for outcomes occurring on study day 26, study day 40, and for every subsequent event that occurs during follow-up. The results from all of these models are averaged together to produce a summary result. As more events are studied, estimates for the variables in the model stabilize, with progressively lower variation.

The important fact to notice is that *follow-up time is held constant for all study subjects at the time of each comparison.* For the first logistic model that was fit for study day 5, all study subjects under comparison have been in the study for exactly 5 days. For the second logistic model, all study subjects have spent exactly 19 days in the study. This procedure is analogous to matching on follow-up time, and will account for differences in follow-up time between treatment groups during follow-up.

What about censoring? Like the Kaplan–Meier method for estimating $S(t)$, the Cox model ignores subjects who dropout of the study in between events, but removes them from analysis when the next failure occurs. For example, consider a study participant in the surgical treatment group who decides to leave the study on day 20. This person will contribute data to the first two logistic models, on days 5 and 19, but will not contribute data to any subsequent models. In this way, each logistic model is performed only among subjects at risk, i.e., subjects who remain in the study. The Cox model forms an elegant solution to differential follow-up time and dropout in clinical studies.

The Cox model yields a measure of risk called the *hazard ratio*, which very closely parallels relative risk. Recall that logistic regression yields the odds ratio of disease, and that the odds ratio approximates the relative risk when the prevalence of disease is low. The Cox model exploits this property by keeping the prevalence of disease very low for each comparison. The Cox model fits separate logistic models for each failure time, ensuring that the prevalence of the outcome is low for each

comparison. Because hazard ratios are calculated by averaging a series of models fit for each failure, hazard ratios represent summary relative risks during follow-up.

Mathematically, $S(t)$ and hazard ratios are related. Specifically, the hazard ratio is the derivative, or instantaneous slope of $S(t)$.

20.6.2 Interpreting Survival Data and the Proportional Hazards Model

The following results were published from the AAA study:

	Adjusted hazard ratio (95% confidence interval)
Surgery (versus surveillance)	0.83 (0.69, 1.00)

Adjusted for age, sex, smoking, initial aneurysm diameter, average of left and right ankle–brachial pressure index, forced expiratory volume in one second, aspirin use, referral source, and type of hospital (teaching or district general).

This result, obtained from a Cox proportional hazards model, can be interpreted as, "the relative hazard of mortality, comparing surgery to surveillance, is 17% lower among subjects assigned to surgery, compared to surveillance, *holding all of the variables in the model constant, and accounting for potential differences in follow-up time between treatment groups*".

Because hazard ratios are computed by averaging instantaneous risks throughout the entire course of a study, the hazard ratio represents a *summary* relative risk of death for the entire study period. However, recall that the survival data demonstrated that surgery leads to decreased survival early in the study, and then improved survival later in the study. In this case, a summary hazard ratio obscures the full extent of the association of surgery with mortality. An accurate approach that more accurately represents the study data is to calculate separate hazard ratios for early and late study times, as shown in Table 20. 6.

Computing separate hazard ratios for different follow-up times provides a more complete picture of the surgery versus surveillance story. Surgery is associated with a 2.5 fold greater risk of death during the first 6 follow-up months, reflecting the high mortality rate associated with AAA surgery, followed by a 23% lower risk of death after 6 months, reflecting the benefit of surgery among subjects who survive

Table 20.6 Hazard ratios for early and late study periods

	Adjusted hazard ratio (95% confidence interval)
Follow-up time 0–6 months:	
Surgery (versus surveillance)	2.52 (1.20, 5.33)
Follow-up time >6 months:	
Surgery (versus surveillance)	0.77 (0.63, 0.93)

the procedure. This example demonstrates that summary hazard ratios can be misleading for studies in which the relative risk changes over time, and that results from such studies are best presented using separate hazard ratios from different time periods of follow-up time. In contrast, summary hazard ratios *are* meaningful for studies in which the relative risk remains roughly constant throughout follow-up.

20.6.3 Survival Versus Hazard Ratio Data

In contrast to the hazard ratios, which change soon after surgery, the survival curves for the AAA study do not cross until nearly three years of follow-up (Fig.20. 2).

The discrepancy between survival and hazard data reflect the natural relationship between $S(t)$ and the hazard ratio. The hazard ratio represents the instantaneous risk of mortality, or the mortality rate, whereas $S(t)$ represents overall survival. The markedly greater risk of death associated with surgery during the first 6 study months indicates that many more subjects assigned to the surgery group died within the first 6 months of follow-up. This survival difference is reflected in $S(t)$, which documents the proportion of subjects remaining alive in each group. After about 6 months of follow-up, the hazard ratio data indicate that the *rate* of death becomes lower in the surgical group; however, it takes another 2.5 years for this lower mortality rate to translate into an equal proportion of subjects remaining alive in each treatment group.

Survival data and hazard ratio results are directly applicable to clinical decision-making and to patient counseling. One perspective on the AAA data is that successful surgery translates fairly quickly into lower rates of mortality for subjects who can survive the surgery. Another perspective is that overall survival does not favor surgery for at least 3 years, such that subjects with multiple medical problems who might not live for three years, may elect to avoid surgery. Clinical studies frequently present Kaplan–Meier survival estimates *and* hazard ratios, because these measures in combination provide the most complete picture of survival and mortality rates.

References

1. Hyman DJ, Henry A, Taylor A. Severe rhabdomyolysis related to cerivastatin without gemfibrozil. *Ann Intern Med.* Jul 2 2002;137(1):74.
2. Psaty BM, Furberg CD, Ray WA, Weiss NS. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: use of cerivastatin and risk of rhabdomyolysis. *JAMA.* Dec 1 2004;292(21):2622–2631.
3. Lindner A, Charra B, Sherrard DJ, Scribner BH. Accelerated atherosclerosis in prolonged maintenance hemodialysis. *N Engl J Med.* Mar 28 1974;290(13):697–701.
4. Munger KL, Levin LI, Hollis BW, Howard NS, Ascherio A. Serum 25-hydroxyvitamin D levels and risk of multiple sclerosis. *JAMA.* Dec 20 2006;296(23):2832–2838.
5. Menacker F. Trends in cesarean rates for first births and repeat cesarean rates for low-risk women: United States, 1990–2003. *Natl Vital Stat Rep.* Sep 22 2005;54(4):1–8.
6. Lydon-Rochelle M, Holt VL, Easterling TR, Martin DP. Risk of uterine rupture during labor among women with a prior cesarean delivery. *N Engl J Med.* Jul 52001;345(1):3–8.
7. Tofovic SP, Dubey R, Salah EM, Jackson EK. 2-Hydroxyestradiol attenuates renal disease in chronic puromycin aminonucleoside nephropathy. *J Am Soc Nephrol.* Nov 2002;13(11): 2737–2747.
8. Dooley AC, Weiss NS, Kestenbaum B. Increased risk of hip fracture among men with CKD. *Am J Kidney Dis.* Jan 2008;51(1):38–44.
9. Ariyo AA, Thach C, Tracy R. Lp(a) lipoprotein, vascular disease, and mortality in the elderly. *N Engl J Med.* Nov 27 2003;349(22):2108–2115.
10. Carpenter RG, Irgens LM, Blair PS, et al. Sudden unexplained infant death in 20 regions in Europe: case control study. *Lancet.* Jan 17 2004;363(9404):185–191.
11. Travis LB, Curtis RE, Glimelius B, et al. Bladder and kidney cancer following cyclophosphamide therapy for non-Hodgkin's lymphoma. *J Natl Cancer Inst.* Apr 5 1995;87(7):524–530.
12. Mell LK, Davis RL, Owens D. Association between streptococcal infection and obsessive-compulsive disorder, Tourette's syndrome, and tic disorder. *Pediatrics.* Jul 2005;116(1): 56–60.
13. Henley DV, Lipson N, Korach KS, Bloch CA. Prepubertal gynecomastia linked to lavender and tea tree oils. *N Engl J Med.* Feb 1 2007;356(5):479–485.
14. Centers for Disease Control and Prevention. Intussusception among recipients of rotavirus vaccine – United States, 1998–1999. *JAMA.* Aug 11 1999;282(6):520–521.
15. Lang IA, Galloway TS, Scarlett A, et al. Association of urinary bisphenol A concentration with medical disorders and laboratory abnormalities in adults. *JAMA*. Sep 17 2008;300(11): 1303–1310.
16. Smith NL, Psaty BM, Heckbert SR, Tracy RP, Cornell ES. The reliability of medication inventory methods compared to serum levels of cardiovascular drugs in the elderly. *J Clin Epidemiol.* Feb 1999;52(2):143–146.
17. Belanger CF, Hennekens CH, Rosner B, Speizer FE. The nurses' health study. *Am J Nurs.* Jun 1978;78(6):1039–1040.

18. Avorn J. In defense of pharmacoepidemiology – embracing the yin and yang of drug research. *N Engl J Med.* Nov 29 2007;357(22):2219–2221.
19. Aronoff S, Rosenblatt S, Braithwaite S, Egan JW, Mathisen AL, Schneider RL. Pioglitazone hydrochloride monotherapy improves glycemic control in the treatment of patients with type 2 diabetes: a 6-month randomized placebo-controlled dose-response study. The Pioglitazone 001 Study Group. *Diabetes Care.* Nov 2000;23(11):1605–1611.
20. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med.* Jun 14 2007;356(24):2457–2471.
21. Manson JE, Hsia J, Johnson KC, et al. Estrogen plus progestin and the risk of coronary heart disease. *N Engl J Med.* Aug 7 2003;349(6):523–534.
22. Prentice RL, Langer R, Stefanick ML, et al. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. *Am J Epidemiol.* Sep 1 2005;162(5):404–414.
23. Michaelsson K, Lithell H, Vessby B, Melhus H. Serum retinol levels and the risk of fracture. *N Engl J Med.* Jan 23 2003;348(4):287–294.
24. Smeeth L, Cook C, Fombonne E, et al. MMR vaccination and pervasive developmental disorders: a case-control study. *Lancet.* Sep 11–17 2004;364(9438):963–969.
25. Fried LP, Borhani NO, Enright P, et al. The Cardiovascular Health Study: design and rationale. *Ann Epidemiol.* Feb 1991;1(3):263–276.
26. Cardo DM, Culver DH, Ciesielski CA, et al. A case-control study of HIV seroconversion in health care workers after percutaneous exposure. Centers for Disease Control and Prevention Needlestick Surveillance Group. *N Engl J Med.* Nov 20 1997;337(21):1485–1490.
27. Papadakis MA, Teherani A, Banach MA, et al. Disciplinary action by medical boards and prior behavior in medical school. *N Engl J Med.* Dec 22 2005;353(25):2673–2682.
28. Wolfe RA, Ashby VB, Milford EL, et al. Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *N Engl J Med.* Dec 2 1999;341(23):1725–1730.
29. Schiffl H, Lang SM, Fischer R. Daily hemodialysis and the outcome of acute renal failure. *N Engl J Med.* Jan 31 2002;346(5):305–310.
30. Palevsky PM, Zhang JH, O'Connor TZ, et al. Intensity of renal support in critically ill patients with acute kidney injury. *N Engl J Med.* Jul 3 2008;359(1):7–20.
31. Waber RL, Shiv B, Carmon Z, Ariely D. Commercial features of placebo and therapeutic efficacy. *JAMA.* Mar 5 2008;299(9):1016–1017.
32. Mehta RL, McDonald B, Gabbai FB, et al. A randomized clinical trial of continuous versus intermittent dialysis for acute renal failure. *Kidney Int.* Sep 2001;60(3):1154–1163.
33. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. *N Engl J Med.* Aug 10 1989;321(6):406–412.
34. Weiss NS, Koepsell TD, Psaty BM. Generalizability of the results of randomized trials. *Arch Intern Med.* Jan 28 2008;168(2):133–135.
35. Juurlink DN, Mamdani M, Kopp A, Laupacis A, Redelmeier DA. Drug–drug interactions among elderly patients hospitalized for drug toxicity. *JAMA.* Apr 2 2003;289(13): 1652–1658.
36. Pitt B, Zannad F, Remme WJ, et al. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. Randomized Aldactone Evaluation Study Investigators. *N Engl J Med.* Sep 2 1999;341(10):709–717.
37. Lassen MR, Ageno W, Borris LC, et al. Rivaroxaban versus enoxaparin for thromboprophylaxis after total knee arthroplasty. *N Engl J Med.* Jun 26 2008;358(26):2776–2786.
38. Hu FB, Stampfer MJ, Manson JE, et al. Dietary fat intake and the risk of coronary heart disease in women. *N Engl J Med.* Nov 20 1997;337(21):1491–1499.
39. Psaty BM, Koepsell TD, Lin D, et al. Assessment and control for confounding by indication in observational studies. *J Am Geriatr Soc.* Jun 1999;47(6):749–754.
40. Mehta RL, Pascual MT, Soroko S, Chertow GM. Diuretics, mortality, and nonrecovery of renal function in acute renal failure. *JAMA.* Nov 27 2002;288(20):2547–2553.

41. Kent DM, Price LL, Ringleb P, Hill MD, Selker HP. Sex-based differences in response to recombinant tissue plasminogen activator in acute ischemic stroke: a pooled analysis of randomized clinical trials. *Stroke.* Jan 2005;36(1):62–65.
42. Terry PD, Miller AB, Rohan TE. Obesity and colorectal cancer risk in women. *Gut.* Aug 2002;51(2):191–194.
43. Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med.* Dec 21 2006; 355(25):2615–2617.
44. United Kingdom Small Aneurysm Trial Participants. Long-term outcomes of immediate repair compared with surveillance of small abdominal aortic aneurysms. *N Engl J Med.* May 9 2002;346(19):1445–1452.References

Author Index

Subject Index