

2. Rothman KJ, Cann CI, Flanders D, Fried MP. Epidemiology of laryngeal cancer. *Epidemiol Rev.* 1980;2:195-209.
3. Murabito JM, D'Agostino RB, Sibershatz H, Wilson PWF. Intermittent claudication: a risk profile from the Framingham Heart Study. *Circulation.* 1997;96:44-49.
4. Rothman KJ. *Modern Epidemiology*. 1st ed. Boston: Little, Brown; 1986.
5. Assman SF, Hosmer DW, Lemeshow S, Mundt KA. Confidence intervals for measures of interaction. *Epidemiology.* 1996;7:286-290.
6. Knol MJ, Vanderweele TJ, Groenwold RH, Klungel OH, Rovers MM, Grobbee DE. Estimating measures of interaction on an additive scale for preventive exposures. *Eur J Epidemiol.* 2011;26:433-438.
7. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med.* 2004;66:411-421.
8. Green SB. How many subjects does it take to do a regression analysis? *Multivar Behav Res.* 1991;26:499-510.
9. Miettinen OS. Stratification by a multivariate confounder score. *Am J Epidemiol.* 1976;104:609-620.
10. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol.* 2006;59:437-447.
11. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *Am J Epidemiol.* 2010;172:843-854.
12. Seeger JD, Williams PL, Walker AM. An application of propensity score matching using claims data. *Pharmacoepidemiol Drug Saf.* 2005;14:465-476.
13. Mauri L, Silbaugh TS, Garg P, et al. Drug-eluting or bare-metal stents for acute myocardial infarction. *N Engl J Med.* 2008;359:1330-1342.
14. Rothman KJ. Drug-eluting versus bare-metal stents in acute myocardial infarction. *N Engl J Med.* 2009;360:301.
15. Mauri L, Normand S-LT. Drug-eluting versus bare-metal stents in acute myocardial infarction. *N Engl J Med.* 2009;360:301-302.

## Epidemiology in Clinical Settings

Clinical epidemiology focuses the application of epidemiologic principles on questions that relate to diagnosis, prognosis, and therapy. It also encompasses screening and other aspects of preventive medicine at both the population and the individual level. Therapeutic thinking has been greatly affected by advances in pharmacoepidemiology, an area that has extended the reach of epidemiologic research from the study of drug benefits to that of adverse effects and has led to the burgeoning fields of *outcomes research* and *comparative effectiveness*. Outcomes research marries epidemiologic methods with clinical decision theory to determine which therapeutic approaches are the most cost-effective, whereas comparative effectiveness aims to evaluate the effect of different interventions against one another in a variety of settings.

### DIAGNOSIS

Assigning a diagnosis is both crucial and subtle. To a large extent, the process of diagnosis may appear to involve intuition, conviction, and guesswork, processes that are opaque to quantification and analysis. Nevertheless, formal approaches to understanding and refining the steps in assigning a diagnosis have helped to clarify the thinking and solidify the foundation for diagnostic decision making. The basis for formulating a diagnosis comprises the data from signs, symptoms, and diagnostic test results that distinguish those with a specific disease from those who do not have that disease.

### The Gold Standard

Diagnosis cannot be a perfect process. Rarely does any sign or symptom, or any combination of them, distinguish completely between those with and those without a disease. Often a diagnosis is considered established when a specific combination of signs and symptoms that has been posed as the criterion for disease



2. Rothman KJ, Cann CI, Flanders D, Fried MP. Epidemiology of laryngeal cancer. *Epidemiol Rev.* 1980;2:195-209.
3. Murabito JM, D'Agostino RB, Sibershatz H, Wilson PWF. Intermittent claudication: a risk profile from the Framingham Heart Study. *Circulation.* 1997;96:44-49.
4. Rothman KJ. *Modern Epidemiology*. 1st ed. Boston: Little, Brown; 1986.
5. Assman SF, Hosmer DW, Lemeshow S, Mundt KA. Confidence intervals for measures of interaction. *Epidemiology.* 1996;7:286-290.
6. Knol MJ, Vanderweele TJ, Groenwold RH, Klungel OH, Rovers MM, Grobbee DE. Estimating measures of interaction on an additive scale for preventive exposures. *Eur J Epidemiol.* 2011;26:433-438.
7. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med.* 2004;66:411-421.
8. Green SB. How many subjects does it take to do a regression analysis? *Multivar Behav Res.* 1991;26:499-510.
9. Miettinen OS. Stratification by a multivariate confounder score. *Am J Epidemiol.* 1976;104:609-620.
10. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol.* 2006;59:437-447.
11. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *Am J Epidemiol.* 2010;172:843-854.
12. Seeger JD, Williams PL, Walker AM. An application of propensity score matching using claims data. *Pharmacoepidemiol Drug Saf.* 2005;14:465-476.
13. Mauri L, Silbaugh TS, Garg P, et al. Drug-eluting or bare-metal stents for acute myocardial infarction. *N Engl J Med.* 2008;359:1330-1342.
14. Rothman KJ. Drug-eluting versus bare-metal stents in acute myocardial infarction. *N Engl J Med.* 2009;360:301.
15. Mauri L, Normand S-LT. Drug-eluting versus bare-metal stents in acute myocardial infarction. *N Engl J Med.* 2009;360:301-302.

## Epidemiology in Clinical Settings

Clinical epidemiology focuses the application of epidemiologic principles on questions that relate to diagnosis, prognosis, and therapy. It also encompasses screening and other aspects of preventive medicine at both the population and the individual level. Therapeutic thinking has been greatly affected by advances in pharmacoepidemiology, an area that has extended the reach of epidemiologic research from the study of drug benefits to that of adverse effects and has led to the burgeoning fields of *outcomes research* and *comparative effectiveness*. Outcomes research marries epidemiologic methods with clinical decision theory to determine which therapeutic approaches are the most cost-effective, whereas comparative effectiveness aims to evaluate the effect of different interventions against one another in a variety of settings.

### DIAGNOSIS

Assigning a diagnosis is both crucial and subtle. To a large extent, the process of diagnosis may appear to involve intuition, conviction, and guesswork, processes that are opaque to quantification and analysis. Nevertheless, formal approaches to understanding and refining the steps in assigning a diagnosis have helped to clarify the thinking and solidify the foundation for diagnostic decision making. The basis for formulating a diagnosis comprises the data from signs, symptoms, and diagnostic test results that distinguish those with a specific disease from those who do not have that disease.

### The Gold Standard

Diagnosis cannot be a perfect process. Rarely does any sign or symptom, or any combination of them, distinguish completely between those with and those without a disease. Often a diagnosis is considered established when a specific combination of signs and symptoms that has been posed as the criterion for disease



is present. A diagnosis meeting this standard may be "definitive" but only in a circular sense, that is, by definition. Another way that a definitive diagnosis can be reached is by expert judgment, often by consensus; but once again this approach makes a diagnosis definitive only by definition. No approach is perfect, and two different approaches to the same disease will not necessarily lead to the same classification for every patient. Nevertheless, even if it is arbitrary, we need to have some definition of disease to use as a "gold standard" by which to judge the findings from individual signs and symptoms or screening tests.

### Sensitivity and Specificity

For years, the diagnosis of tuberculosis (TB) has rested on detection of the *Mycobacterium tuberculosis* organism from smears of acid-fast bacilli and from culture, but this method requires 10,000 bacteria/mL and does not distinguish among various mycobacteria. Catanzaro et al.<sup>1</sup> investigated how well an acid-fast smear predicted the diagnosis of clinical TB among patients who were suspected to have active pulmonary TB solely on the basis of clinical judgment. The diagnosis of TB was established by an expert panel of three judges, who used culture information and clinical information according to specific guidelines to classify patients into those who had and those who did not have TB. The distribution by diagnosis and by outcome of the acid-fast smear results is given in Table 13-1.

A total of 338 patients with suspected active pulmonary TB were studied. Of these, 72 (21%) were diagnosed as having it. Among these 72 TB patients, 43 (60%) had a positive smear. This proportion is known as the *sensitivity* of the smear. The sensitivity of a test, sign, or symptom is defined as the proportion of people with the disease who also have a positive result for the test, sign, or symptom. If everyone who has the disease has a given sign or symptom, the sensitivity of that sign or symptom is 100%. It is easy to find signs or symptoms that have high sensitivities. For example, in diagnosing headache, we might note that all patients have heads, making the sensitivity of having a head 100%. Having a head would have a low *specificity*, however. The specificity of a test, sign, or symptom is the proportion of people among those who do not have the disease

Table 13-1 DISTRIBUTION OF PATIENTS WITH SUSPECTED ACTIVE PULMONARY TUBERCULOSIS, BY DIAGNOSIS AND BY RESULTS OF ACID-FAST BACILLUS SMEAR TESTING

Smear	Present	Tuberculosis Absent	Total
Positive	43	22	65
Negative	29	244	273
Total	72	266	338

$$\text{Sensitivity of smear} = \frac{43}{72} = 60\% \quad \text{Predictive Value Positive of smear} = \frac{43}{65} = 66\%$$

$$\text{Specificity of smear} = \frac{244}{266} = 92\% \quad \text{Predictive Value Negative of smear} = \frac{244}{273} = 89\%$$

who have a negative test, sign, or symptom. The specificity of the acid-fast smear test, based on the data in Table 13-1, was 244/266 (92%). The specificity of having a head in diagnosing a headache would be zero, because everyone has a head. The most useful tests, signs, or symptoms for diagnosing a disease are those with both high sensitivity and high specificity. A test with 100% sensitivity and 100% specificity would be positive for everyone with disease and negative for everyone without disease. Almost all tests, however, fall short of providing perfect separation of those with and without disease.

Tests, signs, and symptoms can be used in combination to improve either the sensitivity or the specificity. Suppose test A had a sensitivity of 80% and a specificity of 90% by itself, and test B also had a sensitivity of 80% and a specificity of 90%. If we used the two tests in combination to indicate disease, we might postulate that a positive result on both tests would be required to indicate the presence of disease. If the tests results were independent of each other, then  $0.8 \times 0.8 = 0.64$  of all patients with disease would test positive on both, making the sensitivity of the combination 64%, worse than the sensitivity of either test alone. On the other hand, the specificity would improve, because those who are negative for the combination of tests would include all those who tested negative on either test. In this example, 90% of those without disease would test negative on the first test, and among the 10% who did not, 90% would test negative on the second test, making the specificity of the combination  $0.9 + (0.1 \times 0.9) = 99\%$ . Therefore, requiring a positive result from two tests increases the specificity but decreases the sensitivity.

The reverse occurs if a positive result on either test is taken to indicate the presence of disease. For the example given, 80% of those with disease would test positive on the first test, and of the remaining 20%, 80% would test positive on the second test, making the sensitivity  $0.8 + (0.2 \times 0.8) = 96\%$ . The price paid to obtain a higher sensitivity is a lower specificity, which would be the proportion of those without disease who test negative on both tests,  $0.9 \times 0.9 = 81\%$ .

This discussion assumes that the test results are independent, which is rarely the case. Nevertheless, the principle always applies that combinations of tests, signs, and symptoms can be used to increase either the sensitivity or the specificity—one at the cost of the other—depending on how a positive outcome for the combination of tests is defined. This principle is used to detect cervical cancer by a Papanicolaou smear, which has a high sensitivity but a lower specificity. As a result, a Pap smear will detect almost all cervical cancers but has a high proportion of false-positive results. By requiring a sequence of positive Pap smears before taking further diagnostic action, however, it is possible to improve the specificity of the smear (ie, reduce the false-positive results) without compromising by much the already high sensitivity. In recent years there has been improvement on the approach of repeated smears: now, a single cervical smear can be simultaneously tested for the DNA of human papilloma virus, another risk factor for cervical cancer, to improve the sensitivity of a single screen rather than having to rely on repeated Pap testing.<sup>2</sup>

### Predictive Value

Sensitivity and specificity describe the characteristics of a test, sign, or symptom in correctly classifying those who have or do not have a disease. *Predictive value* is



a measure of the usefulness of a test, sign, or symptom in classifying people with disease. It can be calculated from the same basic data from which we calculate sensitivity and specificity. Consider the TB example in Table 13-1. We can use these data to calculate the predictive value of a positive smear. Among the 65 people with a positive smear, 43 had TB. Therefore, a positive smear correctly indicated the presence of TB in 43/65 (66%) of people who were tested. This proportion is referred to as the *predictive value positive*, or the predictive value of a positive test, usually abbreviated as PV+. We can also measure the predictive value negative, or the predictive value of a negative test, which is abbreviated PV-. In the same data, of the 273 who had a negative smear, 244 did not have TB, making the predictive value negative of the smear 244/273 (89%).

Sensitivity and specificity should theoretically be constant properties of a test, regardless of the population that is being tested, but in practice they can vary with the mix of patients. In contrast, predictive value varies even theoretically from one population to another, because it is highly dependent on the prevalence of disease in the population being tested. We can illustrate the dependence of predictive value on the prevalence of disease by examining what would result if we added to the population described in Table 13-1 500 people who did not have TB. The effect is similar to the change one would find in moving from a clinic serving a population in which TB was common to a clinic serving a population in which TB was less common. The augmented data are displayed in Table 13-2.

Let us assume that the sensitivity and specificity of the test remain the same. We still have 72 people with TB, of whom 43 have a positive smear. We now have 766 people without TB, which includes the original 266 plus 500 additional people who do not have TB. We have assumed that the specificity of the test remains the same, 92%, which means that 703 of the 766 patients without TB will have a negative smear. The PV+ and PV- are considerably different in this second population, however. The PV+ is  $43/106 = 41\%$ , much less than the PV+ of 66% for the population in Table 13-1. As the prevalence of disease decreases, the predictive value of a positive test will decrease as well. At the same time, the PV- has changed from 89% in the original data to  $703/732 = 96\%$  in the augmented data. As the prevalence of the disease decreases, the PV+ decreases but the PV- increases.

Table 13-2 RESULTS FROM TABLE 13-1 AUGMENTED WITH DATA FROM 500 ADDITIONAL PEOPLE WITHOUT TUBERCULOSIS

Smear	Tuberculosis		Total
	Present	Absent	
Positive	43	63	106
Negative	29	703	732
Total	72	766	838

$$\begin{aligned} \text{Sensitivity of smear} &= \frac{43}{72} = 60\% & \text{Predictive Value Positive of smear} &= \frac{43}{106} = 41\% \\ \text{Specificity of smear} &= \frac{703}{766} = 92\% & \text{Predictive Value Negative of smear} &= \frac{703}{732} = 96\% \end{aligned}$$

These changes in predictive value with changes in prevalence should not be too surprising. If we tested a population in which no one had disease, there would still be some false-positive test results. The predictive value of a positive test in such a population would be zero, because no one in that population actually had the disease. On the other hand, the predictive value of a negative test would be perfect (100%). Taking the other extreme, if everyone in a population had the disease, then the PV+ would be 100% and the PV- would be zero. Changes in predictive value with prevalence of disease have implications for the use of diagnostic and screening tests. Tests that have reasonably good PV+ in a clinic population of patients presenting with symptoms may have little PV+ in an asymptomatic population being screened for disease. For this reason, it may not make sense to convert diagnostic tests into screening tests that would be applied to populations with a low prevalence of disease.

### Screening

The premise of screening for disease is that for many diseases early detection improves the prognosis. Otherwise, there would be no point to screening, because it is expensive both in monetary terms and in terms of the burden it places on the screened population. To be suitable for screening, a disease must be detectable during a preclinical phase by some test, and early treatment must convey a benefit over later treatment (ie, the treatment that would occur after the disease comes to attention without screening).<sup>3</sup> Furthermore, the benefit that early treatment conveys should outweigh the overall costs of the screening. These costs are more than just the expense of administering the screening test to a healthy population. Screening will result in some false-positive tests, saddling those who have the false result with the mistaken prospect of facing a disease that they do not have. Furthermore, a false-positive test usually leads to further tests and sometimes even to treatments that are unnecessary and risky. Another cost comes from false-negative results, which provide false reassurance about the absence of disease. Even for those whom the screening test labels correctly with disease, there is a psychological cost that comes from being labeled earlier in the natural history of the process than would have occurred without screening. Weighing against this cost is the useful reassurance for those who do not have the disease that comes from having tested negative.

For screening to succeed, the disease being screened for should have a reasonably long preclinical phase so that the prevalence of people in this preclinical phase is high. If the preclinical phase is short and people who develop the disease promptly pass through it into a clinical phase, there is little point to screening. In such a situation, the low prevalence of the preclinical phase of the disease in the population will produce a low PV+ for the screening test.

### LEAD-TIME BIAS

Because screening advances the date of diagnosis for a disease, it can be difficult to measure its effect. Suppose the disease is cancer. The success of treating cancer is usually measured by the survival time after diagnosis or the time to recurrence. If early treatment is advantageous, one would expect it to result in longer survival



time or longer time until recurrence. After screening, however, survival time and time to recurrence will increase even if the screening and earlier treatment do no good. The reason is that the time of diagnosis is moved ahead by screening, so that the diagnosis is registered earlier in the natural history of the disease process than it would have been without screening. The difference in time between the date of diagnosis with screening and the date of diagnosis without screening is called the *lead time*. Lead time should not be counted as part of the survival time after disease diagnosis, because it does not represent any real benefit. If it is counted, it will erroneously inflate the survival time, a problem known as *lead-time bias*. Lead time can be estimated by comparing the course of disease among a screened population with the course of disease among a similar population that has not been screened.

### PROGNOSTIC SELECTION BIAS

In addition to lead-time bias, another difficulty in evaluating the success of a screening effort is bias that comes from self-selection of subjects who decide to be screened. This bias is called *prognostic selection bias*. Because screening programs are voluntary, those who volunteer to get screened will differ in many ways from those who refuse to be screened. Volunteers are likely to be more interested in their health, to be more eager to take actions that improve their health, and to have a more favorable clinical course even in the absence of a benefit from screening. One way to avoid this bias, as well as lead-time bias and the effect of length-biased sampling (see next section), is to evaluate the screening test or program in a randomized trial. In nonexperimental studies, however, these biases are important issues that must be taken into account to obtain a valid assessment of screening efficacy.

### LENGTH-BIASED SAMPLING

Another difficulty in measuring the effect of screening comes from *length-biased sampling*, which results from natural variability in the progression rate of disease. To simplify the issue, suppose that breast cancer comes in two types, fast-progressing and slow-progressing. Those with fast-progressing breast cancer have the worse prognosis; their disease goes quickly through the preclinical phase into a clinical phase and spreads rapidly, leading to an early demise for many patients. Slow-progressing breast cancer is more benign, taking many more months or years to progress through the preclinical phase into a clinical phase that also is characterized by slow progression. Women with slow-progressing breast cancer have a better prognosis, even without treatment, although they are also more likely to benefit from treatment.

Let us assume that an equal number of cases of slow-progressing and fast-progressing breast cancer occur in a population. Despite the equal incidence, the prevalence of slow-progressing cases would be greater, because prevalence reflects duration as well as incidence. Thus, more individuals with slow-progressing breast cancer will be in the preclinical phase of disease, because each case takes longer to pass through that stage of the disease process. A screening program, therefore, would tend to identify more slow-progressing cases than fast-progressing cases. Even if early identification and treatment of breast cancer had no effect on the disease, cases identified in a screening program would tend to have a better

prognosis than the average of all cases because of length-biased sampling: the screening tends to favor identification of slow-progressing cases, which have a better prognosis.

### PROGNOSIS

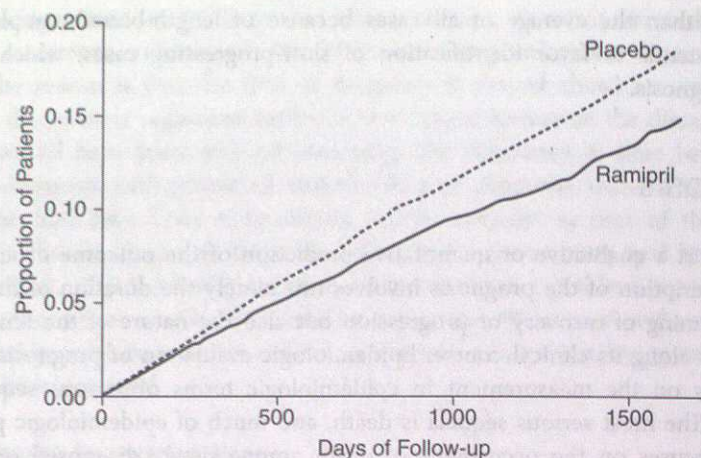
Prognosis is a qualitative or quantitative prediction of the outcome of an illness. A full description of the prognosis involves not merely the duration of the illness and the timing of recovery or progression but also the nature of the illness as it progresses along its clinical course. Epidemiologic evaluation of prognosis focuses specifically on the measurement in epidemiologic terms of serious sequelae or recovery. The most serious sequela is death, and much of epidemiologic prognostication focuses on the occurrence of death among newly diagnosed or treated patients.

The simplest epidemiologic measure of prognosis is the *case-fatality rate*. Despite the name, this measure is an incidence proportion rather than a true rate. It is the proportion of people with newly diagnosed disease who die from the disease. Strictly speaking, the case-fatality rate should be measured over a fixed and stated time period, such as 3 months or 12 months. Traditionally, however, the measure has been used to describe the clinical course of acute infectious illnesses that progress toward recovery or death within a short time. The time period implicit in the measure is the period of active infection and its aftermath and is often left unspecified. For example, we might describe typhoid fever as having a case-fatality rate of 0.01, paralytic poliomyelitis as having a case-fatality rate of 0.05, and Ebola disease as having a case-fatality rate of 0.75, with each disease having its own characteristic time period during which the patient either dies or recovers.<sup>4</sup> Eventually, of course, all the patients with any disease will die from one cause or another. The presumption of the case-fatality rate is that essentially all of the deaths that occur promptly after disease onset are a consequence of the disease.

For diseases with a long clinical course, it becomes more important to specify the time period over which the case-fatality rate is measured. When it is measured over longer periods, the term case-fatality rate is often not even used. Instead, we use terms such as *5-year survival rate* to refer to the proportion of patients surviving for 5 years after diagnosis. This is simply the complement of the proportion who die during the same period. Beyond a simple incidence proportion or survival proportion, we can derive a survival curve, which gives the survival probability according to time since diagnosis. The survival curve conveys information about the survival proportion for all time periods up to the limit of what has been observed, thus providing greater information than any single survival proportion. (A common method for obtaining a survival curve is the *Kaplan-Meier product-limit method*, which is a variant of the life-table approach described in Chapter 4. The Kaplan-Meier method recalculates the proportion of survivors at the time of each death in a cohort.<sup>5</sup>)

The complement and close cousin to the survival curve is the curve that expresses the cumulative proportion of patients who reach a specific end point. Figure 13-1 exemplifies a pair of such cumulative incidence curves. They





**Figure 13-1** Cumulative proportion of patients experiencing a myocardial infarction, stroke, or death from cardiovascular causes, by treatment group. (Adapted with permission from Heart Outcomes Prevention Evaluation Study Investigators. Effects of an angiotensin-converting enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. *N Engl J Med.* 2000;342:145-153, copyright © 2000, Massachusetts Medical Society. All rights reserved.)

describe the results of a randomized trial that compared the effect of ramipril, an angiotensin-converting enzyme inhibitor, with placebo in preventing the occurrence of myocardial infarction, stroke, or death in patients with certain cardiovascular risk factors.<sup>6</sup> In this example, the curves show the cumulative proportion who experienced any of the end points over a 5-year follow-up period.

## THERAPY

It has been said that it was not until the early 20th century that the average patient visiting the average physician was likely to benefit from the encounter. The course of illness today is often greatly affected by the choice of treatment options. The large clinical research enterprise that evaluates new therapies is heavily dependent on epidemiology. In fact, a large part of clinical research is clinical epidemiology.

### Clinical Trials

The randomized clinical trial is the epidemiologic centerpiece of clinical epidemiology. Although the clinical trial is but one type of epidemiologic experiment (the others are field trials and community intervention trials), it is by far the most common. (See Chapter 5 for a discussion of the types of epidemiologic experiments.) A full discussion of clinical trials merits a separate textbook; here, I will only touch on some highlights that bear on the interpretation of trial results.

The central advantage of trials over nonexperimental studies is their ability to control confounding effectively. A particularly knotty problem in therapeutics

is the problem of *confounding by indication*. When nonexperimental studies are conducted to compare the outcomes of different treatments, confounding by indication can present an insuperable problem. Confounding by indication is a bias that stems from inherent differences in prognosis between patients given different therapies. For example, suppose a new antibiotic shows promise in treating resistant strains of meningitis-causing bacteria but has common adverse effects and is costly. It is likely that the new treatment will be reserved for patients who face the greatest risk of a fatal outcome. Even if the drug is highly effective, the mortality rate among those who receive it could be greater than that among those receiving the standard drugs, because those who get the new drug are at the highest risk. A valid evaluation of the new drug can be achieved only if the prognostic differences can be adjusted or otherwise controlled in the epidemiologic comparison. Nonexperimental studies can deal with such confounding by indication if there is sufficiently good information on measured risk factors for the disease complication that the therapy aims to prevent. Nevertheless, the best efforts to control confounding by indication often fail to remove all of the bias. This problem is the primary motivation to conduct experiments that compare therapies. With the random assignment that is possible in a clinical trial, prognostic factors can be balanced between groups receiving different therapies.

### BLINDING AND USE OF PLACEBOS

*Blinding* refers to hiding information about treatment assignment from the key participants in a trial. The concern is that knowledge of the treatment assignment will influence the evaluation of the outcome. This concern relates most directly to the person or persons who are supposed to make judgments or decisions regarding the outcome. For example, if the outcome is hospitalization for an exacerbation of the disease, the physician who makes the determination about hospitalization might be influenced by knowledge about which treatment was assigned to a given patient. This concern is amplified if the physician has a strong view about the merits of the new therapy. If the physician does not know which treatment the patient has received, then the evaluation should be free of this source of potential bias.

Blinding is not always necessary. If the only outcome of interest is death, there is little reason to be concerned about biased classification of the outcome, because judgment is not an important factor in determining whether someone is dead. In other instances, blinding may be infeasible. If the treatment is an elaborate intervention, such as major surgery, it may be neither possible nor ethical to provide a sham procedure that would allow blinding.

Some trials are described as *double-blind*. This term implies that the evaluator assessing the patient for the possible outcome does not know the treatment assignment, and the patient also does not know the treatment assignment. The person who administers the treatment may also be kept unaware of which treatment is being assigned, in which case the study might be described as *triple-blind*. In all of these situations, the goal is to keep the information about treatment assignment a secret so that the evaluation of the outcome will not be affected.

One method that is often used to facilitate blinding is *placebo* treatment for the comparison group. A placebo (from the Latin, "I shall please") is intended to have no biologic effect outside the offer of treatment itself. Placebo pills typically contain sugar or other essentially inert ingredients. Such pills can be manufactured to



be indistinguishable from the new therapy being offered. Other types of placebo treatment involve sham procedures. For example, in a trial of acupuncture, the placebo treatment could involve the application of acupuncture needles at points that are, according to acupuncture theory, not correct. Placebo treatments need to be adapted to the particular experiment in which they are used.

Although a placebo treatment facilitates blinding, that is not the primary reason it is used. It has long been known that even if a treatment has no effect, offering that treatment may have a salubrious effect. An offer of treatment is an offer of hope, and it may bring the expectation of treatment success. Expectations are thought to have a powerful influence on outcome. If so, a new treatment may have an effect that comes only through the lifting of patient expectations. According to some scientists, "The history of medical treatment until recently is largely the history of the placebo effect."<sup>7</sup> The use of a placebo comparison in a trial is intended to distinguish treatments that have only a placebo effect from those that have a greater therapeutic effect. The placebo effect itself is highly variable, depending on the nature of the outcome and the nature of the treatment.

#### ETHICS OF PLACEBO USE IN RANDOMIZED TRIALS

Only decades ago, it was common for physicians to prescribe placebos so that patients could benefit from improved expectations. Today such practice is rare, and many would consider it unethical. Placebo use continues in randomized trials, however, where the biggest concern is also an ethical one. According to the 1964 Declaration of Helsinki of the World Medical Association,<sup>8</sup> the interests of patients must come before the interests of science and society. Furthermore, every patient in a trial should be assured of getting the equivalent of the best available treatment, even those assigned to the comparison group. Therefore, it is unethical to use a placebo in any trial if there is already an accepted treatment for the condition under study. Instead, an investigator must test a new therapy against the existing standard, to see if it beats the current best treatment.

According to the principles embodied in the Declaration of Helsinki, no researcher should deny a patient the best available treatment solely for the purpose of learning whether a new treatment is better than placebo. Identifying new treatments that are better than placebo but worse than the current best treatment is of less interest than identifying new treatments that are better than the best existing treatment. As medicine progresses, there should be fewer and fewer conditions for which a placebo-controlled study is ethical, because standard therapies that are better than placebo will exist for more and more conditions. Unfortunately, the use of placebos in trials has achieved paradigm status in the minds of many researchers and even official agencies.<sup>9</sup> The paradigm should certainly include a comparison, but not necessarily a placebo comparison.<sup>8,9</sup>

#### THREATS TO VALIDITY IN TRIALS

Despite the strengths of randomized trials, there are several issues that can lead to biases in assessment. As mentioned, blinding is intended to reduce some of

these biases, by reducing opportunities for subjective evaluations to be influenced by knowledge of treatment. Some other sources of bias in trials are incomplete follow-up, intent-to-treat analysis, and confounding imbalances that stem from random assignment.

#### Incomplete Follow-up

Randomized trials are susceptible to many of the same biases that afflict other types of cohort studies. One source of bias is differential follow-up of the treatment groups. The ideal situation regarding follow-up is for there to be no subjects lost to follow-up, which prevents any bias from this source. In most trials, however, some subjects are not followed to the intended study end point. Reasons for incomplete follow-up are the same ones that occur in other cohort studies, which include subjects moving from the study area, withdrawing their consent to participate in the study, or dying from a disease that is not one of the study end points. If some study subjects are lost to follow-up for any of these reasons, the count of events will be underestimated compared with what it would have been had there been no losses to follow-up.

To deal with this potential source of bias, investigators may analyze the data under the assumption that the experience of those who were lost to follow-up is similar to that of those who remained in the study. This assumption, however, is not always reasonable. For example, subjects with worsening symptoms may be more inclined to drop out of the study than those with a better prognosis. In that case, the risk of the outcome in each treatment group would be underestimated if it were based on the experience of those with complete follow-up. Alternatively, those with the worst prognosis may be less likely to drop out of a study if they believe that they will receive better care by remaining in it. In that case, the study will overestimate the risks of the study outcome, because those dropping out are at lower risk than those remaining in the study. If follow-up is incomplete and is related to both the study intervention and the study outcome, the result is differential loss to follow-up between study groups, a type of selection bias. Differential loss to follow-up can lead to study results that are biased in either direction.

#### Intent-to-Treat Analysis

As described in Chapter 5, an intent-to-treat analysis is often employed in randomized trials. In this type of analysis, the random assignment at the outset of the trial determines the treatment group in which a subject will be included for the analysis, regardless of whether the subject adhered to that treatment assignment. Therefore, patients who get assigned to a new therapy but for various reasons decide to discontinue it, or never to begin taking it, will still be considered as part of that treatment group for the analysis. This approach maintains the benefits of random assignment for the comparison of a new treatment against an older treatment, but at the cost of misclassification of actual treatment. Those who "cross over" from their assigned treatment to the other treatment group, for example, will be analyzed with their assigned treatment, ignoring the crossover. As a result, the analysis using the intent-to-treat principle incorporates some misclassification of actual exposure. To the extent that the misclassification is independent of the study outcome, the misclassification will be nondifferential and will lead to underestimation of the effect of actual treatment.



Underestimation of the actual treatment effect is often considered acceptable, because it implies that a successful treatment is even better than the value estimated with the intent-to-treat approach. Nevertheless, as mentioned in Chapter 5, adverse effects of a treatment will also be underestimated by this method. This underestimation of risks is a serious drawback to using an intent-to-treat analysis for trials evaluating the safety of a treatment. In such trials, an analysis that classifies subjects according to their actual treatment may be preferred. Because an analysis based on actual treatment would not have all the benefits of a randomized comparison, the usual array of epidemiologic methods would have to be employed to assess and control confounding in the data analysis.

### Confounding Imbalances

Baseline risk factors are prognostic factors for the outcome that are measured at the time of random assignment. If randomization succeeds in achieving its goal, the frequency of the outcome will be similar in the various treatment groups created by randomization, apart from the effect of the intervention, because the overall risk for the outcome is balanced between groups. Although there is no direct way to measure whether such a balance in overall prognosis for the treatment groups has been achieved, it is possible to measure the distribution of individual prognostic factors in the compared groups to see how well balanced they are. Any imbalance in a baseline risk factor represents confounding, because a confounding factor is a risk factor that is associated with exposure. To say that a risk factor is imbalanced means that it is not distributed equally in the compared treatment groups and therefore is associated with the assigned treatment.

Randomization is intended to prevent confounding. The outcome of a random process, however, is predictable only if aggregated over many repetitions. In a specific case or in a particular trial, unlikely distributions can result from the randomization. In the University Group Diabetes Program,<sup>10</sup> the group that was randomly assigned to receive tolbutamide was older on average than the group randomly assigned to receive placebo. As a result, there was confounding by age in the evaluation of the tolbutamide effect. This age confounding was illustrated in Chapter 7: the crude difference in mortality proportion between tolbutamide and placebo, ignoring the age imbalance, was 0.045 (Table 7-7), whereas, after stratification into two age strata (Table 7-8), the tolbutamide effect was estimated as 0.035.

Distributions are rarely identical, so how can we tell when the imbalance in a baseline risk factor is severe enough to warrant treating the variable as a confounding factor? If a factor that is severely imbalanced has only a small effect on the outcome, there will be little confounding even with the large imbalance. On the other hand, even a modest imbalance in a strong risk factor for the outcome might lead to worrisome confounding. Therefore, the amount of imbalance in the risk factor is not, by itself, a good guide to the amount of confounding that the baseline imbalance introduces. The best way to assess the confounding is to use the same approach that epidemiologists use in other situations, which is basically the method that was used to compare the effects for tolbutamide estimated in Tables 7-7 and 7-8. Comparison of the crude estimate of effect, which is obtained without control of confounding, with an unconfounded estimate reveals how much confounding is removed when the variable is treated as

a confounder (see Chapter 10). It may seem cumbersome that one has to control the confounding to measure how much there is, but no evaluation of the imbalance in the baseline risk factor alone can reveal the amount of confounding, which depends on the interplay between that imbalance and the relation of the risk factor to the outcome.

A common mistake in conducting and reporting clinical trials is to use statistical significance testing to assess imbalances in baseline risk factors. Chapter 8 explains the problems with statistical significance testing in general and suggests that it be avoided. Table 8-3 from that chapter displays the results from a prominently published clinical trial that was misinterpreted because the authors relied on statistical significance for their inference. Use of statistical significance testing for interpretation of the results of a study is undesirable, but it is even less desirable to use statistical significance testing to assess baseline differences in a trial.

Aside from the usual problems with statistical significance testing that are described in Chapter 8, its use in the assessment of baseline imbalances introduces further problems. Perhaps the most obvious problem is that an imbalance in baseline risk factors by itself does not reflect the amount of confounding, as explained earlier. A second problem is that the amount of confounding is the result of the strength of the associations between the baseline risk factor and the two main study variables, treatment (exposure) and outcome (disease). In contrast, the result of a statistical significance test depends not just on the strength of the association being tested but also on the size of the study: for a given strength of association, more data results in a smaller *P* value. Thus, a given amount of confounding in a large study might yield a statistically significant difference in a baseline risk factor, whereas the same amount of confounding in a small study might not. For these reasons, it does not make much sense to use statistical significance testing to evaluate confounding. Instead, one should simply compare the crude effect estimate with the estimate after controlling for the possible confounder and assess the difference between the two results.

If an imbalance of baseline risk factors is serious enough to induce worrisome confounding into the effect estimate of a trial, how should it be handled? One school of thought holds that any imbalance should be ignored, because the intent of a randomized trial is to compare the experience of the randomized groups, period. According to this theory, one simply hopes that randomization will control successfully for all possible confounding factors, and then one relies on conducting a crude analysis without any control of confounding, no matter what happens after the randomization. The motivation for this view is that if the researcher does control for confounding, problems can be introduced into the analysis that can nullify the benefits of random assignment.

It is true that in an ideal setting randomization will prevent confounding. But if randomization has failed to prevent confounding, the options that the investigator faces are either to rely on a biased comparison of the crude data or to conduct an analysis that controls for the confounding that has been identified. Given the expense and effort of a trial, it makes little sense to ignore confounding that has been identified and thereby risk having the results of the study ignored because critics claim that the study is biased. It makes much more sense to attempt to control for any confounding that has been identified. Critics may still claim that



### AN UNREJECTABLE NULL HYPOTHESIS

There is yet another reason why the use of statistical significance testing to evaluate baseline imbalances in a clinical trial makes no sense. If such a statistical test is applied, one might ask what null hypothesis it tests. The answer must be that the null hypothesis is that any observed imbalance is just the result of chance. If a statistically significant result is observed, those who focus on significance testing might take that to mean that the null hypothesis is rejected. In the case of baseline imbalances in a randomized trial, that would mean rejecting the hypothesis that chance produced the imbalances. But we cannot reject that hypothesis! Apart from the possibility of chicanery or incompetence, we know that chance did in fact produce the imbalance: the imbalance is the result of a randomized allocation. Random assignment can produce unusual results, but we already know in a trial that the imbalances that do occur are due to chance. Therefore it makes no sense to test the null hypothesis. Actually, it makes no difference whether the imbalance was caused by chance or not. What matters is that the imbalance exists, and what is important to know is how much confounding it causes. Statistical significance testing cannot reveal that, but the straightforward application of epidemiologic rules for assessment of confounding can.

the randomization has "failed" (although it has not really failed). Nevertheless, the hope that random assignment will prevent confounding has already been defeated if confounding has been identified in the data. The question is how to proceed now that randomization has not prevented confounding.

Some might argue that if an identified confounder is controlled, that process itself can introduce confounding by some other, possibly unidentified factor. Although that is possible, there is no basis to assume that control of a known bias will introduce an unknown bias. Instead, it is more reasonable to control all identified confounders and treat the analysis like any other epidemiologic study.<sup>11,12</sup>

### Example: The Alzheimer's Disease Cooperative Study of Selegiline and $\alpha$ -Tocopherol

The question of how to deal with baseline differences arose in a trial<sup>13</sup> of selegiline and  $\alpha$ -tocopherol, two treatments intended to slow progression of Alzheimer's disease. The trial followed a factorial design; that is, participants were assigned to groups so that every combination of treatments was studied. In this study with two treatments, there were four groups: one group received only  $\alpha$ -tocopherol, one received only selegiline, one received both  $\alpha$ -tocopherol and selegiline, and one received a placebo. The mean score on the Mini-Mental State Examination (MMSE) at the start of the trial for the patients randomly assigned to receive  $\alpha$ -tocopherol alone was 11.3 on a scale from 0 to 30, whereas the placebo group had a mean score of 13.3 (higher scores indicate better cognitive function). Thus, the random assignment resulted in lower cognitive function at baseline in the group assigned to  $\alpha$ -tocopherol compared with the placebo group.

At first the investigators disregarded this difference, and they found that the  $\alpha$ -tocopherol group had a risk ratio of 0.7 with respect to the occurrence of at least one of several primary end points, including death, institutionalization, and onset of severe dementia. Thus the estimate of the crude effect of  $\alpha$ -tocopherol indicated a substantial benefit. Adjustment for the baseline difference in MMSE score would be expected to increase the estimated benefit even further, because the  $\alpha$ -tocopherol group had more signs of dementia to begin with, and this was indeed the case: the estimated rate ratio after adjusting for baseline differences was 0.47, representing an even greater benefit.

These findings were challenged by a correspondent,<sup>14</sup> who claimed that the adjusted results were biased. The critic did not offer a clear rationale for the supposed bias, nor did he discuss its magnitude or direction. When a critic suggests that a result is biased, it is incumbent on that person to quantify the effect of the bias. In this case, the critic implied that the adjusted results should be ignored and that the results from the crude analysis should be used for interpretation. Recall that even the crude effect, with no adjustment for the baseline difference, showed a worthwhile benefit, with a rate ratio of 0.70, indicating a 30% reduction in the occurrence of the primary end-point events. Nevertheless, the critic stated that "no true effect of treatment has been proved," suggesting that  $\alpha$ -tocopherol had no effect at all. This conclusion was apparently based not on the effect estimate, which showed a 30% reduction in occurrence of the adverse end points, but rather on a lack of statistical significance. This misinterpretation of the findings was aided by the authors of the original report, who themselves placed great emphasis on statistical significance. They also assessed the baseline differences in terms of their statistical significance rather than the amount of confounding that they produced.

In this example, which estimate of effect should be relied on as the best estimate of the effect of  $\alpha$ -tocopherol on Alzheimer's disease? The crude estimate for the rate ratio is 0.70, and the adjusted estimate is 0.47, but we know that the crude estimate is biased because of baseline differences in the MMSE score. It does not matter what the *P* value is for these baseline differences, nor exactly how they arose; what matters is the amount of confounding that they introduce. Contrary to what the correspondent asserted, the estimate of the  $\alpha$ -tocopherol effect after adjustment for those baseline differences contains less bias, not more bias, than the crude estimate of the effect. With adjustment, the estimated benefit of  $\alpha$ -tocopherol in slowing the progression of Alzheimer's disease is striking. In this example, distrust for an analysis that removed confounding and reliance on statistical significance testing for interpretation wrongly called into question a striking benefit.

### Pharmacoepidemiology

Drug epidemiology, also known as *pharmacoepidemiology*, is an active area of epidemiologic research that focuses on the effectiveness and safety of therapeutic drugs and devices. Although randomized trials are, strictly speaking, under the umbrella of pharmacoepidemiology, this discipline is commonly thought to comprise nonexperimental research on drugs and devices. Safety studies are often nonexperimental, because adverse effects are typically much less common than



the intended effects of drugs, and the randomized trials that are conducted to evaluate the efficacy of new drugs are seldom large enough to provide an adequate assessment of drug safety. Consequently, most of the epidemiologic information on drug safety comes from studies that are conducted after a drug is marketed.

This research activity is usually referred to as *postmarketing surveillance*. Much of it is not surveillance in the traditional sense; instead, it is based on discrete studies aimed at evaluating specific hypotheses. In the United States, however, the Food and Drug Administration (FDA) encourages the voluntary reporting of suspected adverse drug effects. These *spontaneous reports* are challenging to interpret. First, only a small, but unknown, proportion of suspected adverse drug effects are reported spontaneously; presumably, unexpected deaths, liver or kidney failure, and other serious events are more likely to be reported than skin rashes, but even so it is widely believed that only a small fraction of serious events are reported spontaneously. Second, it is difficult to know whether the number of spontaneously reported exposed cases, who represent only one cell in a  $2 \times 2$  table of exposure versus disease, represent an actual excess of exposed cases or just the number that chance would predict.

Case reports such as those submitted to the FDA as part of their surveillance effort are presumed to represent cases that are attributed to a given drug exposure; that is, the reporting process requires the reporter to make an inference about whether a specific drug exposure caused the adverse event. Although this type of inference is encouraged in clinical practice, it runs counter to the thinking that prevails in an epidemiologic study. As discussed in Chapter 3, it is not possible to infer logically whether a specific factor was the cause of an observed event. We can only theorize about the causal connection and test our theories with data.

Epidemiologists typically collect data from many people before making inferences about a causal connection, and we usually do not apply the inference to any specific person. If a person receives a drug and promptly dies of anaphylactic shock, a causal inference about the connection between the drug and the death may appear strong; but many inferences for individual events are tenuous, based more on conviction than anything else. The danger of thinking in terms of causal inferences in regard to individuals is that if this approach is applied to epidemiologic data, it defeats the validity of the epidemiologic process. If case inclusion in any epidemiologic evaluation takes into account information on exposure, it is apt to lead to biases. Instead, disease should be defined on the basis of criteria that have nothing to do with exposure, and the inferences in an epidemiologic study should relate to the general causal theory rather than what happened to any single person.

One way in which this problem can get out of hand is if a disease is defined in terms of an exposure. Once that occurs, a valid epidemiologic evaluation may be impossible. Consider the example of "analgesic nephropathy." This "disease" refers to kidney failure that is supposedly induced by the effect of analgesic drugs, based on the theory that analgesic drugs cause kidney failure in some people. Although there may be no reason to doubt the theory, if it is applied by defining cases of analgesic nephropathy to be kidney failure in people who have taken analgesics for a specified time, it will be impossible to evaluate epidemiologically the relation of analgesics to kidney failure. A valid evaluation would require that

kidney failure be defined and diagnosed on the basis of disease-related criteria alone, with information about analgesic use excluded from the disease definition and diagnosis. Even if the disease is not called analgesic nephropathy, as long as the information on analgesic use is taken into account in making the diagnosis, an epidemiologic evaluation of the relation between analgesics and kidney failure will be biased.

#### WHEN THE DISEASE DEFINITION INCLUDES AN EXPOSURE

It is not only in the epidemiologic study of drugs that one encounters disease definitions that refer to exposures. If a clear understanding of a causal relation exists, it is a natural tendency to refine the definition of disease to reflect this insight. On the other hand, if the "insight" is only a presumption that a researcher would like to study, it is essential to apply disease definitions that are independent of the exposure. The following is a list of some examples of diseases defined on the basis of an exposure. (Most infectious diseases, such as syphilis, malaria, and influenza, could also be included.)

- Analgesic nephropathy
- Asbestosis
- Berylliosis
- Food poisoning
- Frostbite
- Heatstroke
- Hypervitaminosis D
- Iron-deficiency anemia
- Motion sickness
- Protein-calorie malnutrition
- Radiation sickness
- Silicosis
- Smoker's cough
- Strep throat
- Tennis elbow
- Tuberculosis

Much of the work in pharmacoepidemiology today is conducted using health databases, which allow investigators to design studies from computerized files that include information on drug prescriptions, demographic factors, and health data from medical records or from claims that deal with reimbursement. The Boston Collaborative Drug Surveillance Program was a pioneering effort in pharmacoepidemiology, starting with hospital-based interviews of inpatients using nurse monitors.<sup>15</sup> As the medical world gradually became more computerized, this work and that of other pharmacoepidemiologists evolved to use data that were already entered into computers as part of the record-keeping system, such as in some private prepaid health plans in the United States and in governmental plans such as that of the province of Saskatchewan in Canada. Pharmacoepidemiology is now an active field of research that has established itself as a separate specialty area with its own textbooks.<sup>16,17</sup>



## HEALTH OUTCOMES RESEARCH

Health outcomes research and the related field of pharmacoeconomics are comparatively new research areas with lofty goals. Randomized trials and other medical research studies typically focus on a primary end point, such as survival or disease recurrence. Therapeutic evaluations based on narrowly defined end points have been subject to the criticism that they do not adequately take into account the overall quality of life that patients face based on the combination of therapeutic outcomes and unintended effects that a given treatment produces. Furthermore, classic therapeutic research typically does not take into account the economic costs of different therapeutic options. The economic costs are borne either directly by the patient or insurers or indirectly by the government and thus by society as a whole. In either case, there is strong motivation to find therapies that offer desirable results for patients at costs that are attractive to patients or society relative to the therapeutic alternatives. These are the goals that health outcomes research and pharmacoeconomics address, using methods such as meta-analysis, cost-effectiveness analysis, decision analysis, and sensitivity analysis in addition to more traditional epidemiologic methods. The interested reader should consult the text by Petitti<sup>18</sup> for a comprehensive overview.

## QUESTIONS

1. Predictive value depends on disease prevalence, but sensitivity and specificity do not. What might cause the sensitivity and specificity of a test to vary from one population to another?
2. Suppose that you wished to conduct a prospective cohort study to evaluate the benefits of prostate-specific antigen testing as a screening tool for prostate cancer. What outcome would most interest you? What biases would affect the study results? Would these biases also affect the results of a randomized trial?
3. Because everyone eventually dies, why would we not say that the case-fatality rate among patients with any disease is 100%?
4. Under what conditions might one find that the baseline difference in a variable in a clinical trial is "statistically significant" but, nevertheless, not confounding? Under what conditions might we find that the baseline difference is not "statistically significant" but, nevertheless, is confounding?
5. The Alzheimer's disease cooperative trial manifested confounding by MMSE score. If the trial were repeated, would you expect that this same risk factor would be confounding again?
6. Equipoise is a state of genuine uncertainty as to which of two treatments is better. Ethicists consider equipoise to be an ethical requirement for

conducting a randomized therapeutic trial: if the researcher is already of the view that one treatment is better than the other, it would be unethical for that researcher to assign patients to the treatment that he or she believes is inferior. Under what conditions can equipoise be achieved in a placebo-controlled trial?

## REFERENCES

1. Catanzaro A, Perry S, Clarridge JE, et al. The role of clinical suspicion in evaluating a new diagnostic test for active tuberculosis: results of a multicenter prospective trial. *JAMA*. 2000;283:639-645.
2. Manos MM, Kinney WK, Hurley LB, et al. Identifying women with cervical neoplasia: using human papillomavirus DNA testing for equivocal Papanicolaou results. *JAMA*. 1999;281:1605-1610.
3. Cole P, Morrison AS. Basic issues in population screening for cancer. *J Natl Cancer Inst*. 1980;64:1263-1272.
4. Chin JE, ed. *Control of Communicable Diseases Manual*. 17th ed. Washington, DC: American Public Health Association; 2000.
5. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008:42-43.
6. Heart Outcomes Prevention Evaluation Study Investigators. Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. *N Engl J Med*. 2000;342:145-153.
7. Shapiro AK, Shapiro E. The placebo: is it much ado about nothing? In: Harrington A, ed. *The Placebo Effect: An Interdisciplinary Exploration*. Cambridge, MA: Harvard University Press; 1997:19.
8. World Medical Association. *Declaration of Helsinki*. <http://www.wma.net/en/30publications/10policies/b3/>. Accessed October 21, 2011.
9. Rothman KJ, Michels KB. The continuing unethical use of placebo controls. *N Engl J Med*. 1994;331:394-398.
10. University Group Diabetes Program. A study of the effects of hypoglycemic agents on vascular complications in patients with adult onset diabetes. *Diabetes*. 1970;19(suppl 2):747-830.
11. Rothman KJ. Epidemiologic methods in clinical trials. *Cancer*. 1977;39:1771-1775.
12. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials*. 3rd ed. St. Louis, MO: Mosby; 1996:297-302.
13. Sano M, Ernesto C, Thomas RG, et al. A controlled trial of selegiline, alpha-tocopherol, or both as treatment for Alzheimer's disease. *N Engl J Med*. 1997;336:1216-1222.
14. Pincus MM. Alpha-tocopherol and Alzheimer's disease [Letter to the editor]. *N Engl J Med*. 1997;337-572.
15. Allen MD, Greenblatt DJ. Role of nurse and pharmacist monitors in the Boston Collaborative Drug Surveillance Program. *Drug Intell Clin Pharm*. 1975;9:648-654.
16. Hartzema AG, Porta MS, Tilson HH. *Pharmacoepidemiology: An Introduction*. Cincinnati, OH: Harvey Whitney Books; 1998.
17. Strom BL. *Pharmacoepidemiology*. New York: John Wiley & Sons; 2000.
18. Petitti DB. *Meta-analysis, Decision Analysis and Cost-effectiveness Analysis*. New York: Oxford University Press; 2000.



Appendix P VALUES CORRESPONDING TO VALUES OF THE STANDARD NORMAL DISTRIBUTION ( $\chi$  OR  $Z$ ) RANGING FROM 0.00 TO 3.99

$\chi$  Value in Hundredths

$\chi$ Value in Tenths	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	1.000000	0.992021	0.984043	0.976067	0.968093	0.960122	0.952156	0.944194	0.936237	0.928287
0.1	0.920344	0.912409	0.904483	0.896566	0.888660	0.880765	0.872881	0.865010	0.857152	0.849309
0.2	0.841480	0.833668	0.825871	0.818092	0.810330	0.802587	0.794864	0.787160	0.779477	0.771816
0.3	0.764177	0.756561	0.748968	0.741400	0.733856	0.726339	0.718847	0.711382	0.703945	0.696536
0.4	0.689156	0.681806	0.674485	0.667196	0.659937	0.652710	0.645516	0.638355	0.631227	0.624134
0.5	0.617075	0.610051	0.603063	0.596112	0.589197	0.582319	0.575479	0.568678	0.561914	0.555190
0.6	0.548506	0.541862	0.535258	0.528694	0.522172	0.515692	0.509254	0.502858	0.496504	0.490194
0.7	0.483927	0.477704	0.471525	0.465390	0.459300	0.453254	0.447254	0.441300	0.435391	0.429528
0.8	0.423711	0.417940	0.412216	0.406539	0.400908	0.395325	0.389789	0.384300	0.378859	0.373466
0.9	0.368120	0.362822	0.357572	0.352371	0.347217	0.342112	0.337055	0.332046	0.327086	0.322174
1.0	0.317310	0.312495	0.307728	0.303010	0.298340	0.293718	0.289144	0.284619	0.280142	0.275713
1.1	0.271332	0.266999	0.262714	0.258476	0.254286	0.250144	0.246048	0.242001	0.238000	0.234046
1.2	0.230139	0.226279	0.222465	0.218697	0.214975	0.211299	0.207669	0.204084	0.200545	0.197050
1.3	0.193601	0.190196	0.186835	0.183518	0.180245	0.177016	0.173830	0.170687	0.167586	0.164528
1.4	0.161513	0.158539	0.155607	0.152717	0.149867	0.147058	0.144290	0.141561	0.138873	0.136224
1.5	0.133614	0.131043	0.128511	0.126016	0.123560	0.121141	0.118760	0.116415	0.114106	0.111834
1.6	0.109598	0.107398	0.105232	0.103101	0.101005	0.098943	0.096914	0.094919	0.092957	0.091028
1.7	0.089131	0.087266	0.085432	0.083630	0.081859	0.080118	0.078407	0.076727	0.075076	0.073454
1.8	0.071860	0.070295	0.068759	0.067250	0.065768	0.064313	0.062885	0.061483	0.060108	0.058758
1.9	0.057433	0.056133	0.054858	0.053606	0.052379	0.051176	0.049995	0.048838	0.047703	0.046591
2.0	0.045500	0.044431	0.043383	0.042356	0.041350	0.040364	0.039398	0.038452	0.037525	0.036617
2.1	0.035728	0.034858	0.034006	0.033171	0.032354	0.031555	0.030772	0.030006	0.029257	0.028524
2.2	0.027806	0.027105	0.026418	0.025747	0.025090	0.024449	0.023821	0.023207	0.022607	0.022021

2.3	0.021448	0.020888	0.020340	0.019806	0.019283	0.018773	0.018274	0.017788	0.017312	0.016848
2.4	0.016395	0.015952	0.015520	0.015098	0.014687	0.014285	0.013893	0.013511	0.013138	0.012774
2.5	0.012419	0.012073	0.011735	0.011406	0.011085	0.010772	0.010467	0.010170	0.009880	0.009597
2.6	0.009322	0.009054	0.008793	0.008538	0.008290	0.008049	0.007814	0.007585	0.007362	0.007145
2.7	0.006934	0.006728	0.006528	0.006333	0.006144	0.005959	0.005780	0.005605	0.005436	0.005270
2.8	0.005110	0.004954	0.004802	0.004654	0.004511	0.004372	0.004236	0.004104	0.003976	0.003852
2.9	0.003731	0.003614	0.003500	0.003389	0.003282	0.003177	0.003076	0.002978	0.002882	0.002789
3.0	0.002699	0.002612	0.002527	0.002445	0.002365	0.002288	0.002213	0.002140	0.002070	0.002001
3.1	0.001935	0.001870	0.001808	0.001748	0.001689	0.001632	0.001577	0.001524	0.001472	0.001422
3.2	0.001374	0.001327	0.001282	0.001238	0.001195	0.001154	0.001114	0.001075	0.001038	0.001002
3.3	0.000966	0.000933	0.000900	0.000868	0.000837	0.000808	0.000779	0.000751	0.000724	0.000698
3.4	0.000674	0.000649	0.000626	0.000603	0.000581	0.000560	0.000540	0.000520	0.000501	0.000483
3.5	0.000465	0.000448	0.000431	0.000415	0.000400	0.000385	0.000370	0.000357	0.000343	0.000330
3.6	0.000318	0.000306	0.000294	0.000283	0.000272	0.000262	0.000252	0.000242	0.000233	0.000224
3.7	0.000215	0.000207	0.000199	0.000191	0.000184	0.000176	0.000170	0.000163	0.000156	0.000150
3.8	0.000144	0.000139	0.000133	0.000128	0.000123	0.000118	0.000113	0.000108	0.000104	0.000100
3.9	0.000096	0.000092	0.000088	0.000084	0.000081	0.000078	0.000074	0.000072	0.000068	0.000066