CHAPTER 19

Randomized Clinical Trials

19.1 INTRODUCTION

If Alexander Pope is correct that "the proper study of mankind is man" [Pope, 1733], then the development of new therapeutic and prophylactic measures for humans is one of the more proper uses of biostatistics. In addition, it is one of the most active and highly used areas of biostatistics. In this chapter we consider primarily randomized clinical trials in humans, although we mention other uses of the techniques. The use of *clinical* refers to the evaluation of clinical measures, for example, drug treatments or surgical treatments. If an experiment is randomized—that is, treatment assignments given by some random process—it necessarily implies more than one treatment is being considered or tested. Thus, the trials are comparative. And, of course, the term *trial* means that we test, or try, the treatments considered. The acronym RCT has been used for both a randomized *controlled* trial and a randomized *clinical* trial. Randomized clinical trials are examples of randomized controlled trials, but not necessarily vice versa, as we shall see below. Here we use the abbreviation RCT for both. For the most part we shall be discussing clinical trials, although it will be clear from the context which is referred to.

In addition to the statistical methods we have discussed before there are a number of practical issues in clinical trials that are now accepted as appropriate for the best scientific inference. The issues of trial design to some extent "fall between the cracks" in clinical research. They are not an obvious part of a medical education—not being biological per se—and also not an obvious portion of biostatistics, as they do not explicitly involve the mathematics of probability and statistics. However, the issues are important to successful implementation of good scientific clinical studies (and other studies as well) and are a necessary and appropriate part of biostatistical training. Some of these issues are discussed in less detail in Chapter 2 and in Chapter 8, in which we discuss permutation and randomization tests in Section *8.9. Here we give background on why the design features are needed as well as some discussion of how to implement the design features.

The use of RCTs and new drug development is big business. At the end of 2001, the cost for evaluating an approved new chemical entity was estimated at approximately $800 million [*Wall Street Journal*, 2001], and the time for development is often 10 years or more.

19.2 ETHICS OF EXPERIMENTATION IN HUMANS

The idea of experimenting on humans and other animals is distasteful at first blush. This is especially so in light of the Nazi experiments during the World War II period (see, e.g., Lifton [1986]).

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

Yet it is clear that if new and improved therapies and treatments are to be developed, they must be tried initially at some point in time on humans and/or animals. Whether designated so or not, such use does constitute experimentation. This being the case, it seems best to acknowledge this fact and to try to make such experiments as appropriate, justified, and useful as possible. Considerable work has been devoted to this end. The ethics of experimentation on humans has been the subject of intense study in recent decades. Ethics was touched on in Section 2.5, and because of its importance, we return to the subject here. A good introduction is Beauchamp and Childress [2001]. They review four principles for biomedical ethics: respect for autonomy, nonmaleficence, beneficence, and justice. Briefly summarized:

- The *principle of autonomy* recognizes a person's right to "hold views, make choices, and take actions based on personal values and beliefs."
- The *principle of nonmaleficence* is not to inflict harm to others.
- The *principle of beneficence* "asserts an obligation to help others further their important and legitimate interests."
- The *principle of justice* is more difficult to characterize briefly and may mean different things to different people. As Beauchamp and Childress note: "The only principle common to all theories of justice is a minimal principle traditionally attributed to Aristotle: Equals must be treated equally, and un-equals must be treated unequally."

One of the cornerstones of modern clinical research is *informed consent* (consistent with the respect for autonomy). This seemingly simple concept is difficult and complex in application. Can someone near death truly give informed consent? Can prisoners truly give informed consent? Biologically, children are not small adults; drugs may have very different results with children. How can one get informed consent when studying children? Do parents or legal guardians really suffice? How can one do research in emergency settings with unconscious persons who need immediate treatment (e.g., in cardiac arrest)? Do people really understand what they are being told?

The issues have given rise to declarations by professional bodies (e.g., the Declaration of Helsinki, [World Medical Association, 1975], the Nuremberg Code [Reiser et al., 1947], and worldwide regulatory authorities (e.g., Federal Regulations [1988] on Institutional Review Boards). The Health Insurance Portability and Accountability Act (HIPAA) was passed by the U.S. Congress in 1996. The rules resulting from this act have been published and refined since that time. The revised final privacy rules were published in 2002. Much information is protected health information (PHI) and researchers in the United States need to be aware of these regulations and conform to the rules. In the United States, anyone involved in research on humans or animals needs to be familiar with the legal as well as the more general ethical requirements. Without a doubt there is great tension for medical personnel involved in research. Their mandate is to deliver the best possible care to their patients as well as to do good research. See Fisher [1998a] for a brief discussion and some references. In addition, some statistical professional societies have given ethical guidelines for statisticians [Royal Statistical Society, 1993; American Statistical Association, 1999].

All agree that ethical considerations must precede and take precedence over the science. What this means in practice can lead to legitimate differences of opinion. Further continuing scientific advances (such as genetics, cloning, or fetal research) bring up new and important issues that require a societal resolution of what constitutes ethical behavior.

19.3 OBSERVATIONAL AND EXPERIMENTAL STUDIES IN HUMANS

In this section we consider some reasons why randomized studies are usually required by law in the development of new drugs and biologics. Rather than a systematic development, we begin with a few examples and possible lessons to be learned from them.

Example 19.2. If taking a drug helps you survive, it must be effective! During a National Institutes of Health (NIH) Randomized Clinical Trial [Coronary Drug Project Research Group, 1980] a drug was found to have about half the mortality among those who took the drug (defined as taking 80% or more of the assigned medication) vs. those who did not take the drug consistently. The five-year mortality in the men with coronary heart disease was 15.1% of the "good adherers" to drug and 28.2% in the "poor adherers" to drug. Although it certainly seems that the drug is effective (after all counting bodies is not subject to bias), it is possible that those who were good adherers were different when the study started. Fortunately, this was an NIH study with excellent detailed data collected for the known risk factors in this population. There were some differences at baseline between the good and poor adherers. Thus, a multiple linear regression analysis of five-year mortality was run, adjusting for 40 baseline variables in the 2695 patients taking the drug.

The analysis adjusting for these 40 variables led to adjusted five-year mortality of 16.4% for good adherers vs. 25.8% for the poor adherers. This would seem to clearly indicate a survival

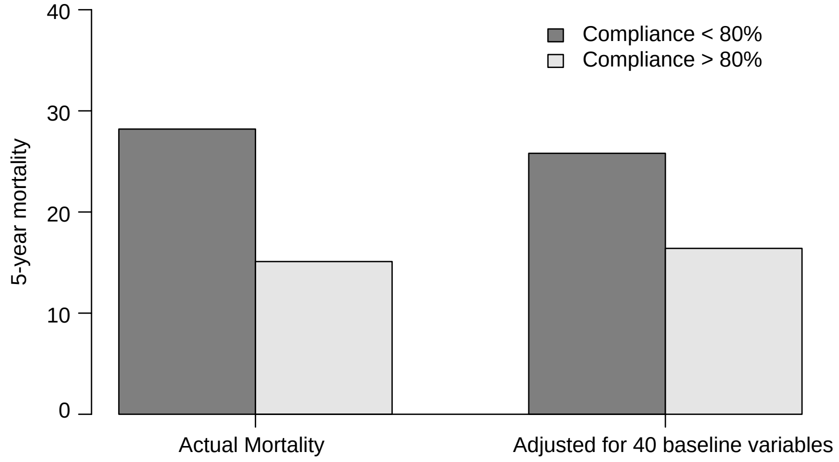

Figure 19.1 Five-year mortality among good and poor adherers to treatment.

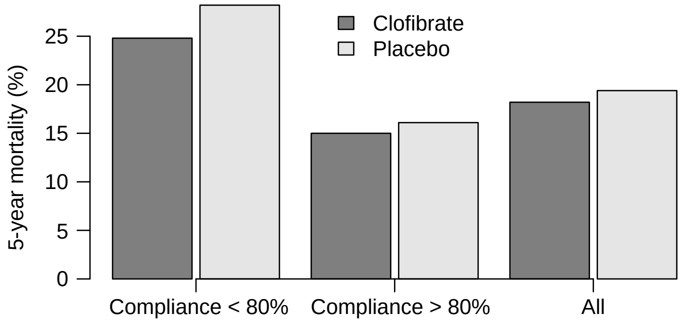

Figure 19.2 Five-year mortality by compliance and treatment in the Coronary Drug Project.

benefit of the drug—thus negating the need for a controlled study, although the data were collected for one arm of a controlled study. The only problem with this result is that the drug above was the placebo! In fact, the good and poor adherers of the active drug, clofibrate, had a very similar pattern. Figures 19.1 and 19.2 give the five-year mortality for the placebo arm of the trial and then for both arms of the trial. The two treatment arms did not differ statistically. The reason for the difference between the placebo mortality for the good and poor adherers was never fully understood.

Results such as this show how difficult it can be to assess a drug effect correctly from observational data. This is one reason why randomized clinical trials are the regulatory gold standard for most drug approvals. This is fine as far as it goes. We are then left with a very difficult consideration. Why, then, does this book give the majority of space to observational data analyses? If we cannot trust such analyses, why bother? The answer is that we do the best we can in any situation. If observational data analyses are the only practical method (due to cost or other feasibility factors), or the only ethical method (as the epidemiology of smoking risk became clear, it would not been considered ethical to randomize to smoking and nonsmoking treatment arms—not to mention the difficulty of execution), observational data must be used.

Example 19.3. If we stop the thing that appears to cause the deaths, we must be prolonging life (or are we?). One of the wonders of the body is our heart; it beats steadily minute after minute, year after year. If the average number of beats is 60 per minute, there are 86,400 beats/day or 31,536,000 beats/year. In a 65-year-old, the heart may have delivered over 2 billion heartbeats. The contraction of the heart muscle to force blood out into the body is triggered by electrical impulses that depolarize and thus contract the heart in a fixed pattern. As the heart muscle becomes damaged, there can be problems with the electrical trigger that leads to the contraction of the heart. The electrical changes in the heart are monitored when a physician takes an electrocardiogram (ECG) of the heart. If the depolarization starts inappropriately someplace other than the usual trigger point (the sinus atrial node), the heart can contract early; such a resulting irregular heartbeat, or arrhythmic beat, is called a *ventricular premature depolarization* (VPD). Although most people have occasional VPDs, after a heart attack or myocardial infarction (MI), patients may have many more VPDs and complex patterns of irregular heart beats, called *arrhythmias*. The VPDs place patients at an increased risk of sudden cardiac death. To monitor the electrical activity of the heart over longer time periods, ambulatory electrocardiographic monitors (AECGMs) may be used. These units, also called *Holter monitors*, measure and record the electrical activity of the heart over approximately 24-hour periods. In this way, patients' arrhythmic patterns may be monitored over time. Patients have suffered sudden cardiac death, or sudden death, while wearing these monitors, and the electrical sequence of events is usually the following: Patients experience numerous VPDs and then a run of VPDs that occur rapidly in succession (say, at a rate greater than or equal to 120 beats/min); the runs are called *ventricular tachycardia* (VT). Now many coronary patients have runs of VT; however, before death, the VT leads to rapid, irregular, continuous electrical activity of the heart called *ventricular fibrillation* (VF). Observed in a cardiac operation, VF is a fluttering, or quivering, of the heart. This irregular activity interrupts the blood flow and the patient blacks out and if not resuscitated, invariably dies. In hospital monitoring settings and cities with emergency rescue systems, the institution of *cardiopulmonary resuscitation* (CPR) has led to the misnomer of *sudden death survivors*. In a hospital setting and when emergency vehicles arrive, electrical defibrillation with paddles that transmit an electrical shock is used. Individuals with high VPD counts on AECGMs are known to be at increased risk of sudden death, with the risk increasing with the amount and type of arrhythmia.

This being the case, it was natural to try to find drugs that reduced, or even abolished, the arrhythmia in many or most patients. A number of such compounds have been developed. In patients with severe life-threatening arrhythmia, if an antiarrhythmic drug can be found that controls the arrhythmia, the survival is greatly superior to the survival if the arrhythmia cannot be controlled [Graboys et al., 1982]. Graboys and colleagues examined the survival of patients with severe arrhythmia defined as VF (outside the period of an MI) or VT that compromised the blood flow of the heart to the degree that the patients were symptomatic. Figure 19.3 gives the survival from cardiac deaths in 98 patients with the arrhythmia controlled and 25 patients in whom the arrhythmia was not controlled.

Thus, there was a very compelling biological scenario. Arrhythmia leads to runs of VT, which leads to VF and sudden death. Drugs were developed, and could be evaluated using AECGMs, that reduced the amount of arrhythmia and even abolished arrhythmia on AECGMs in many patients. Thus, these people with the reduced or abolished arrhythmia should live longer. One would then rely on the *surrogate endpoint* of the arrhythmia evaluation from an AECGM. A surrogate endpoint is a measurement or event that is thought to be closely associated with the real endpoint of interest such that inducing changes in the surrogate endpoint would imply similar changes in the "real" endpoint of interest. Usually, the surrogate endpoint is a measurement or event that is not of direct benefit to a patient or subject, but that is presumably related to direct benefit and can be used to establish benefit. Prentice [1989] defines the issue statistically: "I define a surrogate endpoint to be *a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint*." Antiarrhythmic drugs were approved by the U.S. Food and Drug Administration (FDA) based on this surrogate endpoint. It is important to point out that antiarrhythmic drugs may have other benefits than preventing sudden death.

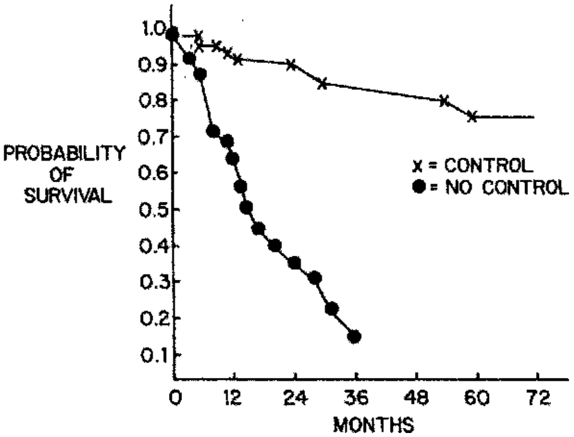

Figure 19.3 Survival free 17 cardiac mortality in patients with severe arrhythmia. The curves are for those whose arrhythmia was controlled by antiarrhythmic drugs and for those in whom the arrhythmia was not controlled by antiarrhythmic drugs.

For example, some patients have such severe runs of VT that they faint. Prevention of fainting spells is of direct benefit to the patient. However, asymptomatic or mildly symptomatic patients with arrhythmia were being prescribed antiarrhythmics with the faith(?), hope(?) that the drugs would prolong their life.

Why, then, would anyone want to perform a randomized survival trial in patients with arrhythmia? How could one perform such a trial ethically? There were a number of reasons: (1) the patients for whom arrhythmia could be controlled by drugs have selected themselves out as biologically different; thus, the survival *even without antiarrhythmic therapy* might naturally be much better than patients for whom no drug worked. That is, modification of the surrogate endpoint of arrhythmia had never been shown to improve the results of the real endpoint of interest (sudden death). (2) Some trials had disturbing results, with adverse trends in mortality on antiarrhythmic drugs [IMPACT Research Group, 1984; Furberg, 1983]. (3) All antiarrhythmic drugs actually produce more arrhythmia in some patients, a *proarrhythmic effect*.

The National Heart, Lung and Blood Institute decided to study the survival benefit of antiarrhythmic drugs in survivors of a myocardial infarction (MI). The study began with a pilot phase to see if antiarrhythmic drugs could be found that reduced arrhythmia by a satisfactory amount. If this could be done, the randomized survival trial would begin. The first study, by the Cardiac Arrhythmia Pilot Study (CAPS) Investigators [1988], showed that three of the drugs studied—encainide, flecainide, and moricizine—suppressed arrhythmias adequately to allow proceeding with the primary survival trial, the Cardiac Arrhythmia Suppression Trial (CAST). Patients within six weeks to two years of an MI needed six VPDs per hour to be eligible for the study. There was an open label, dose titration period where drugs were required to reduce VPDs by at least 80% and runs of VT by at least 90%. (For more detail, see the Cardiac Arrhythmia Suppression Trial (CAST) Investigators [1989] and Echt et al. [1991].) Patients for whom an effective drug was found were then randomized to placebo or to the effective drug (Figure 19.4). Such was the confidence of the investigators that the drugs at least were doing no harm that the test statistic was one-sided to stop for a drug benefit at the 0.025 significance level. The trial was not envisioned as stopping early for excess mortality in the antiarrhythmic drug groups.

The first results to appear were a tremendous shock to the cardiology community. The encainide and flecainide arms were dropped from the study because of excess mortality! Strictly speaking, the investigators could not conclude this with their one-sided design. However, the

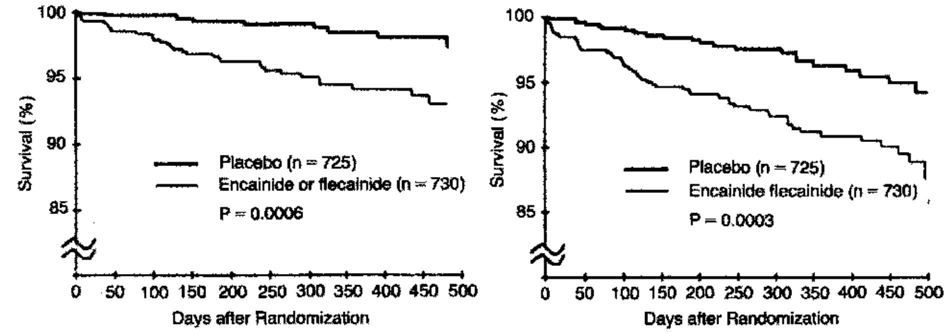

Figure 19.4 The panel on the left shows the survival, free of an arrhythmic death, among 1455 patients randomized to either placebo or one of encainide or flecainide. The second panel is based on all-cause mortality. (From the Cardiac Arrhythmia Suppression Trial (CAST) Investigators [1989].)

evidence was so strong that the investigators, and almost everyone else, were convinced of the harmful effects of these two antiarrhythmic drugs as used in this patient population.

The results of the study have been addressed by Pratt et al. [1990] and Pratt [1990]; the timing of the announcement of the results is described in Bigger [1990]; this paper gives a feeling for the ethical pressure of quickly promulgating the results. Ruskin [1989] conveys some of the impact of the trial results: "The preliminary results ... have astounded most observers and challenge much of the conventional wisdom about antiarrhythmic drugs and some of the arrhythmias they are used to treat. ... Although its basis is not entirely clear, this unexpected outcome is best explained as the result of the induction of lethal ventricular arrhythmias (i.e., a proarrhythmic effect) by encainide and flecainide."

This trial has saved, and will continue to save lives by virtue of changed physician behavior. In addition, it clearly illustrates that consistent, plausible theories and changes in surrogate endpoints cannot be used to replace trials involving the endpoints of importance to the patient, at least not initially. Finally, it is important to note that one should not overextrapolate the results of a trial; the study does not apply directly to patients with characteristics other than those in the trial; it does not imply that other antiarrhythmic drugs have the same effect in this population. However, it does make one more suspicious about the role of antiarrhythmic therapy, with a resulting need for even more well-controlled randomized data for other patient populations and/or drugs.

The trial illustrates the difficulty of relying on very plausible biological theories to generate new drug therapy. New therapies should be tested systematically in a controlled fashion on humans following ethical guidelines and laws. Note also that the arrhythmia itself is not the true focus of the therapy. It was thought to be a good "surrogate" for survival. The use of surrogate endpoints as a guide to approving new therapies is very risky, as the example shows [Temple, 1995; Fleming and DeMets, 1996].

Example 19.4. Epidemiological studies have shown that higher than normal blood pressure in humans is associated with shorter life span [Kesteloot and Joosens, 1980]. The decrease is due especially to increased cardiovascular events, such as a heart attack, stroke, or sudden death due to arrhythmia. Early clinical trials showed that lowering blood pressure by drug therapy resulted in fewer heart attacks, strokes, and cardiovascular deaths. Subsequently, it was considered unethical to treat persons with high blood pressure, called *hypertensive individuals*, with a placebo or sham treatment for a long period of time. Thus blood pressure–lowering drugs, *antihypertensive drugs*, were studied for relatively short periods, six to 12 weeks, in subjects with mild to moderate hypertension. The surrogate endpoint of blood pressure reduction is used for approval of antihypertensive drugs. As blood pressure tends to rise with physical or emotional stress it is subject to change in response to subtle clues in the environment. For this reason trials use placebo (*inactive*) pills or capsules that are in appearance, smell, and so on, the same as tested *active treatment* pills or capsules, as discussed in Chapter 1. In addition, to prevent the transmission of clues that might affect blood pressure, the subject is not informed if she or he is taking the active drug or the placebo drug. If only the subject does not know the treatment, the trial is a *single-blind trial*.

However, since subtle clues by those treating and/or evaluating the subjects could affect blood pressure, those treating and/or evaluating the subjects also are not told if the subject is getting the active or placebo treatment. A study with both the subject and medical personnel blinded is called a *double-blind study*.

At the beginning of the study, subjects are usually all started on placebo during an initial single-blind period. This period serves multiple purposes: (1) it allows the effect of prior therapy to *wash out* or disappear; (2) it allows identification of subjects who will take their medication to be used in the comparative part of the trial; (3) it lessens the effect of raised blood pressure due to the unsettling medical setting (the *white coat hypertension* effect); (4) it helps to remove a regression to the mean effect of patient selection; and (5) multiple readings can assure relative stability of and measurement of the baseline blood pressure.

Figure 19.5 shows the data of the placebo arm in such a trial. Since subjects were on placebo the entire time, the explanation for the stable mean pressure during the single-blind *run-in period* and the drop during the double-blind portion of the trial was thought to be subtle clues being given to the patients by the medical personnel when they knew that some patients would be getting active therapy. It should be emphasized that subjects were never told in the single-blind portion of the trial that they were not potentially receiving active therapy. (The subjects did sign an informed consent and knew that they might receive placebo or active therapy during portions of the trial.) This figure illustrates the need for blinding in some clinical trials.

Figure 19.6 shows data from a second trial of an antihypertensive drug. The trial was a dose escalation study. That is, the dose of a drug was increased in individual patients until they had a satisfactory blood pressure response. Again the data are from the placebo arm of the trial. The increasing "benefit" observed as the "dose" of placebo escalates illustrates the need for a control group.

Example 19.5. In the United States, the National Institutes of Health (NIH) administers most federal funds for health sciences research as well as having its own (intramural) programs of research. Most of its employees thus value and are aware of the importance of well-conducted medical research. Thus, the NIH population would seem the ideal place to study an intervention

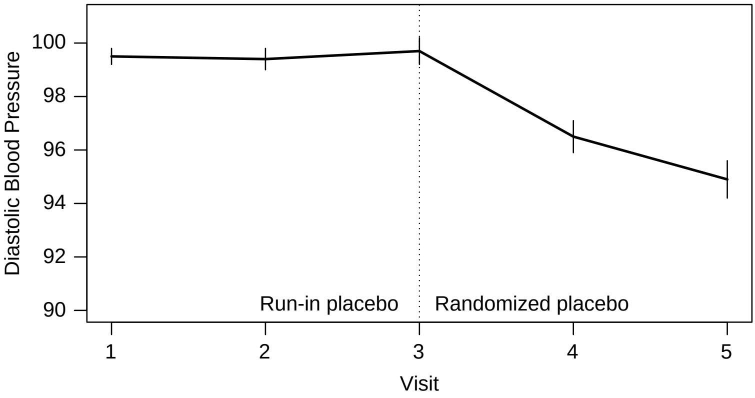

Figure 19.5 Average diastolic blood pressure (±1 standard error) during single-blind run-in and double-blind treatment with placebo.

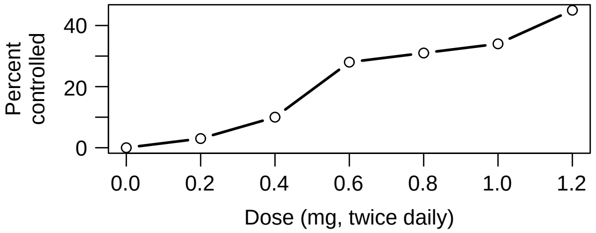

Figure 19.6 Response to escalating doses of placebo antihypertensive.

if there were sufficient numbers of NIH employees experiencing the malady in question. The results of a study on the use of ascorbic acid (vitamin C) for the common cold were published by Karlowski et al. [1975]. Most aspects of the study will not be presented here, in order to concentrate on the difficulty of performing a good experiment. There were four groups in the study. As a preventive (prophylactic) measure there was random assignment to either ascorbic acid or placebo (with capsules containing the study medication), and when a cold was thought to occur (with a clear definition), the study participants were assigned at random (the same for all colds if multiple colds occurred) to either ascorbic acid or placebo. Thus, there were four groups. Three hundred and eleven persons were randomized to therapy (discounting 12 subjects who dropped out early "before taking an appreciable number of capsules"). During the study the investigators learned that some subjects had opened the capsules and tasted the contents to see if they were taking ascorbic acid or placebo. More prophylactic placebo subjects (69) dropped out than ascorbic acid prophylactic subjects (52). At the end of the study the investigators queried the subjects about whether they thought they knew their study drug; of 102 subjects who thought they knew, 79 (77%) guessed correctly. The study results showed no statistical difference in the number of colds, but there was a trend for less severity of a cold if one took ascorbic acid. Unfortunately, this trend disappeared if one took into account those who knew their therapy. The NIH investigators comment under the heading the *power of suggestion*: "Depending upon one's point of view, it is either an unfortunate or fortunate aspect of the study. It would have been gratifying to have performed a flawless clinical trial; on the other hand, it has turned out to be a unique opportunity to gain some insight into the importance of perfect blinding in trials with subjective endpoints. An association between severity and duration of symptoms and knowledge of the medication taken seems to have been clearly established."

These examples above illustrate:

1. The need for a control group to be compared with an active therapy
2. The need for a "fair" or unbiased control, or comparison, group or appropriate mathematical adjustment to make a fair comparison. Appropriate mathematical adjustment is very difficult to do in this setting (as Example 19.2 illustrates)
3. The need for blinding to avoid introducing bias into clinical trials
4. The need for an endpoint of a trial that has clinical relevance (e.g., Temple [1995])

19.4 OBTAINING A FAIR OR UNBIASED COMPARISON: RANDOMIZED CLINICAL TRIAL

We now turn to two aspects of the clinical trial. The first is summarized by the question: How can we assign subjects to unbiased, or comparable, groups at the start of a clinical trial? The idea of random selection to get a "fair" choice or comparison goes back a long time in human history. Lots were used in Old Testament times, the idea of "drawing the short straw," taking a card from a well-shuffled pack, and so on, all show the intuitive appeal of this type of procedure. However, the formal introduction of *randomization* was made in the 1930s by the British statistician and geneticist Sir Ronald Aylmer Fisher [Box, 1978]. In one of the great intellectual advances of the twentieth century, he combined the methodology of probability theory with the intuitive appeal of randomization to begin the *randomized experiment*. The idea is in some ways counterintuitive. As seen previously in this book, a theme of good observational data analysis or experimentation is to eliminate variability in order to make comparisons as precise as possible. Randomness, or "unexplained noise," does just the opposite. Think of the simplest type of random assignment between two treatments: Each eligible patient has her or his therapy determined (after informed consent) by the flip of an unbiased coin (i.e., the probability of each treatment is $\frac{1}{2}$). The different flips are statistically independent, and if there are n assignments, the number on treatment A, or B for that matter, is a binomial variable. Further, any particular pattern of assignment is equally likely ($1/2^n$).

What are the benefits of this random assignment? First, the assignment to treatment is fair. Human biases, whether conscious or unconscious, are eliminated. Second, on average, the two assignments have the same number of easy or difficult-to-treat assignments; that is, patient characteristics are balanced (statistically). Third, if we assume that treatment is unrelated to our outcome, we can assume that the outcomes were preordained to be good or bad. We can find the probability under this random assignment that each treatment arm had outcomes as extreme or more extreme than that actually observed with the actual assignments because we know that each assignment of cases is equally likely. (see Chapter 8). That is, we can compute a p-value that is not dependent on assumptions about the population we are observing. This is called using the *randomization distribution* (see Edgington [1995]). We do, however, need to be sure that the randomization is done appropriately.

The benefits of the randomized trial are so widely recognized that by law and regulation, in most countries new drugs or biologics need to be evaluated by a randomized clinical trial in order to gain regulatory approval to market the new advance legally. See Note 19.4 for a few references on the need for and benefits of the RCT.

19.4.1 Intent to Treat

There are complications to RCTs in practice. Suppose, in fact, that many patients assigned to one, or both, of the treatments do *not* get the assigned therapy? Does it make sense to compare the treatments as randomized? How can patients who do not receive a therapy benefit from it? Thus, does it not seem odd to keep such patients in a comparison of two therapies? This sticking point has led to some difficult considerations: If we consider only patients who received their assigned or randomized therapy, we can introduce bias since those who do not receive their therapy are usually different (and unfortunately, possibly in unknown ways) from those who do receive their assigned therapy. The issue then becomes one of avoiding bias (include all patients who are randomized into their assigned group) vs. biological plausibility (only count those who actually receive a treatment). At its worst this might pit biostatisticians vs. clinicians. At this point in time, including all subjects in the analysis into the group to which they are randomized is considered standard; such analyses are called *intent-to-treat* (ITT) *analyses*. The name arises from the fact that under the randomized assignment there is an implied initial intent to treat the subject in the manner to which he or she was randomized. The best way to avoid the conflict between bias and biology is to perform an excellent experiment where those randomized to a treatment do receive the treatment. For this reason the assignment to randomization should be accomplished at the last possible moment.

If those subjects who do not begin treatment do so for reasons that cannot have been due to the randomized assignment (e.g., nonbreakable double blinding), the subjects who at least begin therapy can be included into the analysis with all the benefits of the randomization process listed above. Such analyses are called *modified intent to treat* (mITT) and are acceptable provided that one can be assured that the lack of therapeutic delivery *cannot* have been related to the treatment assignment. In practice, modified intent-to-treat analyses are often also called intent to treat.

19.4.2 Blinding

We have seen above that using a randomized assignment does not accomplish the full task of assuring a fair comparison. If the outcome is affected by biased behavior due to the treatment assignment, we can have misleading results despite the fact that the treatments were assigned at random. Bias can still ruin an RCT. We have seen this in both the blood pressure and vitamin C examples above. Wherever possible, double blinding should be used. The more subjective the endpoint, the more important blinding is to a trial. However, even with very "hard" endpoints that would not seem to need blinding (e.g., mortality), blinding can be important. The reason is that if the blinding is not effective, there may be treatment biases that change the way subjects in the assigned groups are treated (e.g., hospitalized, given other medications) and this may affect even hard endpoints such as mortality. It is difficult to blind in many trials [e.g., a drug may induce physiologic changes (in heart rate or blood pressure)] and those seeing and treating a patient may have reasonable guesses as to the therapy. Added steps can be taken. For example, those involved in evaluating a patient for outcome might be required to be different from those treating a patient. Often, outcomes for a trial are evaluated by an external classification committee to reduce bias in the determination of events.

19.4.3 Missing Data

Missing data are one of the most common and difficult issues in the analysis of RCTs. Even a modest discussion of the ways to approach and handle missing data in RTCs goes beyond the scope of this book. However, a few partial solutions, based on the concepts introduced in Section 10.5.2 and Chapter 18 are presented here.

The first and most important thing to understand is that there is no totally satisfactory method of dealing with the issue. The best course is not to have any missing data, but often, that wonderful counsel cannot possibly be implemented. For example, in studies performed in a population of street people with illicit drug use, complete data are virtually unknown if the study requires patient cooperation over a moderate length of time. Subjects simply disappear and are extremely difficult to find. Some turn up in jail or hospitals, but follow-up is difficult. It they are to return for follow-up visits, adherence can be quite low. What are those running such a trial, as well as the general society, with its interest in the outcome, to do? We do the best we can but realize that there will be many missing data. Another example: One studies treadmill walking time in a population of congestive heart failure patients. The primary study endpoint is the change in treadmill time from the baseline measurement to the final visit (at some fixed interval from the time the subject was randomized). Some subjects will die: How should their data be treated in the final analysis? Clearly, the missing information (the impossible final treadmill test) is not independent of patient status. This is known as *informative censoring*. Others may have their heart failure progress to a stage where it is too difficult to come in for the test or to perform the test. Other subjects may become discouraged and exercise their right to withdraw from the study. Others may go on vacation and not be around at the correct time for their evaluation. The possibilities go on and on.

First, one might assume that the missing data do not bias the conclusions and analyze only those who have all appropriate data. This is usually not an acceptable approach unless there are only minimal missing data. However, it is often used as an additional analysis. Data may also be "missing" for legitimate medical reasons. In a trial of blood-pressure-lowering medication, patients may present with greatly elevated blood pressures that require immediate, or perhaps after a week's delay, treatment with known effective drug or drugs. In many trials there are more such subjects in the placebo group. If their data are not taken into account, there is a bias against the active therapy. Further, their data at the end of the scheduled therapy period are not unbiased, as strong active therapy is used to lower blood pressure. In this case the endpoint used is the last observation on the assigned randomized therapy. In effect, the last observation is carried forward to the time for final evaluation. Not surprisingly, such analyses are called *last observation carried* *forward* (LOCF). This is often used as a method of analysis when the primary parameter of the study is collected at regularly scheduled visits. Sometimes the missing data are replaced by the mean of the known values for the study. In other cases, more sophisticated methods are used to estimate, *impute*, the missing values. Such strategies can be quite complex. For a discussion of the implications of different reasons that data are missing, the implications for missing data, and analysis methods, see Little and Rubin [2002] and Section 18.6.

If the data are extremely strong, a *worst-case analysis* can be used and an effect still established. For example, in a survival analysis study that is placebo controlled, the comparison to a new therapy, the worst case (for establishing the new therapy), would assume that placebo patients not observed for the full observation period lived to the end of the follow-up period and that those assigned to the new active therapy died immediately after the last time they were known to be alive before being lost to follow-up.

The robustness of the study data to the missing data is sometimes assessed with some type of *sensitivity analysis*. Such analyses make a variety of assumptions about the actual pattern of the missing data and see how extreme it must be to change the study results in some important manner.

19.5 PLANNING AN RCT

19.5.1 Selection of the Study Population

In clinical studies the selection of the study population is critical. The understanding of the drug, biologic, or device mechanism will suggest a population of subjects where efficacy is to be shown. Selection of the highest-risk population is often the most logical choice to demonstrate the effect; however, if only such subjects are studied, the approval for use will usually be limited to such subjects. This may limit use of the new treatment. As a result, this may narrow the range of subjects getting a benefit, as well as lowering the sales potential for the sponsor developing the new therapy.

19.5.2 Special Populations

Historically, many important special populations either were not investigated at all or had very limited data—despite the fact that any realistic appraisal of usage patterns would anticipate such use. For example, women of childbearing potential were avoided; in large part, this was to avoid law suits if there were any birth defects in the children conceived, developing, or born during or close to the trial. The expense of a lifetime of care was avoided by not studying such women. The most infamous example of a drug causing birth defects was thalidomide in Europe and the United States. Nevertheless, many medications are used by pregnant women. Over-the-counter (OTC) products are the most obvious; analgesics (i.e., pain relievers) are one clear class. Now the FDA strongly recommends, and sometimes requires, such studies. Another underevaluated population was children. One might think that from a pharmacological point of view, children are merely small adults and that smaller doses would clearly work if the drug worked in adults. Unfortunately, this idea is simply not true. Children differ in many important ways in addition to size, and care is needed in extrapolating adult results to children. Historically, minorities, especially African-Americans, had limited experimental results in drug development (except in obvious special cases such as sickle cell anemia). In part, this was related to limited access to health care. There are genetic differences in the way that drugs affect humans, and minorities are now studied more systematically. Often, some clinical sites in studies are selected to test a therapy on a more diverse population. The elderly were also underrepresented in RCTs. In part, this is because the elderly have more trouble showing up for clinic visits and complying with their therapy (as they may forget to take their medication). However, the elderly are a particularly important population to study because (1) they take many medications, and drug–drug interactions that cause trouble are more likely to occur in this population; (2) drugs are often metabolized in the liver, so poor liver function can cause problems (the elderly have more liver impairment); (3) elimination is often through the kidneys, and the elderly are more likely to have kidney problems; and (4) the changing world demography shows that a larger proportion of the world population will be elderly in the next few decades.

19.5.3 Multicenter Clinical Trials

Many clinical trials use multiple clinical centers to enroll patients or subjects. There are several reasons for this. The most obvious is the need to enroll many patients in a timely fashion. There are also other reasons, perhaps not as obvious. Most new drugs are developed to be registered (approved for marketing) in many markets around the world: the United States, the European Union, Japan, and Canada, among others. Thus, the studies often have clinical sites from around the world to aid in approval under the various regulatory authorities. Using "influential" physicians at different centers as investigating clinicians in the research program can also be an aid to marketing when approval is granted. Other benefits of using multiple clinics include (1) showing that there is a benefit in different settings, and (2) assessing therapy under a variety of concomitant medical therapeutic settings.

In addition to the benefits, there are numerous additional challenges to multicenter clinical trials. Standardization of treatment and data recording often require extensive education and monitoring. The randomization process needs to be available over a wide range of times if subjects are enrolled around the world. Forms and data collection may be complicated by the number of languages and cultures involved. Data are analyzed for clinical site heterogeneity in response; often, this is done for different delivery settings (e.g., North America, Europe, and the rest of the world). Security of data, monitoring of the raw data (often in clinical files), and investigator and staff training are all quite complicated.

19.5.4 Practical Aspects of Randomization

The process of randomizing subjects in an RCT involves choices. To simplify the discussion we consider only *two-arm trials*, but similar considerations can be used with more than two treatment arms. The simplest random allocation is a fair coin flip, allocating each subject to one arm or the other. (In practice, the "flips" are done using a *pseudorandom number generator* on a computer.) There are drawbacks to the coin-flip approach. If there are clinical sites, each enrolling a small number of subjects, a number of such sites may involve only one treatment. This makes it impossible to see the variability in treatment effect within such sites. Therefore, the randomization is done using randomized blocks. If the ratio of subjects randomized to each arm is to be the same, even-numbered blocks are used. If the size is $2n$, then among each $2n$ randomizations, n will be to one arm and n to the other. Potentially, this can lead to bias, since if the study is unblinded or one can unblind with a reasonable probability, the probabilities for subsequent patients is no longer $\frac{1}{2}$ to $\frac{1}{2}$. To see this, consider an unblinded study: If we know the first $2n - 1$ treatment assignments, we know what the next subject will receive as a treatment. To get around this problem partially, blocks of different size are sometimes used, being chosen with some probability. For example, one might choose a block of size 4 half the time and a block of size 6 half the time.

Often, the blocks are not used to get balance within a site. If there is an important factor that determines the risk of the trial outcome, blocks with some strata for the risk factor may be used. This "forces" some balance with respect to the important prognostic factor. If more than one factor exists, combinations of two or more factors might be used. There is a limitation, however; if one had five factors, each of which had three levels, and we took all combinations, there would be $5^3 = 125$ possible strata. As the number goes up, we tend to get cells with zero or one subject actually randomized within a cell. When we are using only the first element of each block, randomization is the same as if we did not block at all! For this reason, more complex schemes have been developed for forcing balance on a number of factors; this technique is known as *adaptive randomization*. For blocking and adaptive randomization, one needs to know selected information about a subject before an assignment can be given. This is often done through either an interactive voice randomization system that uses touchtone phones or through the Internet. In either case, the needed information will be entered, eligibility may be checked, and the database is quickly informed of the randomization, and may check for subsequently expected data. See Efron [1971], Friedman et al. [1999, Chap. 5], or Meinert [1986, Sec. 10.2].

19.5.5 Data Management and Processing

Data management of randomized clinical trials is challenging, particularly so for international multicenter trials. In most instances, data are entered on *case report forms* (CRFs). Often, clinical sites are visited to compare the forms with the official medical records for consistency and documentation. Inspections are made by those sponsoring a study as well as by regulatory authorities if the trial aims to register a drug. Forms are usually submitted to a central data processing unit. They may be carried by hand using monitors, faxed after data entry at the clinical site (*remote data entry*), transferred electronically, entered via the Internet, or (more and more rarely) mailed in batches. To minimize data-entry errors, the data are often entered twice by two different people, and the entries compared for consistency with resolution in the case of disagreement. Entered files usually undergo extensive *consistency checks* [e.g., are the dates possible? Is a datum plausible (in that it is in a reasonable range)? If a discrete variable, is the code a legal one?] One of the worst errors is to have an incorrect patient identifier for a form or forms; for this reason, patient-identifying information (which only identifies the patient uniquely, not allowing the actual person to be identified) often has redundant checking information. When an entry fails a check, a process is instituted to resolve the problem. Tracking the resolution and any changes is documented for possible subsequent review. Problem resolution can be quite extensive and time consuming.

The database often allows identification of the timing of needed follow-up visits, examinations, or contact. For complex studies the database is sometimes used for notifying clinical sites of the expected upcoming data collection. Missing forms (i.e., those expected from subsequent visits) are asked for after some time interval. Some possible inconsistencies may arise externally (e.g., from a blinded committee used to classify endpoints that need resolution), and these are also tracked and recorded. Before a study is analyzed, or unblinded, all outstanding data issues are resolved to the extent possible, and the data file is then *frozen* for the analysis and interpretation of data. In studies that need ongoing monitoring for ethical reasons, there may be an independent *data and safety monitoring board* to review interim data (see Ellenberg et al. [2002]). To avoid introducing bias into the study, a group, independent of the sponsor, often provides tables, lists, and materials. The complexity and effort needed for such processes is hard to appreciate unless one has been through it. (See also Sections 2.6 to 2.9.)

19.6 ANALYSIS OF AN RCT

19.6.1 Preservation of the Validity of Type I Error

Because drug development costs so much and because the financial reward for a successful new drug in the right setting is so great, there is an apparent conflict between the sponsors and regulators. Stated statistically, the sponsors want to maximize the power of a study (i.e., minimize Type II error), and the regulators want to minimize and preserve the appropriateness and interpretability of the Type I error or p-value. Some areas of particular related concern are discussed below.

19.6.2 Interim Analysis of an Ongoing Clinical Trial

New investigational therapies hold potential for both benefit and harm. Experience has shown that no matter how thorough the prior work in other animal species, the results in humans may differ in unexpected ways. This is especially true with respect to adverse events. This requires looking at outcomes during the study—carrying out *interim analyses*. Similarly, when serious irreversible endpoints, such as death or permanent disability, are being considered, if a therapy is beneficial, there is an ethical requirement to stop the trial. But repeated interim analyses inflate the Type I error. This problem has been dealt with extensively in the biostatistical literature under the rubric of *sequential analysis*. Boundaries for values of a test statistic that would stop the trial at different times have been studied extensively (e.g., O'Brien and Fleming [1979]; Whitehead [1983]; Jennison and Turnbull [2000]; Lan and DeMets [1983]). In recent years, methods have been developed that allow examination of the results by treatment arm, with resulting modifications of the trial that still preserve the Type I error (e.g., Fisher [1998b]; Cui et al. [1999]). A basic strategy is to parcel out the Type I error over the trial. For example, suppose that two interim analyses are planned during the course of a study. Then test the results for the two interim analyses at the 0.001 level and the final analysis at the 0.048 level. This still ensures an overall level of 0.05.

19.6.3 Multiple Endpoints, Multivariate Endpoints, and Composite Endpoints

In some situations, multiple endpoints may be used to demonstrate the benefit of a new therapy. Of course, one cannot simply look at all of them and claim success if any one of them meets the significance level used in the RCT because the multiple comparisons inflate the Type I error. Several strategies have been used:

1. Select one of the possible beneficial endpoints to be the primary analysis for trial.
2. Adjust the p-value to account for the multiple comparisons. A conservative adjustment is to use the Bonferroni inequality and its refinements (Chapter 12) [Wright, 1992]. If the possible endpoints are positively correlated, as is usually the case, less severe adjustments are possible using the randomization distribution for the RCT.
3. The various components of possible endpoints can be considered to be a vector (i.e., arranged in sequence), and methods are available to test all the endpoints at once.
4. Sometimes an index, a weighted sum of the endpoints, is used as the one primary endpoint (see Schouten [2000]).
5. When a number of endpoints occur as distinct events in time, the first occurrence of any of them can be used as one event. Comparisons may be made using the methods of survival, or time to event, analysis (Chapter 16).

These issues are discussed in more detail in Chapter 12.

19.7 DRUG DEVELOPMENT PARADIGM

The following points introduce some of the ideas and terminology used in the development of drugs and biologics (see Mathieu [2002] for more). The first step is to identify a potential drug (a molecule). This used to be accomplished largely by chance (e.g., the discovery of penicillin) or through large screening programs, but because of recent substantial advances in genetics, molecular biology, and computer modeling, more and more compounds are being designed for specific purposes. Compounds may be screened for *in vitro* (i.e., "in glass") reaction with known molecules to identify candidates.

The first testing is carried out in several animal species. This *preclinical phase* of drug development accomplishes several purposes. Among the purposes are the following: The first is to identify if a drug is toxic at most possible doses (both short-term and longer-term studies in at least two species are done). Second, a range of doses can be evaluated. Are there doses that are not toxic (that have efficacy at the lower doses)? Third, use of an animal species will sometimes allow examination of an efficacy assessment vs. toxicity as a function of the dose. Other tasks performed are to look for the formation of fetal and birth abnormalities (*teratogenicity studies*), to see if drugs cause cancer (*carcinogenicity*, as a function of animal species and dose), and to see if gene abnormality results (*mutagenicity testing*). Of course, usually, the more drug one takes, the greater the amount that enters the body. One studies the time course of the drug [whether administered as a pill or capsule, by injection (intravenously or intramuscularly), by inhalation, etc.] within the body. Almost all drugs change into other molecules (metabolites) when in the body. Study of the time course of adsorption, distribution, metabolism, and elimination of the drug molecule and its metabolites comprises the field of drug *pharmacokinetics*. The relationship of the drug time–concentration value to the magnitude of effect is the field of *pharmacodynamics*.

After the preclinical data have been reviewed (in the United States) and approved by appropriate authorities, testing may begin in humans. *Phase I* of drug development is initial use of the drug in humans. Unfortunately, the preclinical animal testing gives only a rough idea of possible appropriate doses in humans. The animal data are often predictive only to an order of magnitude, so testing in humans usually begins at a very low dose and is slowly escalated. If the drug is anticipated to be well tolerated by normal subjects, the initial testing is usually done in healthy, normal volunteers. Drugs that are harmful by their nature [e.g., cancer (oncology) drugs that kill cells] are tested initially in patients. Some idea of activity may be gained in this initial phase. Slow escalation of the dose given helps to establish a preliminary dose range for the compound.

Phase II studies are reasonably large studies that give preliminary evidence of the efficacy of a drug in humans, to determine reasonable doses, and to get evidence on safety and tolerability in a patient population. These studies are often not blinded.

Phase III studies are large, randomized clinical trials to establish efficacy and safety. For most drugs it is expected that there will be at least two independent RCTs, double-blinded where possible, that establish efficacy at the 0.05 significance level. An increasing number of active control trials are being conducted in which noninferiority is established by showing that the new compound does not differ from the active control by more than a small equivalence margin (see, e.g., Temple and Ellenberg [2000]; Ellenberg and Temple [2000]). Often, the Phase III trials for efficacy do not provide adequate experience to evaluate patient safety. There often are *open label* (i.e., patient and physician know what treatment the subject is getting) extensions, where all patients get the new therapy if they consent to continue in the study. These trials may enroll more subjects, to get additional safety data.

After drugs are approved, *postmarketing*, or *Phase IV*, studies are sometimes performed for a variety of purposes: to collect more safety data, to do additional evaluation of efficacy (sometimes using a different endpoint), or to study efficacy in a broader, representative population.

19.8 SUMMARY

RCTs are difficult, expensive, ethically challenging, and require great attention to planning and monitoring operationally. Still the benefits are generally agreed to be worth the effort. This type of human experimentation gives the most cogent and convincing proof of the benefit of a new therapy. Further, the control group (whether a placebo or a proven active therapy) provides a better comparison of the safety of a new therapy. The benefit and risk must be traded off in the approval of new therapies.

NOTES

19.1 Interventions Other Than Drugs

In the discussion above we have discussed RCTs primarily as if they were for new drugs or biologics. Many interventions, such as medical devices, have been and/or could be investigated using RCTs or analogs. A variety of surgical interventions have been investigated by RCTs.

Prevention programs, such as smoking-cessation programs, can be investigated by randomizing larger experimental units. For example, in an NIH study of smoking prevention, the school district was the unit of randomization [Peterson et al., 2000]. One could randomize to different health care strategies, different modes of psychotherapy, and so on. In these studies the unit of randomization may be much larger; such group randomization is discussed by Feng et al. [2001].

19.2 Drug Approval and Physician Use of Drugs

In the United States, drugs are approved by the Food and Drug Administration. The approval includes labeling that specifies the population the drug is to benefit (i.e., the *indication*) as well as dosing information and warnings about safety, interactions with other drugs, and so on. Physicians may then legally use the drug for other indications (other diseases or patient populations) without violating the law (*off-label use*). If this use is in accord with the practice norms of the community, adequate malpractice defense can often be established. Drug companies selling the drugs are prohibited by law from advertising such off-label use of their product. One suspects that implied off-label uses are sometimes promoted.

19.3 Generic Drugs

In the United States, from the time that human experimentation begins, a sponsor has exclusive rights to sell the drug (assuming approval) for a limited period of time. The rights are for 17 years from the time the application is approved for experimentation on humans. Thus, there is a limited time to recoup research costs and make a profit. After this time, others may manufacture and sell the drug provided that they establish that it is the same drug (*bioequivalence*). These are called *generic drugs*. Equivalence is shown by establishing that the pharmacokinetics is the same for the new version and the original approved version. We do not address the topic of bioequivalence further here (see Chow and Liu [2000]).

19.4 Further Reading: Specific Topics

For more information on informed consent see, for example, Faden and Beauchamp [1986]. For a mathematical discussion of what constitutes an appropriate surrogate endpoint, see the paper of Prentice [1989]. For nice discussions of the history of blinding, see the papers by Kaptchuk [1998] and Chalmers [2001]. Some references on the benefits of the randomized clinical trial are Ederer [1975], Green [1982], Greenberg [1951], and Kempthorne [1977].

Since the 1970s, the number of articles and books about RCTs and statistical analysis has grown exponentially (e.g., books on clinical trials: Bulpitt, 1996; Cato and Sutton, 2002; Chow and Liu, 2003; Cleophas et al., 2002; Duley and Farrell, 2002; Friedman et al., 1999; Matthews, 2000; Meinert, 1986; Mulay, 2001; Norleans, 2001; Piantadosi, 1997; Pocock, 1996; Spilker, 1991).

There are numerous books about particular disease areas (e.g., AIDS [Finkelstein and Schoenfeld, 1999]; cardiology and cardiovascular disease [Hennekens and Zorab, 2000; Pitt et al., 1997]; epilepsy [French et al., 1997]; hypertension [Black, 2001]; multiple sclerosis [Goodkin and Rudick, 1998]; neurology [Guilogg, 2001; Porter and Schoenberg, 1990]; oncology [Green et al., 2002]; opthamology [Kertes and Conway, 1998]); and for material for patients [Giffels, 1996; Slevin and Wood, 1996]; aspects of trials, such as quality of life and pharmacoeconomics [Fairclough, 2002; Spilker, 1995]; data management [McFadden, 1997]; combining data from trials (metaanalysis: [Whitehead, 2002]; evaluating the literature [Ascione, 2001]; and dictionary or encyclopedic entries [Day, 1999; Redmond et al., 2001]).

REFERENCES

American Statistical Association [1999]. *Ethical Guidelines for Statistical Practice.* ASA, Alexandria, VI.

Ascione, F. J. [2001]. *Principles of Scientific Literature Evaluation: Critiquing Clinical Drug Trials*. American Pharmaceutical Association, Washington, DC.

Beauchamp, T. L., and Childress, J. F. [2001]. *Principles of Biomedical Ethics*, 5th ed. Oxford University Press, New York.

Bigger, J. T., Jr., [1990]. Editorial: the events surrounding the removal of encainide and flecainide from the Cardiac Arrhythmia Suppression Trial (CAST) and why CAST is continuing with moricizine. *Journal of the American College of Cardiology*, **15**: 243–245.

Black, H. R. [2001]. *Clinical Trials in the Pharmacologic Management of Hypertension*. Marcel Dekker, New York.

Box, J. F. [1978]. *R. A. Fisher: The Life of a Scientist*. Wiley, New York, p. 146.

Bulpitt, C. J. [1996]. *Randomized Controlled Clinical Trials*. Kluwer Academic, New York.

Cardiac Arrhythmia Pilot Study (CAPS) Investigators [1988]. Effect of encainide, flecainide, imipramine and moricizine on ventricular arrhythmias during the year after acute myocardial infarction: the CAPS. *American Journal of Cardiology*, **61**: 501–509.

Cardiac Arrhythmia Suppression Trial (CAST) Investigators [1989]. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *New England Journal of Medicine,* **321**: 406–412.

Cato, A. E., and Sutton, L. [2002]. *Clinical Drug Trials and Tribulations*, 2nd ed. Marcel Dekker, New York.

Chalmers, I. [2001]. Comparing like with like: some historical milestones in the evolution of methods to create unbiased groups in therapeutic experiments. *International Journal of Epidemiology*, **30**: 1156–1164.

Chow, S.-C., and Liu, J.-P. [2003]. *Design and Analysis of Clinical Trials: Concepts and Methodologies*, 2nd ed. Wiley, New York.

Chow, S.-C., and Liu, J.-P. [2000]. *Design and Analysis of Bioavailability and Bioequivalence Studies*, rev. ed. Marcel Dekker, New York.

Cleophas, T. J., Zwinderman, A. H., and Cleophas, T. F. [2002]. *Statistics Applied to Clinical Trials*. Kluwer Academic, New York.

Coronary Drug Project Research Group [1980]. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. *New England Journal of Medicine*, **303**: 1038–1041.

Cui, L., Hung, H. M. J., and Wang, S.-J. [1999]. Modification of sample size in group sequential clinical trials. *Biometrics*, **55**: 853–857.

Day, S. [1999]. *Dictionary for Clinical Trials*. Wiley, New York.

Duley, L., and Farrell, B. (eds.) [2002]. *Clinical Trials*. British Medical Association, London.

Echt, D. S., Liebson, P. R., Mitchell, B., Peters, R. W., Obias-Manno, D., Barker, A. H., Arensberg, D., Baker, A. Friedman, L., Greene, H. L., Huther, M. L., Richardson, D. W., and the CAST Investigators [1991]. Mortality and morbidity in patients receiving encainide, flecainide, or placebo: the Cardiac Arrhythmia Suppression Trial. *New England Journal of Medicine*, **324**: 781–788.

Ederer, F. [1975]. Why do we need controls? Why do we need to randomize? *American Journal of Ophthalmology*, **79**: 758–762.

Edgington, E. S. [1995]. *Randomization Tests*, 3rd rev. exp. ed. Marcel Dekker, New York.

Efron, B. [1971]. Forcing a sequential experiment to be balanced. *Biometrika*, **58**: 403–417.

Ellenberg, S. S., and Temple, R. [2000]. Placebo-controlled trials and active-control trials in the evaluation of new treatments: 2. Practical issues and specific cases. *Annals of Internal Medicine*, **133**: 464–470.

Ellenberg, S. S., Fleming, T. R., and DeMets, D. L. [2002]. *Data Monitoring Committees in Clinical Trials*. Wiley, New York.

Faden, R. R., and Beauchamp, T. L. [1986]. *A History and Theory of Informed Consent*. Oxford University Press, New York.

Fairclough, D. L. [2002]. *Design and Analysis of Quality of Life Studies in Clinical Trials*. CRC Press, Boca Raton, FL.

Federal Regulations [1988]. 21 CFR Ch. I, Part 56: Institutional Review Boards (4-1-88 ed.). U.S. Government Printing Office, Washington, DC.

Feng, Z., Diehr, P., Peterson, A., and McLerran, D. [2001]. Selected statistical issues in group randomized trials. *Annual Review of Public Health*, **22**: 167–187.

Finkelstein, D. M., and Schoenfeld, D. A. [1999]. *AIDS Clinical Trials*. Wiley, New York.

Fisher, L. D. [1998a]. Ethics of randomized clinical trials. In *Encyclopedia of Biostatistics*, Vol. 2, P. Armitage and T. Colton (eds.). Wiley, New York, pp. 1394–1398.

Fisher, L. D. [1998b]. Self-designing clinical trials. *Statistics in Medicine,* **17**: 1551–1562.

Fisher, L. D, Dixon, D. O., Herson, J., Frankowski, R. F., Hearron, M. S., and Peace, K. E. [1990]. Intention to treat in clinical trials. In *Statistical Issues in Drug Research and Development*, K. E. Peace (ed.). Marcel Dekker, New York, pp. 331–350.

Fleming, T. R., and DeMets, D. L. [1996]. Surrogate endpoints in clinical trials: are we being mislead? *Annals of Internal Medicine*, **125**: 605–613.

French, J. A., Leppik, I. E., and Dichter, M. A. [1997]. *Antiepileptic Drug Trials*, Vol. 76. Lippincott Williams & Wilkins, Philadelphia.

Friedman, L., Furberg, C., and DeMets, D. L. [1999]. *Fundamentals of Clinical Trials*, 3rd ed. Springer-Verlag, New York.

Furberg, C. D. [1983]. Effect of antiarrhythmic drugs on mortality after myocardial infarction. *American Journal of Cardiology*, **52**: 32C–36C.

Giffels, J. J. [1996]. *Clinical Trials: What You Should Know before Volunteering to Be a Research Subject*. Demos Medical Publishing, New York.

Goodkin, D. E., and Rudick, R. A. [1998]. *Multiple Sclerosis: Advances in Clinical Trial Design, Treatment and Perspectives*. Springer, New York.

Graboys, T. B., Lown, B., Podrid, P. J., and DeSilva, R. [1982]. Long-term survival of patients with malignant ventricular arrhythmia treated with antiarrhythmic drugs. *American Journal of Cardiology*, **50**: 437–443.

Green, S. B. [1982]. Patient heterogeneity and the need for randomized clinical trials. *Controlled Clinical Trials*, **3**: 189–198.

Green, S., Benedetti, J., and Crowley, J. [2002]. *Clinical Trials in Oncology*, 2nd ed. CRC Press, Boca Raton, FL.

Greenberg, B. G. [1951]. Why randomize? *Biometrics*, **7**: 309–322.

Guilogg, R. J. (ed.) [2001]. *Clinical Trials in Neurology*. Springer, New York.

Hennekens, C. H., and Zorab, R. [2000]. *Clinical Trials in Cardiovascular Disease: A Companion to Braunwald's Heart Disease*. W. B. Saunders, Philadelphia.

IMPACT Research Group [1984]. International Mexiletine and placebo antiarrhythmic coronary trial: I. Report on arrhythmias and other findings. *Journal of the American College of Cardiology*, **4**: 1148–1163.

Jennison, C., and Turnbull, B. W. [2000]. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall, New York.

Kaptchuk, T. J. [1998]. Intentional ignorance: a history of blind assessment and placebo controls in medicine. *Bulletin of the History of Medicine*, **72**: 389–433.

Karlowski, T. R., Chalmers, T. C., Frenkel, L. D., Zapikian, A. Z., Lewis, T. L., and Lynch, J. M. [1975]. Ascorbic acid for the common cold: a prophylactic and therapeutic trial. *Journal of the American Medical Association*, **231**: 1038–1042.

Kempthorne, O. [1977]. Why randomize? *Journal of Statistical Planning and Inference*, **1**: 1–25.

Kertes, P. J., and Conway, M. D. [1998]. *Clinical Trials in Opthamology: A Summary and Practice Guide*. Lippincott Williams & Wilkins, Philadelphia.

Kesteloot, H., and Joosens, J. V. [1980]. *Epidemiology of Arterial Blood Pressure: Developments in Cardiovascular Medicine*, Vol. 8. Martinus Nijhoff, Dordrecht, The Netherlands.

Lan, K. K. G., and DeMets, D. L. [1983]. Discrete sequential boundaries for clinical trials. *Biometrika*, **70**: 659–663.

Lifton, R. J. [1986]. *The Nazi Doctors: Medical Killing and the Psychology of Genocide*. Basic Books, New York.

Little, R. J. A., and Rubin, D. B. [2002]. *Statistical Analysis of Missing Data*, 2nd ed. Wiley, New York.

Mathieu, M. [2002]. *New Drug Development: Regulation Overview*. Parexel International Corporation, Cambridge, MA.

Matthews, J. N. [2000]. *Introduction to Randomized Controlled Clinical Trials*. Edward Arnold, London.

McFadden, E. [1997]. *Management of Data in Clinical Trials*. Wiley, New York.

Meinert, C. L. [1986]. *Clinical Trials: Design, Conduct, and Analysis*. Oxford University Press, New York.

Mulay, M. [2001]. *A Step-by-Step Guide to Clinical Trials*. Jones & Bartlett, Boston.

Norleans, M. X. [2001]. *Statistical Methods for Clinical Trials*. Marcel Dekker, New York.

O'Brien, P. C., and Fleming, T. R. [1979]. A multiple testing procedure for clinical trials. *Biometrics*, **35**: 549–556.

Peterson, A. V., Mann, S. L., Kealey, K. A., and Marek, P. M. [2000]. Experimental design and methods for school-based randomized trials: experience from the Hutchinson smoking prevention project (HSPP). *Controlled Clinical Trials*, **21**: 144–165.

Piantadosi, S. [1997]. *Clinical Trials: A Methodologic Perspective*. Wiley, New York.

Pitt, B., Julian, D., and Pocock, S. J. [1997]. *Clinical Trials in Cardiology*. W. B. Saunders, Philadelphia.

Pocock, S. J. [1982]. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics*, **38**: 153–162.

Pocock, S. J. [1996]. *Clinical Trials: A Practical Approach*. Wiley, New York.

Pope, A. [1733]. *An Essay on Man*. Cited in [1968] *Bartlett's Familiar Quotations*, 14th ed. Little, Brown, Boston.

Porter, R. J., and Schoenberg, B. S. [1990]. *Controlled Clinical Trials in Neurological Disease*. Kluwer Academic, New York.

Pratt, C. M. (ed.) [1990]. A symposium: the Cardiac Arrhythmia Suppression Trial—does it alter our concepts of and approaches to ventricular arrhythmias? *American Journal of Cardiology*, **65**: 1B–42B.

Pratt, C. M., Brater, D. C., Harrell, F. E., Jr., Kowey, P. R., Leier, C. V., Lowenthal, D. T., Messerlie, F., Packer, M., Pritchett, E. L. C., and Ruskin, J. N. [1990]. Clinical and regulatory implications of the Cardiac Arrhythmia Suppression Trial. *American Journal of Cardiology*, **65**: 103–105.

Prentice, R. L. [1989]. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, **8**: 431–440.

Redmond, C. K., Colton, T., and Stephenson, J. [2001]. *Biostatistics in Clinical Trials*. Wiley, New York.

Reiser, S. J., Dyck, A. J., and Curran, W. J. (eds.). [1947]. The Nuremberg Code. in *Ethics in Medicine: Historical Perspectives and Contemporary Concerns*. MIT Press, Cambridge, MA, pp. 272–274.

Royal Statistical Society [1993]. *Code of Conduct*. RSS, London.

Ruskin, J. N. [1989]. The cardiac arrhythmia suppression trial (CAST) (editorial). *New England Journal of Medicine*, **321**: 386–388.

Schouten, H. J. A. [2000]. Combined evidence from multiple outcomes in a clinical trial. *Journal of Clinical Epidemiology*, **53**: 1137–1144.

Slevin, M., and Wood, S. [1996]. *Understanding Clinical Trials*. Cancer BACUP, London. *http://www.cancerbacup.org.uk/info/trials.htm.* Accessed June 10, 2003.

Spilker, B. [1991]. *Guide to Clinical Trials*. Lippincott Williams & Wilkins, Philadelphia.

Spilker, B. [1995]. *Quality of Life and Pharmacoeconomics in Clinical Trials*. Lippincott Williams & Wilkins, Philadelphia.

Student [1931]. The Lanarkshire milk experiment. *Biometrika*, **23**: 398–406.

Temple, R. J. [1995]. A regulatory authority's opinion about surrogate endpoints. In *Clinical Measurement in Drug Evaluation,* W. S. Nimmo and G. T. Tucker, (eds.). Wiley, New York, pp. 3–22.

Temple, R., and Ellenberg, S. S. [2000]. Placebo-controlled trials and active-control trials in the evaluation of new treatments: 1. Ethical and scientific issues. *Annals of Internal Medicine*, **133**: 455–463.

Thomas, L. [1983]. *The Youngest Science*. pp. 30–31. New York, NY, The Viking Press.

Wall Street Journal [2001]. Cost of drug development found to rise, p. B14, Dec. 3, 2001, taken from the Tufts Center for Drug Development. Also given in Tufts Center for the Study of Drug Development, *Outlook 2002*, Boston.

Whitehead J. [1983]. *The Design and Analysis of Sequential Clinical Trials*. Ellis Horwood, Chichester, West Sussex, England.

Whitehead, A. [2002]. *Meta-analysis of Controlled Clinical Trials*. Wiley, New York.

World Medical Association [1975]. Declaration of Helsinki, revision of original 1964 version. In *Ethics in Medicine: Historical Perspectives and Contemporary Concerns*, S. J., Reiser, A. J. Dyck, and W. J., Curran, (eds.). MIT Press, Cambridge, MA, pp. 328–330.

Wright, S. P. [1992]. Adjusted p-values for simultaneous inference. *Biometrics*, **48**: 1005–1013.

CHAPTER 20

Personal Postscript

20.1 INTRODUCTION

One reviewer of this book felt that it would be desirable to have a final chapter that ended the book with more interesting material than yet another statistical method. This stimulated us to think about all the exciting, satisfying, and interesting things that had occurred in our own careers as biostatisticians. We decided to try to convey some of these feelings through our own experiences. This chapter is unabashedly written from a first-person point of view. The examples do not represent a random sample of our experiences but rather, the most important and/or interesting experiences of our careers. There is some deliberate duplication of background material that appears in other chapters so that this chapter may be self-contained (except for the statistical methods used). We have not made an effort to choose experiences that illustrate the use of many different statistical methods (although this would have been possible). Rather, we want to entertain, and in doing so, show the important collaborative role of biostatistics in biomedical research.

20.2 IS THERE TOO MUCH CORONARY ARTERY SURGERY?

The National Institutes of Health in the United States funds much of the health research in the country. During the late 1960s and early 1970s, an exciting new technique for dealing with anginal chest pain caused by coronary artery disease was developed. Recall that coronary artery disease is caused by fibrous fatty deposits building up within the arteries that supply blood to the heart muscle (i.e., the coronary arteries). As the arteries narrow, the blood supply to the heart is inadequate when there are increased demands because of exercise and/or stress; the resulting pain is called *angina*. Further, the narrowed arteries tend to close with blood clots, which results in the death (infarction) of heart muscle (myocardium), whose oxygen and nutrients are supplied by the blood coming through the artery; these heart attacks are also called *myocardial infarctions* (MIs). *Coronary artery bypass graft* (CABG; pronounced “cabbage”) surgery replumbs the system. Either saphenous veins from the leg or the internal mammary arteries already in the chest are used to supply blood beyond the narrowing, that is, bypassing the narrowing. Figure 20.1 shows the results of bypass surgery. A key measure of damaged arteries is the *ejection fraction* (EF), the proportion of blood pushed out of the pumping chamber of the heart, the left ventricle. A normal value is 0.5 or greater. EF values between 0.35 and 0.49 are considered evidence of mild to moderate impairment. When the heart muscle is damaged, say by an MI, or has a limited blood supply, the EF decreases.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

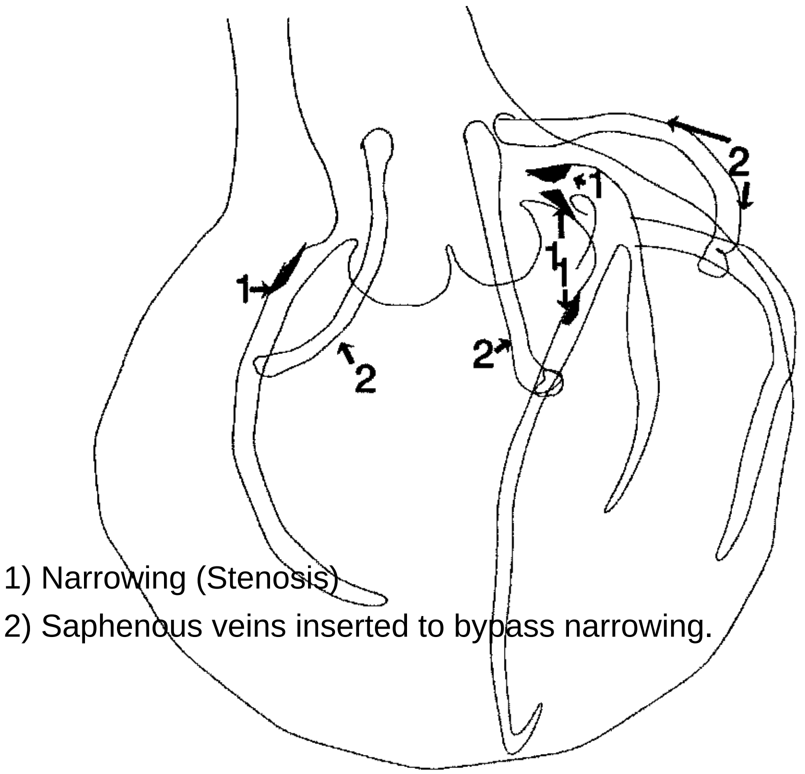

Figure 20.1 Schematic display of coronary artery bypass graft surgery. Here saphenous veins from the leg are sewn into the aorta where the blood is pumped out of the heart and then sewn into coronary arteries beyond narrowings in order to deliver a normal blood supply.

Because the restored blood flow should allow normal function, it was conjectured that surgery would both remove the anginal pain and also prolong life by reducing both the stress on the heart and the number of myocardial infarctions. It became clear early on that surgery did help to relieve angina pain (although even this has been debated; see Preston [1977]). However, the issue of prolonging life was more debatable. The amount of surgery had important implications for the health care budget, since in the early 1970s the cost per operation ranged between $12,000 and $50,000, depending on the location of the clinic, complexity of the surgery, and a variety of other factors. The number of surgeries by year up to 1972 is shown in Figure 20.2.

Because of the potential savings in lives and the large health resources requirements, the National Heart, Lung and Blood Institute (NHLBI; at that time the National Heart Institute) decided that it was appropriate to obtain firm information about which patients have improved survival with CABG surgery. Such therapeutic comparisons are best addressed through a randomized clinical trial, and that was the approach taken here with randomization to early surgery or early medical treatment. However, because not all patients could ethically be randomized, it was also decided to have a registry of patients studied with coronary angiography so that observational data analyses could be performed on other subsets of patients to compare medical and surgical therapy. When the NHLBI has internally sponsored initiatives, they are developed through a request for proposals (RFP), which recruits investigators to perform the collaborative research. This trial and registry, called the Coronary Artery Surgery Study (CASS), had two RFPs; one was for clinical sites and the other for a coordinating center. The RFP for the

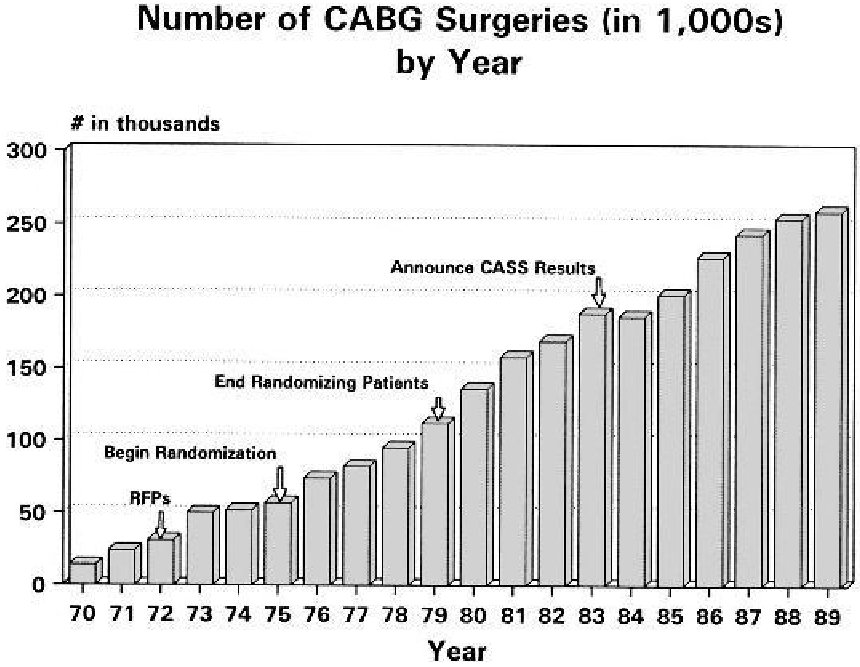

Figure 20.2 Number of coronary artery bypass graft surgeries in thousands of operations by year, 1970–1989. Marked are some of the key time points in the Coronary Artery Surgery Study. (Data courtesy of the cardiac diseases branch of the National Heart, Lung and Blood Institute from the National Hospital Discharge Survey, National Center for Health Services.)

clinical sites was issued in November 1972 and described the proposed study, both randomized and registry components, and asked for clinics to help complete the design and to enroll patients in the randomized and registry components of the study. The coordinating center RFP requested applications for a center to help with the statistical design and analysis of the study, to receive and process the study forms with a resultant database, to produce reports for monitoring the progress of the study and to otherwise participate in the quality assurance of the study, and finally, to collaborate in the analysis and publication of the randomized study and registry results. The organization of such a large multicenter study had a number of components: The NHLBI had a program office with medical, biostatistical, and financial expertise to oversee operation of the study; there were 15 cooperating clinical sites in the United States and Canada; the Coordinating Center was at the University of Washington under the joint direction of Lloyd Fisher and Richard Kronmal; a laboratory to read electrocardiograms (ECG lab) was established at the University of Alabama.

The randomized study enrolled 780 cases with mild angina or no angina with a prior MI, and significant disease (defined as a 70% or greater narrowing of the internal diameter of a coronary artery that was suitable for bypass surgery). There were a variety of other criteria for eligibility for randomization. The registry, including the patients randomized, enrolled 24,959 patients. Extensive data were collected on all patients. The first patients were enrolled in July 1974, with randomization beginning in August 1975 [CASS Principal Investigators and Their Associates, 1981]. Follow-up of patients within the randomized study ended in 1992. Needless to say, such a large effort cost a considerable amount of money, over $30,000,000. It will be shown that the investment was very cost-effective.

Results of the survival analysis and indicators of the quality of life were made public in 1983 [CASS Investigators, 1983a,b, 1984b]. The survival estimates for the subjects randomized to

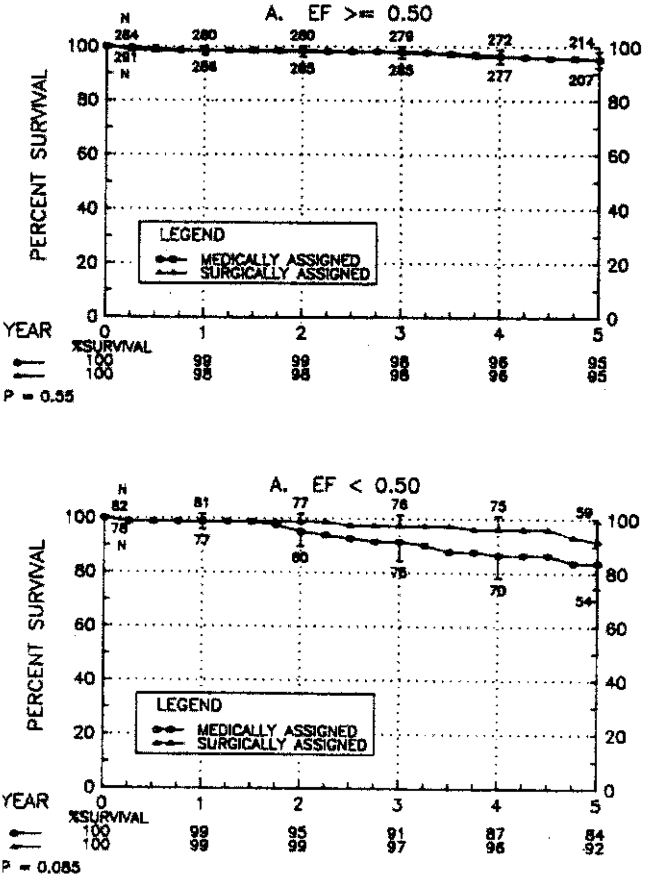

Figure 20.3 Data from the CASS randomized clinical trial; the bottom panel is for patients with ejection fractions less than 0.50; the top panel is for patients with ejection fractions of 0.50 or above. The p-values are the log-rank statistic for the comparison.

initial medical and surgical treatment are given in Figure 20.3. For patients with an EF of 0.5 or more, the survival curves were virtually identical; for subjects with lower EF values, there was a trend toward favorable mortality in the surgery group ($p = 0.085$ by the log-rank test).

A number of points were important in interpreting these data:

1. The CASS investigators agreed before the study started that the surgery was efficacious in relieving angina. Thus, if a patient started to have severe angina that could not be controlled by medication, the patient was allowed to "cross over" to surgery. By year 5, 24% of the patients assigned to initial medical therapy had crossed over to the CABG surgery group. If surgery is, in fact, having a beneficial effect and there is much crossover, the statistical power of the comparison is reduced. Is this a bad thing? The issue is a complex one (see Peto et al. [1977]; Weinstein and Levin [1989]; Fisher et al. [1989, 1990]). We know that one of the benefits of randomization is that we are assured of comparable groups (on average) even with respect to unrecorded and unknown variables. If we manipulated people, or parts of their experience, between groups by using events that occurred after the time of randomization, bias can enter the analysis. Thus, people should be included only in the group to which they are randomized; this is called an *intent-to-treat analysis* since they are counted with the group whose treatment was intended. (Does such an approach avoid bias? Does it always make biological sense?) The CASS investigators favored an intent-to-treat analysis not only because it avoided possible bias but also because of the ethical imperative to perform CABG surgery for pain relief when the pain became intractable under medical treatment. Thus, including all the experience of those assigned to initial medical treatment, including CABG surgery and subsequent events, mirrored what would happen to such a group in real life. This is the question that the trial should answer: Is early surgery helpful when patients will receive it anyway when the pain becomes too severe? However, the power of such a comparison will be diminished by the crossovers. The interpretation of such intent-to-treat analyses must acknowledge that without the crossover, the results could have been substantially different.

2. Because bypass surgery is such a big industry (e.g., 200,000 surgeries per year at $30,000 per operation adds up to $6 billion per year), with many careers and much professional prestige committed to the field, one could expect a counter reaction if surgery did not look beneficial. Such reactions did occur, and a number of editorials, reviews, and sessions at professional meetings were given to consideration of the results. One of the authors (LF) appeared on the CBS national news as well as going to New York City to be interviewed by Mike Wallace and appearing on the TV program *60-minutes*. Based largely on the CASS results, the program suggested that there was too much CABG surgery.

3. It is important to keep the findings in context. They did not apply to all, or even most, patients. The CASS was one of three major randomized trials of CABG surgery. One study showed definitively that the surgery prolonged life in patients with left main disease [Takaro et al., 1976]. This study excluded patients with severe angina and thus had nothing to say about differential survival in such patients. In fact, there is observational data to suggest that early elective CABG surgery prolonged life in such patients [Kaiser et al., 1985; Myers et al., 1989].

4. Even though the findings may apply to a *relatively* small number of patients, the results could have a very substantial impact on the national health scene. Subsequent CASS papers showed that the trend toward increased survival with surgery in the low ejection fraction patients was real [Passamani et al., 1985; Alderman et al., 1990]. Thus, suppose that we restrict ourselves to those patients with EFs of at least 0.5. This accounted for 575 of the 780 randomized patients. Suppose that the randomized study had not been in effect; how many of these patients might have received early surgery? In the CASS study, there were 1315 patients who met the eligibility criteria and might have been randomized but in fact were not randomized [CASS Principal Investigators, 1984a; Chaitman et al., 1990]; these patients were called the *randomizable patients*. In this group, 43% (570/1315) received early elective surgery. Of those who did not receive early surgery and had good ejection fractions, by 10 years, 38% had received surgery. That is, 60% or so did not receive surgery. Assuming that the CASS clinics were representative of the surgical practice in the country (they may have been more conservative than many centers because they were willing to participate in research to assess the appropriate role of bypass surgery), about 4.4% of the surgery in the United States might be prevented by applying the results of the study. In a year with 188,000 CABGs costing $30,000 each, this would lead to a savings of over $245 million. Over a 4-year period over $1 billion could be saved in surgical costs. However, because the patients treated medically have more anginal pain, they have higher drug costs; they might have higher hospitalization costs (but they do not; see CASS Principal Investigators [1983b] and Rogers et al. [1990]). Without going into detail, it is my (L.F.) opinion that the study saved several billion dollars in health care costs without added risk to patient lives.

5. The issues are more complex than presented here; we have not discussed the findings and integration of results with the other major randomized studies of CABG surgery. Further, it is important to note that a number of other proven and/or promising techniques for dealing with coronary artery disease (CAD) have been developed. These include drug and/or dietary therapy; blowing up balloons in the artery to "squish" the narrowing into the walls of the artery [**p**ercutaneous **t**ransluminal **c**oronary **a**ngioplasty (PTCA)]; introducing lasers into the coronary arteries to disintegrate the plaques that narrow the arteries; using a roto-rooter in the arteries to replumb by grinding up the plaques; and stents. Although all of these alternatives have been or are being used, the number of CABG surgeries did not decrease but leveled off up to 1989.

6. The surgery may improve with time as techniques and skills improve. Further, it became apparent that the results of the surgery deteriorated at 10 to 12 years or so. The disease process at work in the coronary arteries also was at work in the grafts that bypassed the narrowed areas; thus the grafts themselves narrow and close, often requiring repeat CABG surgery. Internal mammary grafts have a longer lifetime and are now used more often, suggesting that current long-term results will be better.

In summary, the CASS study showed that in patients with selected characteristics, CABG surgery is not needed immediately to prolong life and can often be avoided. The study was a bargain both in human and economic terms, illustrating the need and benefits of careful evaluation of important health care procedures.

20.3 SCIENCE, REGULATION, AND THE STOCK MARKET

In the United States, foods, drugs, biologics, devices, and cosmetics are regulated by the Food and Drug Administration (FDA). To get a new drug or biologic approved for marketing within the United States, the sponsor (usually, a pharmaceutical company or biotechnology company) must perform adequate and well-controlled clinical trials that show the efficacy and safety of the product. The FDA is staffed with personnel who have expertise in a number of areas, including pharmacology, medicine, and biostatistics. The FDA staff reviews materials submitted and rules on the approval or nonapproval of a product. The FDA also regulates marketing of the compounds. Marketing before approval is not allowed. The FDA uses the services of a number of advisory committees composed of experts in the areas considered. The deliberations of the advisory committees are carried out in public, often with large audiences in attendance. At the meetings, the sponsor makes a presentation, usually with both company and clinical experts, and answers questions from the committee. The FDA has a presence, asks questions, particularly of the advisory committee, but usually does not play a dominant role. At the end of its deliberations the committee votes on whether the drug or biologic should be approved, should be disapproved, or should be disapproved at least temporarily because further information is needed before final approval or disapproval is appropriate.

Two of the authors have been members of FDA advisory committees, G.vB. with the peripheral and central nervous system drugs advisory committee and L.F. with the cardiovascular and renal drugs advisory committee. Here we discuss the consideration of one biologic: tissue plasminogen activator (tPA). A *biologic* is a compound that occurs naturally in the human body, whereas a *drug* is a compound that does not occur naturally but is introduced artificially, solely for therapeutic purposes. For example, insulin is a biologic, whereas aspirin is a drug. Here we will use the term *drug* for tPA because that is the more common usage, although within the FDA, drugs and biologics go to different divisions. We turn next to the background and rationale for the use of tPA.

As discussed above, when coronary artery disease occurs, it narrows the arteries, changing the fluid flow properties of the blood, leading to clotting within the coronary arteries. These clots then block the blood supply to the heart muscle, resulting in heart attacks, or myocardial infarctions (MIs), as discussed above. The clot is composed largely of fibrin. When converted to plasmin, plasminogen converts insoluble fibrin into soluble fragments. One conceptual way to treat a heart attack would be to dissolve the blood clot, thus reestablishing blood flow to the heart muscle and preventing the death of the muscle, saving the heart and often saving the life. Should a drug be approved for dissolving blood clots alone? Although biologically plausible, does this assure that the drug will work? In other words, is this an acceptable surrogate endpoint?

Returning to the thrombolytic (i.e., to *lyse*, or break up, the blood clot, the thrombosis) tPA therapy, it is clear that lysing the coronary arterial blood clot is a surrogate endpoint. Should this surrogate endpoint be appropriate for approving the drug? After all, there is such a clearcut biological rationale: Coronary artery clots cause heart attacks; heart attacks damage the heart, often either impairing the heart function, and thus lowering exercise capacity, or killing the person directly. But experience has shown that very convincing biological scenarios do not always deliver the benefits expected; below we present an important example of a situation where an obvious surrogate endpoint did not work out.

Let us return now to the tPA cardiorenal advisory committee meeting and decision. In addition to tPA, another older thrombolytic drug, streptokinase, was also being presented for approval for the same indication. Prior to the meeting, there was considerable publicity over the upcoming meeting and possible approval of the drug tPA. The advisory committee meeting was to take place on Friday, May 29, 1987. On Thursday, May 28, 1987, the *Wall Street Journal* published an editorial entitled “The TPA Decision.” The editorial read as follows:

> Profile of a heart-attack victim: 49 years old, three children, middle-manager, in seemingly good health. Cutting the grass on a Saturday afternoon, he is suddenly driven to the ground with severe chest pain. An ambulance takes him to the nearest emergency room, where he receives drugs to reduce shock and pain.
>
> At this point, he is one of approximately 4000 people who suffer a heart attack each day. If he has indeed had a heart attack, he will experience one of two possible outcomes. Either he will be dead, joining the 500,000 Americans killed each year by heart attack. Or, if he’s lucky, he will join the one million others who go on to receive some form of therapy for his heart disease.
>
> Chances of survival will depend in great part on the condition of the victim’s heart, that is, how much permanent muscular damage the heart sustained during the time a clot prevented the normal flow of blood into the organ. Heart researchers have long understood that if these clots can be broken up early after a seizure’s onset, the victim’s chances of staying alive increase significantly. Dissolving the clot early enhances the potential benefits of such post-attack therapies as coronary bypass surgery or balloon angioplasty.
>
> Tomorrow morning, a panel of the Food and Drug Administration will review the data on a blood-clot dissolver called TPA, for tissue-type plasminogen activator. In our mind, TPA—not any of the pharmaceutical treatments for AIDS—is the most noteworthy, unavailable drug therapy in the United States. Put another way, the FDA’s new rules permitting the distribution of experimental drugs for life-threatening diseases came under pressure to do something about the AIDS epidemic. But isn’t it as important for the government to move with equal speed on the epidemic of heart attacks already upon us?
>
> This isn’t to say that TPA is more important than AIDS treatments. Both have a common goal: keeping people alive. The difference is that while the first AIDS drug received final approval in about six months, TPA remains unapproved and unavailable to heart-attack victims despite the fact that the medical community has known for more than two years that it can save lives.
>
> How many lives? Obviously no precise projection is possible, but the death toll is staggering, with about 41,000 individuals killed monthly by heart attacks.
>
> In its April 4, 1985, issue, the *New England Journal of Medicine* carried the first report on the results of the National Institutes of Health’s TIMI study comparing TPA’s clot dissolving abilities with a drug already approved by the FDA. NIH prematurely ended that trial because TPA’s results were so significantly better than the other drug.
>
> In an accompanying editorial, the *Journal*’s editor, Dr. Arnold Relman, said a safe and effective thrombolytic “might be of immense clinical value.” In October 1985, a medical-policy committee of California’s Blue Shield recommended that TPA be recognized “as acceptable medical practice.” The following month at the American Heart Association’s meeting, Dr. Eugene Braunwald, chairman of the department of medicine at Harvard Medical School, said, “If R-TPA were available on a wide basis, I would select that drug today.” In its original TIMI report, the NIH said TPA would next be

> tested against a placebo; later, citing ethical reasons, the researchers dropped the placebo and now all heart patients in the TIMI trial are receiving TPA.
>
> It is for these reasons that we call TPA the most noteworthy unavailable drug in the U.S. The FDA may believe it is already moving faster than usual with the manufacturer's new-drug application. Nonetheless, bureaucratic progress [*sic*] must be measured against the real-world costs of keeping this substance out of the nation's emergency rooms. The personal, social and economic consequences of heart disease in this country are immense. The American Heart Association estimates the total costs of providing medical services for all cardiovascular disease at $71 billion annually.
>
> By now more than 4,000 patients have been treated with TPA in clinical trials. With well over a thousand Americans going to their deaths each day from heart attack, it is hard to see what additional data can justify the government's further delay in making a decision about this drug. If tomorrow's meeting of the FDA's cardio-renal advisory committee only results in more temporizing, some in Congress or at the White House should get on the phone and demand that the American public be given a reason for this delay.

The publicity before the meeting of the advisory committee was quite unusual since companies are prohibited from preapproval advertising; thus the impetus presumably came from other sources.

The cardiorenal advisory committee members met and considered the two thrombolytic drugs, streptokinase and tPA. They voted to recommend approval of streptokinase but felt that further data were needed before tPA could be approved. The reactions to the decision were extreme, but probably predictable given the positions expressed prior to the meeting.

The *Wall Street Journal* responded with an editorial on Tuesday, June 2, 1987, entitled "Human Sacrifice." It follows in its entirety:

> Last Friday an advisory panel of the Food and Drug Administration decided to sacrifice thousands of American lives on an altar of pedantry.
>
> Under the klieg lights of a packed hearing room at the FDA, an advisory panel picked by the agency's Center for Drugs and Biologics declined to recommend approval of TPA, a drug that dissolves blood clots after heart attacks. In a 1985 multicenter study conducted by the U.S. National Heart, Lung and Blood Institute, TPA was so conclusively effective at this that the trial was stopped. The decision to withhold it from patients should be properly viewed as throwing U.S. medical research into a major crisis.
>
> Heart disease dwarfs all other causes of death in the industrialized world, with some 500,000 Americans killed annually; by comparison, some 20,000 have died of AIDS. More than a thousand lives are being destroyed by heart attacks every day. In turning down treatment with TPA, the committee didn't dispute that TPA breaks up the blood clots impeding blood flow to the heart. But the committee asked that Genentech, which makes the genetically engineered drug, collect some more mortality data. Its submission didn't include enough statistics to prove to the panel that dissolving blood clots actually helps people with heart attacks.
>
> Yet on Friday, the panel also approved a new procedure for streptokinase, the less effective clot dissolver—or thrombolytic agent—currently in use. Streptokinase previously had been approved for use in an expensive, specialized procedure called intracoronary infusion. An Italian study, involving 11,712 randomized heart patients at 176 coronary-care units in 1984–1985, concluded that administering streptokinase intravenously reduced deaths by 18%. So the advisory panel decided to approve intravenous streptokinase, but not approve the superior thrombolytic TPA. This is absurd.
>
> Indeed, the panel's suggestion that it is necessary to establish the efficacy of thrombolysis stunned specialists in heart disease. Asked about the committee's justification for its decision, Dr. Eugene Braunwald, chairman of Harvard Medical School's department of medicine, told us: "The real question is, do you accept the proposition that the proximate cause of a heart attack is a blood clot in the coronary artery? The evidence is overwhelming, *overwhelming*. It is sound, basic medical knowledge. It is in every textbook of medicine. It has been firmly established in the past decade beyond any reasonable question. If you accept the fact that a drug [TPA] is twice as effective as

> streptokinase in opening closed vessels, and has a good safety profile, then I find it baffling how that drug was not recommended for approval."
>
> Patients will die who would otherwise live longer. Medical research has allowed statistics to become the supreme judge of its inventions. The FDA, in particular its bureau of drugs under Robert Temple, has driven that system to its absurd extreme. The system now serves itself first and people later. Data supersede the dying.
>
> The advisory panel's suggestion that TPA's sponsor conduct further mortality studies poses grave ethical questions. On the basis of what medicine already knows about TPA, what U.S. doctor will give a randomized placebo or even streptokinase? We'll put it bluntly: Are American doctors going to let people die to satisfy the bureau of drugs' chi-square studies?
>
> Friday's TPA decision should finally alert policy makers in Washington and the medical-research community that the theories and practices now controlling drug approval in this country are significantly flawed and need to be rethought. Something has gone grievously wrong in the FDA bureaucracy. As an interim measure FDA Commissioner Frank Young, with Genentech's assent, could approve TPA under the agency's new experimental drug rules. Better still, Dr. Young should take the matter in hand, repudiate the panel's finding and force an immediate reconsideration. Moreover, it is about time Dr. Young received the clear, public support of Health and Human Services Secretary Dr. Otis Bowen in his efforts to fix the FDA.
>
> If on the other hand Drs. Young and Bowen insist that the actions of bureaucrats are beyond challenge, then perhaps each of them should volunteer to personally administer the first randomized mortality trials of heart-attack victims receiving the TPA clot buster or nothing. Alternatively, coronary-care units receiving heart-attack victims might use a telephone hotline to ask Dr. Temple to randomize the trial himself by flipping a coin for each patient. The gods of pedantry are demanding more sacrifice.

Soon after joining the Cardiovascular and Renal Drugs Advisory Committee, L.F. noticed that a number of people left the room at what seemed inappropriate times, near the end of some advisory deliberations. I was informed that often, stock analysts with expertise in the pharmaceutical industry attended meetings about key drugs; when the analysts thought they knew how the vote was going to turn out, they went out to the phones to send instructions. That was the case during the tPA deliberations (and made it particularly appropriate that the *Wall Street Journal* take an interest in the result). Again we convey the effect of the deliberations through quotations taken from the press. On June 1, 1978, the *Wall Street Journal* had an article under the heading "FDA Panel Rejection of Anti-Clot Drug Set Genentech Back Months, Perils Stock." The article said in part:

> A Food and Drug Administration advisory panel rejected licensing the medication TPA, spoiling the summer debut of what was touted as biotechnology's first billion-dollar drug. ... Genentech's stock—which reached a high in March of $64.50 following a 2-for-1 split—closed Friday at $48.25, off $2.75, in national over-the-counter trading, even before the close of the FDA panel hearing attended by more than 400 watchful analysts, scientists and competitors. Some analysts expect the shares to drop today. ... Wall Street bulls will also be rethinking their forecasts. For example, Kidder Peabody & Co.'s Peter Drake, confident of TPA's approval, last week predicted sales of $51 million in the second half of 1987, rising steeply to $205 million in 1988, $490 million in 1989 and $850 million in 1990.

USA Today, on Tuesday, June 2, 1987, on the first page of the Money section, had an article headed "Biotechs Hit a Roadblock, Investors Sell." The article began:

> Biotechnology stocks, buoyed more by promise than products, took one of their worst beatings Monday. Leading the bad-news pack: Biotech giant Genentech Inc., dealt a blow when its first blockbuster drug failed to get federal approval Friday. Its stock plummeted $11\frac{1}{2}$ points to $\$36\frac{3}{4}$, on 14.2 million shares traded—a one-day record for Genentech. "This is very serious, dramatically serious," said analyst Peter Drake, of Kidder, Peabody & Co., who Monday changed his recommendations for the

> group from buy to "unattractive." His reasoning: The stocks are driven by "a blend of psychology and product possibilities. And right now, the psychology is terrible."

Biotechnology stocks as a group dropped with the Genentech panel vote. This seemed strange to me because the panel had not indicated that the drug, tPA, was bad but only that in a number of areas the data needed to be gathered and analyzed more appropriately (as described below). The panel was certainly not down on thrombolysis (as the streptokinase approval showed); it felt that the risk/benefit ratio of tPA needed to be clarified before approval could be made.

The advisory committee members replied to the *Wall Street Journal* editorials both individually and in groups, explaining the reasons for the decision [Borer, 1987; Kowey et al., 1988; Fisher et al., 1987]. This last response to the *Wall Street Journal* was submitted with the title "The Prolongation of Human Life"; however, after the review of the article by the editor, the title was changed by the *Wall Street Journal* to "The FDA Cardio-Renal Committee Replies." The reply:

> The evaluation and licensing of new drugs is a topic of legitimate concern to not only the medical profession but our entire populace. Thus it is appropriate when the media, such as the *Wall Street Journal*, take an interest in these matters. The Food and Drug Administration recognizes the public interest by holding open meetings of advisory committees that review material presented by pharmaceutical companies, listen to expert opinions, listen to public comment from the floor and then give advice to the FDA. The Cardiovascular and Renal Drugs Advisory Committee met on May 29 to consider two drugs to dissolve blood clots causing heart attacks. The *Journal* published editorials prior to the meeting ("The TPA Decision," May 28) and after the meeting ("Human Sacrifice," June 2 and "The Flat Earth Committee," July 13). The second editorial began with the sentence: "Last Friday an advisory committee of the Food and Drug Administration decided to sacrifice thousands of American lives on an altar of pedantry." How can such decisions occur in our time? This reply by members of the advisory panel presents another side to the story. In part the reply is technical, although we have tried to simplify it. We first discuss drug evaluation in general and then turn to the specific issues involved in the evaluation of the thrombolytic drugs streptokinase and TPA.
>
> The history of medicine has numerous instances of well-meaning physicians giving drugs and treatments that were harmful rather than beneficial. For example, the drug thalidomide was widely marketed in many countries—and in West Germany without a prescription—in the late 1950s and early 1960s. The drug was considered a safe and effective sleeping pill and tranquilizer. Marketing was delayed in the U.S. despite considerable pressure from the manufacturer upon the FDA. The drug was subsequently shown to cause birth defects and thousands of babies world-wide were born with grotesque malformations, including seal-like appendages and lack of limbs. The FDA physician who did not approve the drug in the U.S. received an award from President Kennedy. One can hardly argue with the benefit of careful evaluation in this case. We present this, not as a parallel to TPA, but to point out that there are two sides to the approval coin—early approval of a good drug, with minimal supporting data, looks wise in retrospect; early approval, with minimal supporting data, of a poor drug appears extremely unwise in retrospect. Without adequate and well-controlled data one cannot distinguish between the two cases. Even with the best available data, drugs are sometimes found to have adverse effects that were not anticipated. Acceptance of unusually modest amounts of data, based on assumptions and expectations rather than actual observation is very risky. As will be explained below, the committee concluded there were major gaps in the data available to evaluate TPA.
>
> The second editorial states that "Medical research has allowed statistics to become the supreme judge of its inventions." If this means that data are required, we agree; people evaluate new therapies with the hope that they are effective—again, before licensing, proof of effectiveness and efficacy is needed. If the editorial meant that the TPA decision turned on some arcane mathematical issue, it is incorrect. Review of the transcript shows that statistical issues played no substantial role.
>
> We now turn to the drug of discussion, TPA. Heart attacks are usually caused by a "blood clot in an artery supplying the heart muscle with blood." The editorial quotes Dr. Eugene Braunwald, "The real question is, do you accept the proposition that the proximate cause of a heart attack is a blood clot in the coronary artery?" We accept the statement, but there is still a significant question: "What can one then do to benefit the victim?" It is not obvious that modifying the cause after the event

occurs is in the patient's best interest, especially when the intervention has toxicity of its own. Blood clots cause pulmonary embolism; it is the unusual patient who requires dissolution of the clot by streptokinase. Several trials show the benefit does not outweigh the risk.

On May 29 the Cardiovascular and Renal Drugs Advisory Committee reviewed two drugs that "dissolve" blood clots. The drug streptokinase had been tested in a randomized clinical trial in Italy involving 11,806 patients. The death rate in those treated with streptokinase was 18% lower than in patients not given streptokinase; patients treated within six hours did even better. Review of 10 smaller studies, and early results of a large international study, also showed improved survival. It is important to know that the 18% reduction in death rate is a reduction of a few percent of the patients studied. The second drug considered—recombinant tissue plasminogen activator (TPA)—which also was clearly shown to dissolve blood clots, was not approved. Why? At least five issues contributed, to a greater or lesser amount, to the vote not to recommend approval for TPA at this time. These issues were: the safety of the drug, the completeness and adequacy of the data presented, the dose to be used, and the mechanism of action by which streptokinase (and hopefully TPA) saves lives.

Safety was the first and most important issue concerning TPA. Two formulations of TPA were studied at various doses; the highest dose was 150 milligrams. At this dose there was an unacceptable incidence of cerebral hemorrhage (that is, bleeding in the brain), in many case leading to both severe stroke and death. The incidence may be as high as 4% or as low as 1.5% to 2% (incomplete data at the meeting made it difficult to be sure of the exact figure), but in either case it is disturbingly high; this death rate due to side effects is of the same magnitude as the lives saved by streptokinase. This finding led the National Heart, Lung and Blood Institute to stop the 150-milligram treatment in a clinical trial. It is important to realize that this finding was unexpected, as TPA was thought to be relatively unlikely to cause such bleeding. Because of bleeding, the dose of TPA recommended by Genentech was reduced to 100 milligrams. The safety profile at doses of 100 milligrams looks better, but there were questions of exactly how many patients had been treated and evaluated fully. Relatively few patients getting this dose had been reported in full. Without complete reports from the studies there could be smaller strokes not reported and uncertainty as to how patients were examined. The committee felt a substantially larger database was needed to show safety.

The TPA used to evaluate the drug was manufactured by two processes. Early studies used the double-stranded (roller bottle) form of the drug; the sponsor then changed to a predominantly single-stranded form (suspension culture method) for marketing and production reasons. The second drug differed from the first in how long the drug remained in the blood, in peak effect, in the effect on fibrinogen and in the dose needed to cause lysis of clots. Much of the data was from the early form; these data were not considered very helpful with respect to the safety of the recommended dose of the suspension method drug. This could perhaps be debated, but the intracranial bleeding makes the issue an important one. The excessive bleeding may well prove to be a simple matter of excessive dose, but this is not yet known unequivocally.

Data were incomplete in that many of the patients' data had not been submitted yet and much of the data came from treatment with TPA made by the early method of manufacture. There was uncertainty about the data used to choose the 100-milligram dose, i.e., perhaps a lower dose is adequate. When there is a serious dose-related side effect it is crucial that the dose needed for effectiveness has been well-defined and has acceptable toxicity.

Let us turn to the mechanism of action, the means by which the beneficial effect occurs. There may be a number of mechanisms. The most compelling is clot lysis (dissolution). However, experts presented data that streptokinase changes the viscosity of the blood that could improve the blood flow; the importance is uncertain. Streptokinase also lowers blood pressure, which may decrease tissue damage during a heart attack. While there is convincing evidence that TPA (at least by the first method of manufacture) dissolves clots faster than streptokinase (at least after a few hours from the onset of the heart attack), we do not have adequate knowledge to know what portion of the benefit of streptokinase comes from dissolving the clot. TPA, thus, may differ in its effect on the heart or on survival. The drugs could differ in other respects, such as how often after opening a vessel they allow reclosure, and, of course, the frequency of important adverse effects.

These issues delay possible approval. Fortunately, more data are being collected. It is our sincere hope that the drug lives up to its promise, but should the drug prove as valuable as hoped, that would

> not imply the decision was wrong. The decision must be evaluated as part of the overall process of drug approval.
>
> The second editorial suggests that if the drug is not approved, Dr. Temple (director of the Bureau of Drugs, FDA), Dr. Young (FDA commissioner) and Dr. Bowen (secretary of health and human services) should administer "randomized mortality trials of heart-attack victims receiving the TPA clot buster or nothing." This indignant rhetoric seems inappropriate on several counts. First, the advisory committee has no FDA members; our votes are independent and in the past, on occasion, we have voted against the FDA's position. It is particularly inappropriate to criticize Drs. Temple and Young for the action of an independent group. The decision (by a vote of eight against approval, one for and two abstaining) was made by an independent panel of experts in cardiovascular medicine and research from excellent institutions. These unbiased experts reviewed the data presented and arrived at this decision; the FDA deserves no credit or blame. Second, we recommend approval of streptokinase; we are convinced that the drug saves lives of heart-attack victims (at least in the short term). To us it would be questionable to participate in a trial without some treatment in patients of the type shown to benefit from streptokinase. A better approach is to use streptokinase as an active control drug in a randomized trial. If it is as efficacious or better than streptokinase, we will rejoice. We have spent our adult lives in the care of patients and/or research to develop better methods for treatment. Both for our patients and our friends, our families and ourselves, we want proven beneficial drugs available.
>
> In summary, with all good therapeutic modalities the benefits must surely outweigh the risks of treatment. In interpreting the data presented by Genentech in May 1987 the majority of the Cardiovascular and Renal Drugs Advisory Committee members could not confidently identify significant benefits without concomitant significant risk. The review was clouded by issues of safety, manufacturing process, dose size and the mechanism of action. We are hopeful these issues will be addressed quickly, allowing more accurate assessment of TPA's risk-benefit ratio with conclusive evidence that treatment can be recommended that allows us to uphold the physician's credo, *primum non nocere* (first do no harm).

The July 28 1987, *USA Today's* Life section carried an article on the first page entitled "FDA Speeds Approval of Heart Drug." The article mentioned that the FDA commissioner Frank Young was involved in the data gathering. Within a few months of the advisory committee meeting, tPA was approved for use in treating myocardial infarctions. The drug was 5 to 10 times more expensive than streptokinase; however, it opened arteries faster and that was thought to be a potential advantage. A large randomized comparison of streptokinase and tPA was performed (ISIS 3); the preliminary results were presented at the November 1990 American Heart Association meeting. The conclusion was that the efficacy of the two drugs was essentially equivalent. Thus by approving streptokinase, even in retrospect, no period of the lack of availability of a clearly superior drug occurred because of the time delay needed to clear up the questions about tPA. This experience shows that biostatistical collaboration has consequences above and beyond the scientific and humanitarian aspects; large political and financial issues also are often involved.

20.4 OH, MY ACHING BACK!

One of the most common maladies in the industrialized world is the occurrence of low-back problems. By the age of 50, nearly 85% of humans can recall back symptoms; and as someone has said, the other 15% probably forgot. Among persons in the United States, back and spine impairment are the chronic conditions that most frequently cause activity limitation. The occurrence of industrial back disability is one of the most expensive health problems afflicting industry and its employees. The cost associated with back injury in 1976 was $14 billion; the costs are greatly skewed, with a relatively low percent of the cost accrued by a few chronic back injury cases [Spengler et al., 1986]. The costs and human price associated with industrial back injury prompted the Boeing Company to contact the orthopedics department at the University of Washington to institute a collaborative study of back injury at a Boeing factory in western Washington

State. Collaboration was obtained from the Boeing company management, the workers and their unions, and a research group at the University of Washington (including one of the authors, L.F.). The study was supported financially by the National Institutes of Health, the National Institute for Occupational Safety and Health, the Volvo Foundation, and the Boeing Company. The study was designed in two phases. The first phase was a retrospective analysis of past back injury reports and insurance costs from already existing Boeing records; the second phase was a prospective study looking at a variety of possible predictors (to be described below) of industrial back injury.

The retrospective Boeing data were analyzed and presented in a series of three papers [Spengler et al., 1986; Bigos et al., 1986a,b]. The analysis covered 31,200 employees who reported 900 back injuries among 4645 claims filed by 3958 different employees. The data emphasized the cost to Boeing of this malady, and as in previous studies, showed that a small percentage of the back injury reports lead to most of the cost; for example, 10% of the cases accounted for 79% of the cost. The incurred costs of back injury claims was 41% of the Boeing total, although only 19% of the claims were for the back. The most expensive 10% of the back injury claims accounted for 32% of all the Boeing injury claims. Workers were more likely to have reported an acute back injury if they had a poor employee appraisal rating from their supervisor within 6 months prior to the injury.

The prospective study was unique and had some very interesting findings (the investigators were awarded the highest award of the American Academy of Orthopedic Surgeons, the Kappa Delta award, for excellence in orthopedic research). Based on previously published results and investigator conjectures, data were collected in a number of areas with potential ability to predict reports of industrial back injury. Among the information obtained prospectively from the 3020 aircraft employees who volunteered to participate in the study were the following:

- *Demographics:* race, age, gender, total education, marital status, number in family, method, and time spent in commuting to work.
- *Medical history:* questions about treatment for back pain by physicians and by chiropractors; hospitalization for back pain; surgery for back injury; smoking status.
- *Physical examination:* flexibility; spinal canal size by ultrasonography; and anthropometric measures such as height and weight.
- *Physical capacities:* arm strength; leg strength; and aerobic capacity measured by a submaximal treadmill test.
- *Psychological testing:* the MMPI (Minnesota Multiphasic Inventory and its subscales); a schedule of recent life change events; a family questionnaire about interactions at home; a health locus of control questionnaire.
- *Job satisfaction:* subjects were asked a number of questions about their job: did they enjoy their job almost always, some of the time, hardly ever; do they get along well with their supervisor; do they get along well with their fellow employees, etc.

The details of the design and many of the study results may be found in Battie et al. [1989, 1990a,b] and Bigos et al. [1991, 1992a,b]. The extensive psychological questionnaires were given to the employees to be taken home and filled out; 54% of the 3020 employees returned completed questionnaires, and some data analyses were necessarily restricted to those who completed the questionnaire(s). Figure 20.4 summarizes graphically some of the important predictive results.

The results of several stepwise, step-up multivariate Cox models are presented in Table 20.1. There are some substantial risk gradients among the employees. However, the predictive power is not such that one can conclusively identify employees likely to report an acute industrial back injury report. Of more importance, given the traditional approaches to this field, which have been largely biomechanical, work perception and psychological variables are important predictors, and the problem cannot be addressed effectively with only one factor in mind. This is emphasized in Figure 20.5, which represents the amount of information (in a formal sense) in each of the categories of variables as given above. The figure is a Venn diagram of the estimated amount of predictive information for variables in each of the data collection areas [Fisher and Zeh, 1991]. The job perception and psychological areas are about as important as the medical history and physical examination areas. To truly understand industrial back injury, a multifactorial approach must be used.

Among the more interesting aspects of the study is speculation on the meaning and implications of the findings. Since, as mentioned above, most people experience back problems at

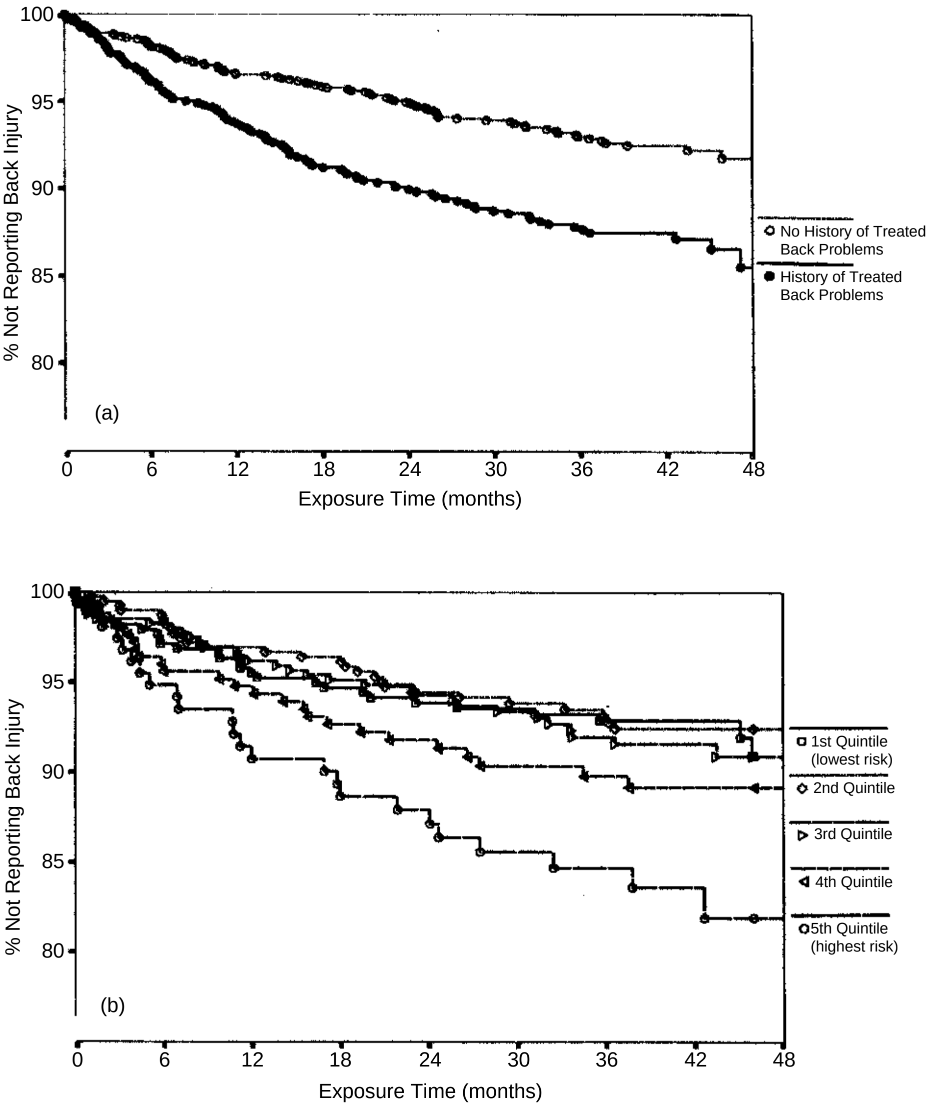

Figure 20.4 Panel (*a*) shows the product limit curves for the time to a subsequent back injury report for those reporting previous back problems and those who did not report such problems. Panel (*b*) divides the MMPI scale 3 (hysteria) values by cut points taken from the quintiles of those actually reporting events. Panel (*c*) divides the subjects by their response to the question: “Do you enjoy your job (1) almost always; (2) some of the time; or (3) hardly ever?” Panel (*d*) gives the results of the multivariate Cox model of Table 20.1; the predictive equation uses the variables from the first three panels. (From Bigos et al. [1991].)

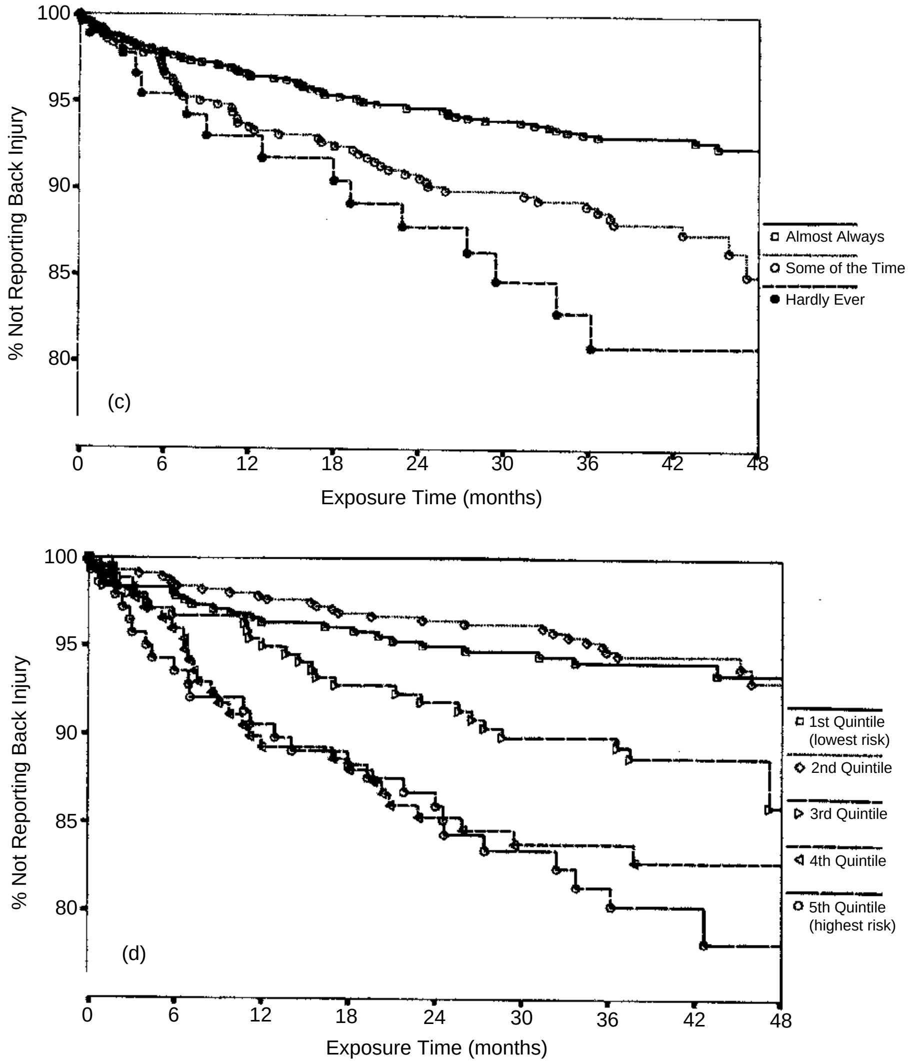

Figure 20.4 *(continued)*

some time in their lives, could legitimate back discomfort be used as an escape if one does not enjoy his or her job? Can the problem be reduced by taking measures to make workers more satisfied with their employment, or do a number of people tend to be unhappy no matter what? Is the problem a mixture of these? The results invite systematic, randomized intervention studies. Because of the magnitude of the problem, such approaches may be effective in both human and financial terms; however, this remains for the future.

20.5 SYNTHESIZING INFORMATION ABOUT MANY COMPETING TREATMENTS

Randomized controlled trials, discussed in Chapter 19, are the gold standard for deciding if a drug is effective and are required before new drugs are marketed. These trials may compare a

Table 20.1 Predicting Acute Back Injury Reports[a]

Variable	Univariate Analysis p-Value	Multivariate Analysis p-Value	Relative Risk	(95% Confidence Interval)
	Entire Population (n = 1326, injury = 117)			
Enjoy job[b]	0.0001	0.0001	1.70	(1.31, 2.21)
MMPI 3[c]	0.0003	0.0032	1.37	(1.11, 1.68)
Prior back pain[d]	0.0010	0.0050	1.70	(1.17, 2.46)
	Those with a History of Prior Back Injury (n = 518, injury = 63)			
Enjoy job[b]	0.0003	0.0006	1.85	(1.30, 2.62)
MMPI 3[c]	0.0195	0.0286	1.34	(1.17, 1.54)
	Those without a History of Prior Back Pain (n = 808, injury = 54)			
Enjoy jobs[b]	0.0220	0.0353	1.53	(1.09, 2.29)
MMPI 3[c]	0.0334	0.0475	1.41	(1.19, 1.68)

[a] Using the Cox proportional hazards regression model.

[b] Only subjects with complete information on the enjoy job question, MMPI, and history of back pain were included in these analyses.

[c] For an increase of one unit.

[d] For an increase of 10 units.

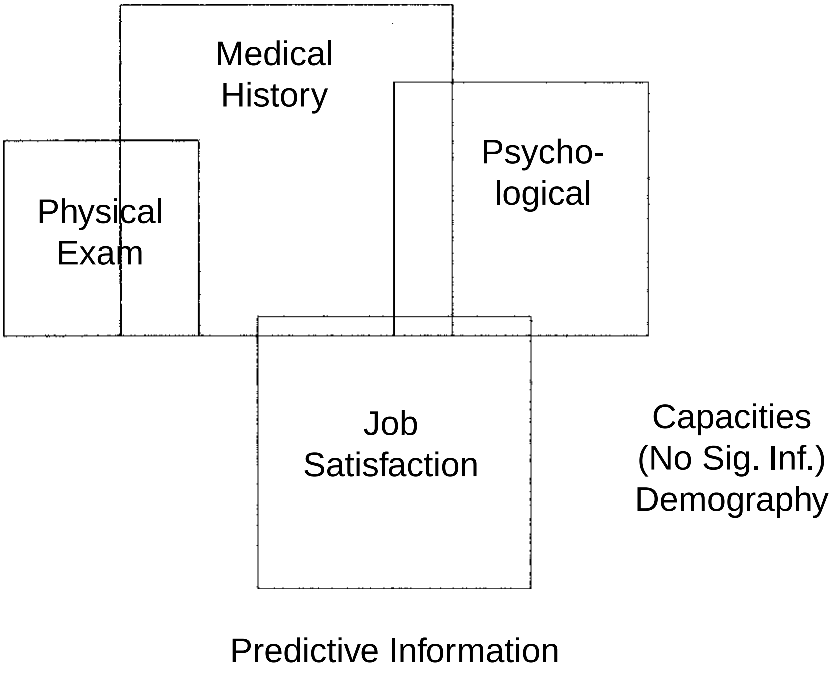

Figure 20.5 Predictive information by type of variable collected. Note that the job satisfaction and psychological areas contribute the same order of magnitude as the more classical medical history and physical examination variables. The relative lack of overlap in predictive information means that at least these areas must be considered if the problem is to be fully characterized. Capacities and demography variables added no information and so have no boxes.

new treatment to a placebo or to an accepted treatment. When many different treatments are available, however, it is not enough to know that they are all better than nothing, and it is often not feasible to compare all possible pairs of treatments in large randomized trials.

Clinicians would find it helpful to be able to use information from "indirect" comparisons. For example, if drug A reduces mortality by 20% compared to placebo, and drug B reduces mortality by 10% compared to drug A, it would be useful to conclude that B was better than placebo. However, indirect comparisons may not be reliable. The International Conference on Harmonisation, a project of European, Japanese, and U.S. regulators and industry experts, says in its document E10 on choice of control groups [2000, Sec. 2.1.7.4]

> "Placebo-controlled trials lacking an active control give little useful information about comparative effectiveness, information that is of interest and importance in many circumstances. Such information cannot reliably be obtained from cross-study comparisons, as the conditions of the studies may have been quite different."

The major concern with cross-study comparisons is that the populations being studied may be importantly different. People who participate in a trial of drug A when no other treatment is available may be very different from those who participate in a trial comparing drug A as an established treatment with a new experimental drug, B. For example, people for whom drug A is less effective may be more likely to participate in the hope of getting a better treatment. The ICH participants are certainly correct that cross-study comparisons *may* be misleading, but it would be very useful to know if they *are* actually misleading in a particular case.

An important example of this comes from the treatment of high blood pressure. There are many classes of drugs to treat high blood pressure, working in different ways on the heart, the blood vessels, and the kidneys. These include α-blockers, β-blockers, calcium channel blockers, angiotensin-converting enzyme (ACE) inhibitors, angiotensin receptor blockers, and diuretics. The availability of multiple treatments is useful because they have different side effects and because a single drug may not reduce blood pressure sufficiently. Some of the drug classes have the advantage of also treating other conditions that may be present in some people (β-blockers or calcium channel blockers for angina, α-blockers for the symptoms of prostatic hyperplasia). However, in many cases it is not obvious which drug class to try first.

Many clinical trials have been done, but these usually compare a single pair of treatments, and many important comparisons have not been done. For example, until late 2002, there had been only one trial in previously healthy people designed to measure clinical outcomes comparing ACE inhibitors with diuretics, although these drug classes are both useful in congestive heart failure and so seem a natural comparison. In a situation such as this, where there is reliable information from within-study comparisons of many, but not all, pairs of drugs, it should be possible to assess the reliability of cross-study comparisons and decide whether they can be used. That is, the possible cross-study comparisons of, say, ACE inhibitors and calcium channel blockers can be compared with each other and with any direct within-study comparisons. The better the agreement, the more confidence we will have in the cross-study comparisons. This technique is called *network metaanalysis* [Lumley, 2002]. The name comes from thinking of each randomized trial as a link connecting two treatments. A cross-study comparison is a path between two treatments composed of two or more links. If there are many possible paths joining two treatments, we can obtain an estimate along each path and see how well they agree.

The statistical model behind network metaanalysis is similar to the random-effects models discussed in Chapter 18. Write Y_{ijk} for a summary of the treatment difference in trial k of drugs i and j, for example, the logarithm of the estimated relative risk. If we could simply assume that trials were comparable, we could model this log relative risk by

$$Y_{ijk} = \beta_i - \beta_j + \epsilon_{ijk}$$

where β_i and β_j measure the effectiveness of drugs i and j, and ϵ_{ijk} represents the random sampling error.

When we say that trials of different sets of treatments are not comparable, we mean precisely that the average log relative risk when comparing drugs i and j is not simply given by $\beta_i - \beta_j$: there is some extra systematic difference. These differences can be modeled as random intercepts belonging to each pair of drugs:

$$Y_{ijk} = \beta_i - \beta_j + \xi_{ij} + \epsilon_{ijk}$$
$$\xi \sim N(0, \omega^2)$$

So, comparing two drugs i and j gives on average $\beta_i - \beta_j - \xi_{ij}$. If ξ_{ij} is large, the metaanalysis is useless, since the true differences between treatments $(\beta_i - \beta_j)$ are masked by the biases ξ_{ij}. The random effects standard deviation, ω, also called the *incoherence*, measures how large these biases are, averaged over all the trials. If the incoherence is large, the metaanalysis should not be done. If the incoherence is small, the metaanalysis may be worthwhile. Confidence intervals for $\beta_i - \beta_j$ will be longer because of the uncertainty in ξ_{ij}, slightly longer if the incoherence is very small, and substantially longer if the incoherence is moderately large.

Clearly, it would be better to have a single large trial that compared all the treatments, but this may not be feasible. There is no particular financial incentive for the pharmaceutical companies to conduct such a trial, and the cost would make even the National Institutes of Health think twice. In the case of antihypertensive treatments, a trial of many of the competing treatments was eventually done. This trial, ALLHAT [ALLHAT, 2002] compared a diuretic, a calcium channel blocker, an ACE inhibitor, and an α-blocker. It found that α-blockers were distinctly inferior (that portion of the trial was stopped early), and that diuretics were perhaps slightly superior to the other treatments.

Before the results of ALLHAT were available, Psaty et al. performed a network metaanalysis of the available randomized trials, giving much the same conclusions but also including comparisons with β-blockers, placebo, and angiotensin receptor blockers. This analysis, updated to include the results of ALLHAT, strengthens the conclusion that diuretics are probably slightly superior to the other options in preventing serious cardiovascular events [Psaty et al., 2003]. The cross-study comparisons showed good agreement except for the outcome of congestive heart failure, where there seemed to be substantial disagreement (perhaps due to different definitions over time). The network metaanalysis methodology incorporates this disagreement into confidence intervals, so the conclusions are weaker than they would otherwise be, but still valid.

The most important limitation of network metaanalysis is that it requires many paths and many links to assess the reliability of the cross-study comparisons. If each new antihypertensive drug had been compared only to placebo, there would be only a single path between any two treatments, and no cross-checking would be possible. Reliability of cross-study comparisons would then be an unsupported (and unsupportable) assumption.

20.6 SOMETHING IN THE AIR?

Fine particles in the air have long been known to be toxic in sufficiently high doses. Recently, there has been concern that even the relatively low exposures permitted by European and U.S. law may be dangerous to sensitive individuals. These fine particles come from smoke (wood smoke, car exhaust, power stations), dust from roads or fields, and haze formed by chemical reactions in the air. They have widely varying physical and chemical characteristics, which are incompletely understood, but the legal limits are based simply on the total mass per cubic meter of air.

Most of the recent concern has come from *time-series studies*, which are relatively easy and inexpensive to carry out. These studies examine the associations between total number of deaths, hospital admissions, or emergency room visits in a city with the average pollution levels.

As the EPA requires regular monitoring of air pollution and other government agencies collect information on deaths and hospital attendance, the data merely need to be extracted from the relevant databases.

This description glosses over some important statistical issues, many of which were pointed out by epidemiologists when the first studies were published:

1. There is a lot of variation in exposure among a group of people.
2. The monitors may be deliberately located in dirty areas to detect problems (or in clean areas so as not to detect problems).
3. The day-to-day outcome measurements are not independent.
4. There is a large seasonal variation in both exposure and outcome, potentially confounding the results.
5. We don't know how much time should be expected between exposure to fine particles and death or illness.

You should be able to think of several other potential problems, but a more useful exercise for the statistician is to classify the problems by whether they are important and whether they are soluble. It turns out that the first two are not important because they are more or less constant from day to day and so cancel out of our comparisons. The third problem is potentially important and led to some interesting statistical research, but it turns out that addressing it does not alter the results.

The fourth problem, seasonal variation, is important, as Figure 20.6 shows. In Seattle, mortality and air pollution peak in the winter. In many other cities the pattern is slightly different, with double peaks in winter and summer, but some form of strong seasonality is the rule. The

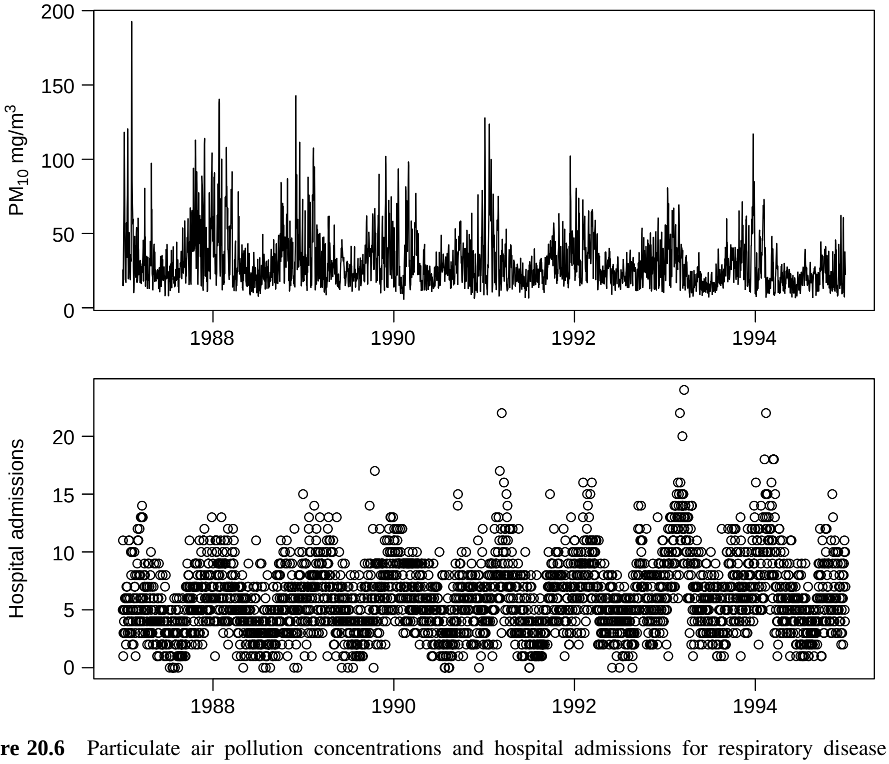

Figure 20.6 Particulate air pollution concentrations and hospital admissions for respiratory disease in Seattle.

solution to this confounding problem is to include these seasonal effects in our regression model. This is complicated: As gardeners and skiers well know, the seasons are not perfectly regular from year to year. Epidemiologists found a statistical solution, the generalized additive model (GAM), which had been developed for completely different problems, and adapted it to these time series. The GAM models allow the seasonal variation to be modeled simply by saying how smooth it should be:

$$\log(\text{mortality rate on day } t) = \alpha(t) + \beta \times \text{fine-particle concentration}$$

The smooth function $\alpha(t)$ absorbs all the seasonal variation and leaves only the short-term day-to-day fluctuations for evaluating the relationship between air pollution and mortality summarized by the log relative risk β. Computationally, $\alpha(t)$ is similar to the scatter plot smoothers discussed in Chapter 3.

With the problem of seasonal variation classified as important but soluble, analyses proceeded using data from many different U.S. cities and cities around the world. Shortly after the EPA had compiled a review of all the relevant research as a prelude to setting new standards, some bad news was revealed. Researchers at Johns Hopkins School of Public Health, who had compiled the largest and most systematic set of time-series studies, reported that they and everyone else had been using the GAM software incorrectly. The software had been written many years before, when computers were much slower, and had been intended for simpler examples than these time-series studies. The computations for a GAM involve iterative improvements to an estimate until it stops changing, and the default criterion for "stops changing" was not tight enough for the air pollution time-series models. At about the same time, researchers in Canada noticed that one of the approximations used in calculating confidence intervals and p-values was also not quite good enough in these time-series models [Ramsay et al., 2003]. When the dust settled, it became clear that the problem of seasonal variation was still soluble—fixes were found for these two problems, many studies were reanalyzed, and the conclusions remained qualitatively the same.

The final problem, the fact that the latency is not known, is just one special case of the problem of model uncertainty—choosing a regression model is much harder than fitting it. It is easy to estimate the association between mortality and today's pollution, or yesterday's pollution, or the previous day's, or the average of the past week, or any other choice. It is very hard to choose between these models. Simply reporting the best results is clearly biased, but is sometimes done. Fitting all the possible models may obscure the true associations among all the random noise. Specifying a particular model a priori allows valid inference but risks missing the true association. This final problem is important, but there is no simple mathematical solution.

20.7 ARE TECHNICIANS AS GOOD AS PHYSICIANS?

The neuropathological diagnosis of Alzheimer's disease (AD) is time consuming and difficult, even for experienced neuropathologists. Work in the late 1960s and early 1970s found that the presence of senile neuritic plaques in the neocortex and hippocampus justified a neuropathological diagnosis of Alzheimer's disease [Tomlinson et al., 1968, 1970]. Plaques are proteins associated with degenerating nerve cells in the brain; they tend to be located near the points of contact between cells. Typically, they are found in the brains of older persons.

These studies also found that large numbers of neurofibrillary tangles were often present in the neocortex and the hippocampus of brains from Alzheimer's disease victims. A tangle is another protein in the shape of a paired helical fragment found in the nerve cell. Neurofibrillary tangles are also found in other diseases. Later studies showed that plaques and tangles could be found in the brains of elderly persons with preserved mental status. Thus, the quantity and distribution of plaques and tangles, rather than their mere presence, are important in distinguishing Alzheimer's brains from the brains of normal aging persons.

A joint conference of 1985 [Khachaturian, 1985] stressed the need for standardized clinical and neuropathological diagnoses for Alzheimer's disease. We wanted to find out whether subjects with minimal training can count plaques and tangles in histological specimens of patients with Alzheimer's disease and controls [van Belle et al., 1997]. Two experienced neuropathologists trained three student helpers to recognize plaques and tangles in slides obtained from autopsy material. After training, the students and pathologists examined coded slides from patients with Alzheimer's disease and controls. Some of the slides were repeated to provide an estimate of reproducibility. Each reader read four fields, which were then averaged.

Ten sequential cases with a primary clinical and neuropathological diagnosis of Alzheimer's disease were chosen from the Alzheimer's Disease Research Center's (ADRC) brain autopsy registry. Age at death ranged from 67 years to 88 years, with a mean of 75.7 years and a standard deviation of 5.9 years.

Ten controls were examined for this study. Nine controls were selected from the ADRC registry of patients with brain autopsy, representing all subjects in the registry with no neuropathological evidence of AD. Four of these did have a clinical diagnosis of Alzheimer's disease, however. One additional control was drawn from files at the University of Washington's Department of Neuropathology. This control, aged 65 years at death, had no clinical history of Alzheimer's disease.

For each case and control, sections from the hippocampus and from the temporal, parietal, and frontal lobes were viewed by two neuropathologists and three technicians. The three technicians were a first-year medical school student, a graduate student in biostatistics with previous histological experience, and a premedical student. The technicians were briefly trained (for several hours) by a neuropathologist. The training consisted of looking at brain tissue (both Alzheimer's cases and normal brains) with a double-headed microscope and at photographs of tissue. The neuropathologist trained the technicians to identify plaques and tangles in the tissue samples viewed. The training ended when the neuropathologist was satisfied that the technicians would be able to identify plaques and tangles in brain tissue samples on their own for the purposes of this study. The slides were masked to hide patient identity and were arbitrarily divided into batches of five subjects, with cases and controls mixed. Each viewer was asked to scan the entire slide to find the areas of the slide with the highest density of plaques and tangles (implied by Khachaturian [1985]). The viewer then chose the four fields on the slide that appeared to contain the highest density of plaques and tangles when viewed at 25×. Neurofibrillary tangles and senile plaques were counted in these four fields at 200×. If the field contained more than 30 plaques or tangles, the viewer scored the number of lesions in that field as 30.

The most important area in the brain for the diagnosis of Alzheimer's is the hippocampus, and the results are presented for that region. Results for other regions were similar. In addition, we deal here only with cases and plaques. Table 20.2 contains results for the estimated number of plaques per field for cases; each reading is the average of readings from four fields. The estimated number of plaques varied considerably, ranging from zero to more than 20. Inspection of Table 20.2 suggests that technician 3 tends to read higher than the other technicians and the neuropathologists, that is, tends to see more plaques. An analysis of variance confirms this impression:

Source of Variation	d.f.	Mean Square	F-Ratio
Patients	9	102.256	—
Observers	4	—	—
Technicians vs. neuropathologists	1	21.31	2.70
Within technicians	2	42.53	5.39
Neuropathologist A vs. neuropathologist B	1	2.556	0.32
Patients × observers	36	7.888	—

Table 20.2 Average Number of Plaques per Field in the Hippocampus as Estimated by Three Technicians and Two Neuropathologists[a]

						Correlations:				
	Technician			Neuropathologist			Technician		Neuropathologist	
Case	1	2	3	A	B		2	3	A	B
1	0.75	0.00	0.00	0.00	0.00	1	0.69	0.63	0.65	0.76
2	7.25	6.50	7.50	4.75	3.75	2		0.77	0.79	0.84
3	5.50	7.25	5.50	5.75	8.75	3			0.91	0.67
4	5.25	8.00	14.30	5.75	6.50	A				0.82
5	10.00	8.25	9.00	3.50	7.75					
6	7.25	7.00	21.30	13.00	8.50					
7	5.75	15.30	18.80	10.30	8.00					
8	1.25	4.75	3.25	3.25	4.00					
9	1.75	5.00	7.25	2.50	3.50					
10	10.50	16.00	18.30	13.80	19.00					
Mean	5.25	7.80	10.50	6.26	6.98					
SD	3.44	4.76	7.21	4.60	5.08					

[a] Averages are over four fields.

You will recognize from Chapter 10 the idea of partitioning the variance attributable to observers into three components; there are many ways of partitioning this variance. The table above contains one useful way of doing this. The analysis suggests that the average levels of response do not vary within neuropathologists. There is a highly significant difference among technicians. We would conclude that technician 3 is high, rather than technician 1 being low, because of the values obtained by the two neuropathologists. Note also that the residual variability is estimated to be $\sqrt{7.888} = 2.81$ plaques per patient. This represents considerable variability since the values represent averages of four readings. Using a single reading as a basis produces an estimated standard deviation of $(\sqrt{4})(2.81) = 5.6$ plaques per reading.

But how shall agreement be measured or evaluated? Equality of the mean levels suggests only that the raters tended to count the same number of plaques on average. We need a more precise formulation of the issue. A correlation between the technicians and the neuropathologists will provide some information but is not sufficient because the correlation is invariant under changes in location and scale. In Chapter 4 we distinguished between precision and accuracy. *Precision* is the degree to which the observations cluster around a line; *accuracy* is the degree to which the observations are close to some standard. In this case the standard is the score of the neuropathologist and accuracy can be measured by the extent to which a technician's readings are from a 45° line. A paper by Lin [1989] nicely provides a framework for analyzing these data. In our case, the data are analyzed according to five criteria: location shift, scale shift, precision, accuracy, and concordance. *Location shift* refers to the degree to which the means of the data differ between technician and neuropathologist. A *scale shift* measures the differences in variability. *Precision* is quantified by a measure of correlation (Pearson's in our case). *Accuracy* is estimated by the distance that the observations are from the 45° line. *Concordance* is defined as the product of the precision and the accuracy. In symbols, denote two raters by subscripts 1 and 2. Then we define

$$\text{location shift} = u = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1 \sigma_2}}$$

$$\text{scale shift} = v = \frac{\sigma_1}{\sigma_2}$$

Table 20.3 Characteristics of Ratings of Three Technicians and Two Neuropathologists[a]

Technician	Pathologist	Location Shift	Scale Shift	Precision	Accuracy	Concordance
1	A	−0.18	0.75	0.95	0.94	0.89
	B	−0.35	0.68	0.76	0.88	0.67
2	A	0.33	1.03	0.79	0.95	0.75
	B	0.17	0.94	0.84	0.98	0.83
3	A	0.74	1.57	0.91	0.73	0.66
	B	0.58	1.42	0.67	0.81	0.55
A	B	−0.14	0.98	0.82	0.99	0.81

[a] Estimated numbers of plaques in the hippocampus of 10 cases, based on data from Table 20.2.

$$\text{precision} = r$$

$$\text{accuracy} = A = \left(\frac{v + 1/v + u^2}{2}\right)^{-1}$$

$$\text{concordance} = rA$$

We discuss these briefly. The location shift is a standardized estimate of the difference between the two raters. The quantity $\sqrt{\sigma_1\sigma_2}$ is the geometric mean of the two standard deviations. If there is no location difference between the two raters, this quantity is centered around zero. The scale shift is a ratio; if there is no scale shift, this quantity is centered around 1. The precision is the usual correlation coefficient; if the paired data fall on a straight line, the correlation is 1. The accuracy is made up of a mixture of the means and the standard deviations. Note that if there is no location or scale shift, the accuracy is 1, the upper limit for this statistic. The concordance is the product of the accuracy and the precision; it is also bounded by 1. The data in Table 20.2 are analyzed according to the criteria above and displayed in Table 20.3. This table suggests that all the associations between technicians and neuropathologists are comparable. In addition, the comparisons between neuropathologists provide an internal measure of consistency. The "location shift" column indicates that, indeed, technician 3 tended to see more plaques than the neuropathologists. Technician 3 was also more variable, as indicated in the "scale shift" column. Technician 1 tended to be less variable than the neuropathologists. The precision of the technicians was comparable to that of the two neuropathologists compared with each other. The neuropathologists also displayed very high accuracy, almost matched by technician 1 and 2. The concordance, the product of the precision and the accuracy, averaged over the two neuropathologists is comparable to their concordance. As usual, it is very important to graph the data to confirm these analytical results by a graphical display. Figure 20.7 displays the seven possible graphs.

In summary, we conclude that it is possible to train relatively naive observers to count plaques in a manner comparable to that of experienced neuropathologists, as defined by the measures above. By this methodology, we have also been able to isolate the strengths and weaknesses of each technician.

20.8 RISKY BUSINESS

Every day of our lives we meet many risks: the risk of being struck by lightning, getting into a car accident on the way to work, eating contaminated food, and getting hepatitis. Many risks have associated moral and societal values. For example, what is the risk of being infected by AIDS through an HIV-positive health practitioner? How does this risk compare with getting infectious hepatitis from an infected worker? What is the risk to the health practitioner in being identified as HIV positive? As we evaluate risks, we may ignore them, despite their being real

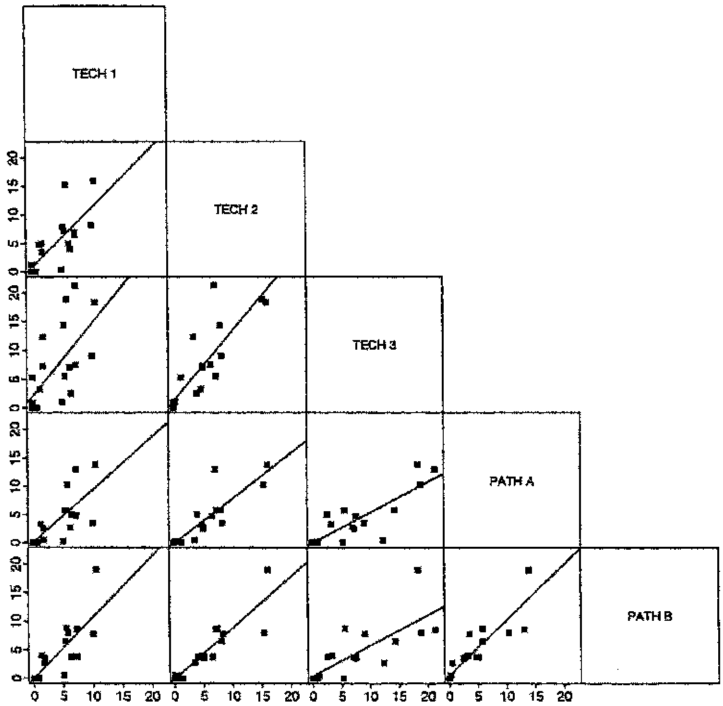

Figure 20.7 Seven possible graphs for the data in Table 20.3, prepared by SYSTAT, a very comprehensive software package. (From Wilkinson [1989].)

and substantial: for example, smoking in the face of the evidence in the Surgeon General's reports. Or we may react to risks even though they are small: for example, worry about being hit by a falling airplane.

What is a risk? A *risk* is usually an event or the probability of the event. Thus, the risk of being hit by lightning is defined to be the probability of this event. The word *risk* has an unfavorable connotation. We usually do not speak of the risk of winning the lottery. For purposes of this chapter, we relate the risk of an event to the probability of the occurrence of the event. In Chapter 3 we stated that all probabilities are conditional probabilities. When we talk about the risk of breast cancer, we usually refer to its occurrence among women. Probabilities are modified as we define different groups at risk. R. A. Fisher talked about *relevant subsets*, that is, what group or set of events is intended when a probability is specified.

In the course of thinking about environmental and occupational risks, one of us (G.vB.) wanted to develop a scale of risks similar to the Richter scale for earthquakes. The advantages of such a scale is to present risks numerically in such a way that the public would have an intuitive understanding of the risks. This, despite not understanding the full basis of the scale (it turns out to be fairly difficult to find a complete description of the Richter scale).

What should be the characteristics of such a scale? It became clear very quickly that the scale would have to be logarithmic. Second, it seemed that increasing risks should be associated with increasing values of the scale. It would also be nice to have the scale have roughly the same numerical range as the Richter scale. Most of its values are in the range 3 to 7. The *risk scale* for events is defined as follows: Let $P(E)$ be the probability of an event; then the risk units, $\text{RU}(E)$, for this event are defined to be

$$\text{RU}(E) = 10 + \log_{10}[P(E)]$$

Table 20.4 Relationship of Risk Units to Probabilities

Probability of Event	Risk Units
1	10
1/10	9
1/100	8
1/1000	7
1/10,000	6
1/100,000	5
1/1,000,000	4
1/10,000,000	3
1/100,000,000	2
1/1,000,000,000	1
1/10,000,000,000	0
1/100,000,000,000	−1

This scale has several nice properties. First, the scale is logarithmic. Second, if the event is certain, $P(E) = 1$ and $\text{RU}(E) = 10$. Given two independent events, E_1 and E_2, the difference in their risks is

$$\text{RU}(E_1) - \text{RU}(E_2) = \log_{10} \frac{P(E_1)}{P(E_2)}$$

that is, the difference in the risk units is related to the relative risk of the events in a logarithmic fashion, that is, a logarithm of the odds (see Table 20.4). Third, the progression is terms of powers of 10 is very simple; and so on. So a shift of 2 risk units represents a 100-fold change in probabilities. Events with risk units of the order of 1 to 4 are associated with relatively rare events. Note that the scale can go below zero.

As with the Richter scale, familiarity with common events will help you get a feeling for the scale. Let us start by considering some random events; next we deal with some common risks and locate them on the scale; finally, we give you some risks and ask you to place them on the scale (the answers are given at the end of the chapter). The simplest case is the coin toss. The probability of, say, a head is 0.5. Hence the risk units associated with observing a head with a single toss of a coin is $\text{RU(heads)} = 10 - \log_{10}(0.5) = 9.7$ (expressing risk units to one decimal place is usually enough). For a second example, the risk units of drawing at random a specified integer from the digits 0, 1, 2, 3, ..., 9 is 1/10 and the RU value is 9. Rolling a pair of sevens with two dice has a probability of 1/36 and are RU value of 8.4. Now consider some very small probabilities. Suppose that you dial at random; what is the chance of dialing your own phone number? Assume that we are talking about the seven-digit code and we allow all zeros as a possible number. The RU value is 3. If you throw in the area code as well, you must deduct three more units to get the value $\text{RU} = 0$. There are clearly more efficient ways to make phone calls.

The idea of a logarithmiclike scale for probabilities appears in the literature quite frequently. In a delightful, little-noticed book, *Risk Watch*, Urquhart and Heilmann [1984] defined the safety unit of an event, E, as

$$\text{safety unit of } E = -\log_{10}[P(E)]$$

The drawback of this definition is that it calibrates events in terms of safety rather than risk. People are more inclined to think in terms of risk; they are “risk avoiders” rather than safety

Table 20.5 The Risk Unit Scale and Some Associated Risks

Risk Unit	Event
10	Certain event
9	Pick number 3 at random from 0 to 3
8	Car accident with injury (annual)
7	Killed in hang gliding (annual)
6	EPA action (life time risk)
5	Cancer from 4 tbsp peanut butter/day (annual)
4	Cancer from one transcontinental trip
3	Killed by falling aircraft
2	Dollar bill has specified set of eight numbers
1	Pick spot on earth at random and land within $\frac{1}{4}$ mile of your house
0	Your phone number picked at random (+ area code)
−0.5	Killed by falling meteorite (annual)

Table 20.6 Events to Be Ranked and Placed on Risk Units Scale[a]

a.	Accidental drowning
b.	Amateur pilot death
c.	Appear on the *Johnny Carson Show* (1991)
d.	Death due to smoking
e.	Die in mountain climbing accident
f.	Fatality due to insect bite or sting
g.	Hit by lightning (in lifetime)
h.	Killed in college football
i.	Lifetime risk of cancer due to chlorination
j.	Cancer from one diet cola per day with saccharin
k.	Ace of spades in one draw from 52-card deck
l.	Win the *Reader's Digest* Sweepstakes
m.	Win the Washington State lottery grand prize (with one ticket)

[a] All risks are annual unless otherwise indicated. Events not ordered by risk.

seekers. But it is clear that risk units and safety units very simply related:

$$\mathrm{RU}(E) = 10 - \mathrm{SU}(E)$$

Table 20.5 lists the risk units for a series of events. Most of these probabilities were gleaned from the risk literature. Beside the events mentioned already, the risk unit for a car accident with injury in a 1-year time interval has a value of 8. This corresponds to a probability of 0.01, or 1/100. The Environmental Protection Agency takes action on lifetime risks of risk unit 6. That is, if the lifetime probability of death is 1/10,000, the agency will take some action. This may seem rather anticonservative, but there are many risks, and some selection has to be made. All these probabilities are estimates with varying degrees of precision. Crouch and Wilson [1982] include references to the data set upon which the estimate is based and also indicate whether the risk is changing. Table 20.6 describes some events for which you are asked to estimate the risk units. The answers are given in Table 20.7, preceding the References.

Table 20.7 Activities Estimated to Increase the Annual Probability of Death by One in a Million[a]

Activity	Cause of Death
Smoking 1.4 cigarettes	Cancer, heart disease
Drinking 0.5 liter of wine	Cirrhosis of the liver
Living 2 days in New York or Boston	Air pollution
Traveling 10 miles by bicycle	Accident
Living 2 months with a cigarette smoker	Cancer, heart disease
Drinking Miami drinking water for 1 year	Cancer from chloroform
Living 150 years within 5 miles of a nuclear power plant	Cancer from radiation
Eating 100 charcoal-broiled steaks	Cancer from benzopyrene

Source: Condensed from Wynne [1991].
[a] All events have a risk unit value of 4.

How do we evaluate risks? Why do we take action on some risks but not on others? The study of risks has been become a separate science with its own journals and society. The Borgen [1990] and Slovic [1986] articles in the journal *Risk Analysis* are worth examining. The following dimensions about evaluating risks have been mentioned in the literature:

Voluntary	Involuntary
Immediate effect	Delayed effect
Exposure essential	Exposure a luxury
Common hazard	“Dread” hazard
Affects average person	Affects special group
Reversible	Irreversible

We discuss these briefly. Recreational scuba diving has an annual probability of death of 4/10,000, or a risk unit of 6.6 [Crouch and Wilson, 1982, Table 7.4]. Compare this with some of the risks in Table 20.5. Another dimension is the timing of the effect. If the effect is delayed, we are usually willing to take a bigger risk; the most obvious example is smoking (which also is a voluntary behavior). If the exposure is essential, as part of one’s occupation, then again, larger risks are acceptable. A “dread” hazard is often perceived as of greater risk than a common hazard. The most conspicuous example is an airplane crash vs. an automobile accident. But perversely, we are less likely to be concerned about hazards that affect special groups to which we are not immediately linked. For example, migrant workers have high exposures to pesticides and resulting increased immediate risks of neurological damage and long-term risks of cancer. As a society, we are not vigorous in reducing those risks. Finally, if the effects of a risk are reversible, we are willing to take larger risks.

Table 20.7 lists some risks with the same estimated value: Each one increases the annual risk of death by 1 in a million; that is, all events have a risk unit value of 4. These examples illustrate that we do not judge risks to be the same even though the probabilities are equal. Some of the risks are avoidable; others may not be. It may be possible to avoid drinking Miami drinking water by drinking bottled water or by moving to Alaska. Most of the people who live in New York or Boston are not aware of the risk of living in those cities. But even if they did, it is unlikely that they would move. A risk of 1 in a million is too small to act on.

How can risks be ranked? There are many ways. The primary one is by the probability of occurrence as we have discussed so far. Another is by the expected loss (or gain). For example, the probability of a fire destroying your home is fairly small but the loss is so great that it pays to make the unfair bet with the insurance company. An unfair bet is one where the expected gain is negative. Another example is the lottery. A typical state lottery takes more than 50 cents from every dollar that is bet (compared to about 4 cents for roulette play in a casino). But the reward is so large (and the investment apparently small) that many people gladly play this unfair game.

Table 20.8 Answers to Evaluation of Risks in Table 20.5

	Risk Units	Source/Comments
a.	5.6	Crouch and Wilson [1982, Table 7.2]
b.	7.0	Crouch and Wilson [1982, Table 7.4]
c.	4.3	Siskin et al. [1990]
d.	7.5	Slovic [1986, Table 1]
e.	6.8	Crouch and Wilson [1982, Table 7.4]
f.	3.4	Crouch and Wilson [1982, Table 7.2]
g.	4.2	Siskin et al. [1990]
h.	5.5	Crouch and Wilson [1982, Table 7.4]
i.	4.0	Crouch and Wilson [1982, Table 7.5 and pp. 186–187]
j.	5.0	Slovic [1986, Table 1]
k.	8.3	$10 + \log(1/52)$
l.	1.6	From back of announcement; $10 + \log(1/250{,}000{,}000)$
m.	3.0	From back of lottery ticket; $10 + \log(1/10{,}000{,}000)$

How can risks be changed? It is clearly possible to stop smoking, to give up scuba diving, quit the police force, never drive a car. Many risks are associated with specific behaviors and changing those behaviors will change the risks. In the language of probability we have moved to another subset. Some changes will not completely remove the risks because of lingering effects of the behavior. But a great deal of risk reduction can be effected by changes in behavior. It behooves each one of us to assess the risks we take and to decide whether they are worth it.

The *Journal of the Royal Statistical Society*, Series A devoted the June 2003 issue (Volume 166) to statistical issues in risk communication. The journal *Risk Analysis* address risk analysis, risk assessment, and risk communication.

REFERENCES

Alderman, E. L., Bourassa, M. G., Cohen, L. S., Davis, K. B., Kaiser, G. C., Killip, T., Mock, M. B., Pettinger, M., and Robertson, T. L. [1990]. Ten-year follow-up of survival and myocardial infarction in the randomized Coronary Artery Surgery Study. *Circulation*, **82**: 1629–1646.

ALLHAT Officers and Coordinators [2002]. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic. The antihypertensive and lipid-lowering treatment to prevent heart attach trial (ALLHAT). *JAMA*, **288**: 2981–2997.

Battie, M. C., Bigos, S. J., Fisher, L. D., Hansson, T. H., Nachemson, A. L., Spengler, D. M., Wortley, M. D., and Zeh, J. [1989]. A prospective study of the role of cardiovascular risk factors and fitness in industrial back pain complaints. *Spine*, **14**: 141–147.

Battie, M. C., Bigos, S. J., Fisher, L. D., Spengler, D. M., Hansson, T. H., Nachemson, A. L., and Wortley, M. D. [1990a]. Anthropometric and clinical measures as predictors of back pain complaints in industry: a prospective study. *Journal of Spinal Disorders*, **3**: 195–204.

Battie, M. C., Bigos, S. J., Fisher, L. D., Spengler, D. M., Hansson, T. H., Nachemson, A. L., and Wortley, M. D. [1990b]. The role of spinal flexibility in back pain complaints within industry: a prospective study. *Spine*, **15**: 768–773.

Bigos, S. J., Spengler, D. M., Martin, N. A., Zeh, J., Fisher, L., Nachemson, A., and Wang, M. H. [1986a]. Back injuries in industry—a retrospective study: II. Injury factors. *Spine*, **11**: 246–251.

Bigos, S. J., Spengler, D. M., Martin, N. A., Zeh, J., Fisher, L., Nachemson, A., and Wang, M. H. [1986b]. Back injuries in industry—a retrospective study: III. Employee-related factors. *Spine*, **11**: 252–256.

Bigos, S. J., Battie, M. C., Spengler, D. M., Fisher, L. D., Fordyce, W. E., Hansson, T. H., Nachemson, A. L., and Wortley, M. D. [1991]. A prospective study of work perceptions and psychosocial factors affecting the report of back injury. *Spine*, **16**: 1–6.

Bigos, S. J., Battie, M. C., Fisher, L. D., Fordyce, W. E., Hansson, T. H., Nachemson, A. L., and Spengler, D. M. [1992a]. A longitudinal, prospective study of industrial back injury reporting in industry. *Clinical Orthopaedics*, **279**: 21–34.

Bigos, S. J., Battie, M. C., Fisher, L. D., Hansson, T. H., Spengler, D. M., and Nachemson, A. L. [1992b]. A prospective evaluation of commonly used pre-employment screening tools for acute industrial back pain. *Spine*, **17**: 922–926.

Borer, J. S. [1987]. t-PA and the principles of drug approval (editorial). *New England Journal of Medicine*, **317**: 1659–1661.

Borgen, K. T. [1990]. Of apples, alcohol, and unacceptable risks. *Risk Analysis*, **10**: 199–200.

CASS Principal Investigators and Their Associates [1981]. *National Heart, Lung, Blood Institute Coronary Artery Surgery Study*, T. Killip, L. D. Fisher, and M. B. Mock (eds.). American Heart Association Monograph 79. *Circulation*, **63**(p. II): I-1 to I-81.

CASS Principal Investigators and Their Associates: Coronary Artery Surgery Study (CASS) [1983a]. A randomized trial of coronary artery bypass surgery: survival data. *Circulation*, **68**: 939–950.

CASS Principal Investigators and Their Associates: Coronary Artery Surgery Study (CASS) [1983b]. A randomized trial of coronary artery bypass surgery: quality of life in patients randomly assigned to treatment groups. *Circulation*, **68**: 951–960.

CASS Principal Investigators and Their Associates: Coronary Artery Surgery Study (CASS) [1984a]. A randomized trial of coronary artery bypass surgery: comparability of entry characteristics and survival in randomized patients and nonrandomized patient meeting randomization criteria. *Journal of the American College of Cardiology*, **3**: 114–128.

CASS Principal Investigators and Their Associates [1984b]. Myocardial infarction and mortality in the Coronary Artery Surgery Study (CASS) randomized trial. *New England Journal of Medicine*, **310**: 750–758.

Chaitman, B. R., Ryan, T. J., Kronmal, R. A., Foster, E. D., Frommer, P. L., Killip, T., and the CASS Investigators [1990]. Coronary Artery Surgery Study (CASS): comparability of 10 year survival in randomized and randomizable patients. *Journal of the American College of Cardiology*, **16**: 1071–1078.

Crouch, E. A. C., and Wilson, R. [1982]. *Risk Benefit Analysis*. Ballinger, Cambridge, MA.

Fisher, L. D., Giardina, E.-G., Kowy, P. R., Leier, C. V., Lowenthal, D. T., Messerli, F. H., Pratt, C. M., and Ruskin, J. [1987]. The FDA Cardio-Renal Committee replies (letter to the editor). *Wall Street Journal*, Wed., Aug. 12, p. 19.

Fisher, L. D., Kaiser, G. C., Davis, K. B., and Mock, M. [1989]. Crossovers in coronary bypass grafting trials: desirable, undesirable, or both? *Annals of Thoracic Surgery*, **48**: 465–466.

Fisher, L. D., Dixon, D. O., Herson, J., and Frankowski, R. F. [1990]. Analysis of randomized clinical trials: intention to treat. In *Statistical Issues in Drug Research and Development*, K. E. Peace (ed.). Marcel Dekker, New York, pp. 331–344.

Fisher, L. D., and Zeh, J. [1991]. An information theory approach to presenting predictive value in the Cox proportional hazards regression model (unpublished).

International Conference on Harmonisation [2000]. *ICH Harmonised Tripartite Guideline: E10. Choice of Control Group and Related Issues in Clinical Trials. http://www.ich.org*

Kaiser, G. C., Davis, K. B., Fisher, L. D., Myers, W. O., Foster, E. D., Passamani, E. R., and Gillespie, M. J. [1985]. Survival following coronary artery bypass grafting in patients with severe angina pectoris (CASS) (with discussion). *Journal of Thoracic and Cardiovascular Surgery*, **89**: 513–524.

Khachaturian, Z. S. [1985]. Diagnosis of Alzheimer's disease. *Archives of Neurology*, **42**: 1097–1105.

Kowey, P. R., Fisher, L. D., Giardina, E.-G., Leier, C. V., Lowenthal, D. T., Messerli, F. H., and Pratt, C. M. [1988]. The TPA controversy and the drug approval process: the view of the Cardiovascular and Renal Drugs Advisory Committee. *Journal of the American Medical Association*, **260**: 2250–2252.

Lin, L. I. [1989]. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**: 255–268.

Lumley, T. [2002]. Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, **21**: 2313–2324

Myers, W. O., Schaff, H. V., Gersh, B. J., Fisher, L. D., Kosinski, A. S., Mock, M. B., Holmes, D. R., Ryan, T. J., Kaiser, G. C., and CASS Investigators [1989]. Improved survival of surgically treated patients with triple vessel coronary disease and severe angina pectoris: a report from the Coronary Artery Surgery Study (CASS) registry. *Journal of Thoracic and Cardiovascular Surgery*, **97**: 487–495.

Passamani, E., Davis, K. B., Gillespie, M. J., Killip, T., and the CASS Principal Investigators and Their Associates [1985]. A randomized trial of coronary artery bypass surgery: survival of patients with a low ejection fraction. *New England Journal of Medicine*, **312**: 1665–1671.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. L., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. [1977]. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: II. Analysis and examples. *British Journal of Cancer*, **35**: 1–39.

Preston, T. A. [1977]. *Coronary Artery Surgery: A Critical Review.* Raven Press, New York.

Psaty, B., Lumley, T., Furberg, C., Schellenbaum, G., Pahor, M., Alderman, M. H., and Weiss, N. S. [2003]. Health outcomes associated with various anti-hypertensive therapies used as first-line agents: a network meta-analysis. *Journal of the American Medical Association*, **289**: 2532–2542.

Ramsay, T. O., Burnett, R. T., and Krewski, D. [2003]. The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology*, **14**: 18–23.

Rogers, W. J., Coggin, C. J., Gersh, B. J., Fisher, L. D., Myers, W. O., Oberman, A., and Sheffield, L. T. [1990]. Ten-year follow-up of quality of life in patients randomized to receive medical therapy or coronary artery bypass graft surgery. *Circulation*, **82**: 1647–1658.

Siskin, B., Staller, J., and Rornik, D. [1990]. *What Are the Chances? Risk, Odds and Likelihood in Everyday Life*. Crown Publishers, New York.

Slovic, P. [1986]. Informing and educating the public about risk. *Risk Analysis*, **6**: 403–415.

Spengler, D. M., Bigos, S. J., Martin, N. A., Zeh, J., Fisher, L. D., and Nachemson, A. [1986]. Back injuries in industry: a retrospective study: I. Overview and cost analysis. *Spine*, **11**: 241–245.

Takaro, T., Hultgren, H., Lipton, M., Detre, K., and participants in the Veterans Administration Cooperative Study Group [1976]. VA cooperative randomized study for coronary arterial occlusive disease: II. Left main disease. *Circulation*, **54**(suppl. 3): III-107.

Tomlinson, B. E., Blessed, G., and Roth, M. [1968]. Observations on the brains of non-demented old people. *Journal of Neurological Science*, **7**: 331–356.

Tomlinson, B. E., Blessed, G., and Roth, M. [1970]. Observations on the brains of demented old people. *Journal of Neurological Science*, **11**: 205–242.

Urquhart, J., and Heilmann, K. [1984]. *Risk Watch: The Odds of Life.* Facts on File Publications, New York.

van Belle, G., Gibson, K., Nochlin, D., Sumi, M., and Larson, E. B. [1997]. Counting plaques and tangles in Alzheimer's disease: concordance of technicians and pathologists. *Journal of neurological Science*, **145**: 141–146.

Wall Street Journal [1987a]. The TPA decision (editorial). *Wall Street Journal*, Thurs., May 28, p. 26.

Wall Street Journal [1987b]. Human sacrifice (editorial). *Wall Street Journal*, Tues., June 2, p. 30.

Weinstein, G. S., and Levin, B. [1989]. Effect of crossover on the statistical power of randomized studies. *Annals of Thoracic Surgery*, **48**: 490–495.

Wilkinson, L. [1989]. *SYGRAPH: The System for Graphics*. SYSTAT, Inc., Evanston, IL.

Wynne, B. [1991]. Public perception and communication of risk: what do we know? *NIH Journal of Health*, **3**: 65–71.