CHAPTER 1

What Is Bias?

Two Red Sox fans were discussing the finer points of baseball strategy one day while driving to Fenway Park in Boston. Burt had read a statistical study about the effectiveness of the sacrifice bunt. In this maneuver, the batter tries to advance a base-runner from first to second base by tapping the ball a few feet in front of home plate. He is willing to be thrown out at first base in exchange for helping the runner to reach second base safely. The data in the study revealed that a runner on first base scored less frequently when the batter attempted to bunt. This implied, Burt insisted, that a batter should never attempt to sacrifice. Harry disagreed. Situations in which managers called for a sacrifice bunt, he argued, were not the same as those in which batters were allowed to swing away. Somehow, Harry knew intuitively that he was right and that some deeper principle of logic was involved, but he was never able to convince his friend.

Burt was unaware that by comparing the frequency of scoring between two different sets of at-bats, he was making a biased comparison. A lower success rate observed after attempting to bunt than when "swinging away" would not necessarily mean that bunting always, or even sometimes, *causes* a decrease in the probability of scoring the runner. Perhaps less proficient batters often bunt, whereas stronger hitters nearly always swing away. Then the success rate of the bunters would have been lower even if they had not bunted. So, was the lower observed success rate really caused by bunting?

The remainder of this book focuses on more consequential (and often controversial) analyses of causation that arise in many scientific contexts. In particular, we will concentrate on the problem of trying to reach a valid conclusion about some *factor* that might affect human health, behavior, or well-being. Sometimes we will denote this causal factor as F. For example, F might be an innovative educational program, and the outcome of interest some measure of academic achievement. Mathematically, we will treat F as

Bias and Causation: Models and Judgment for Valid Comparisons, By Herbert I. Weisberg

an indicator variable, with $F = 1$ if the causal factor is present and $F = 0$ if it is not.

This introductory chapter defines the problem of bias in a general sense. Bias is intrinsically a problem related to causality. We explain how traditional statistical methods are severely limited as a way to address causality in general, and questions related to bias in particular. Consequently, a new approach to data analysis is needed. Subsequent chapters describe a theoretical framework within which such a "new paradigm" has begun to evolve. For concreteness, this chapter includes six illustrative case studies that motivate and provide context for the ideas developed throughout the book.

1.1 APPLES AND ORANGES

Typically, the scientific community weighs the evidence provided by one or more *comparative studies* in order to decide whether a causal relationship between F and the outcome exists and to measure the strength of this effect. A comparative study examines some relevant aspect of a specified population of individuals. The aim is to ascertain whether and how a particular characteristic of individuals in the population (e.g., academic achievement) tends to respond when the factor is introduced, either deliberately (intervention) or unintentionally (risk factor). To provide statistical evidence, the study obtains data on individuals under two alternative conditions: exposure to the factor and nonexposure. Of course, the actual sets of individuals in the two groups being compared will differ. So, the critical question becomes whether the two *study groups* are sufficiently similar for the comparison to be interpreted as a *causal* effect of the factor, not as an "apples-to-oranges" comparison.

In a comparative study, a difference between groups that is not attributable to the factor under study can result from either *random* or *systematic* variability. Random variability can occur for a variety of reasons, but does not tend to favor the exposed or unexposed group. In large groups, these random variations tend to even out. If we imagine the size of the groups to increase without limit, the error in estimating the causal effect eventually becomes negligible. Moreover, in smaller groups, the amount of variability can at least be calculated and taken into account. Therefore, uncertainty related to random variability can be "managed" through statistical methods based on probability theory. These methods (significance testing, confidence intervals, regression modeling, Bayesian posterior distributions, etc.) represent the principal triumph of twentieth-century statistical theory.

Our focus in this book will be on the ways in which a comparison can be *systematically* (i.e., nonrandomly) distorted. An estimated effect that deviates systematically from the actual causal effect of interest is said to be biased. Unlike random variability, *bias* is a structural tendency that does not balance out, even with extremely large study groups. Unlike random error, bias cannot be reduced by increasing the sample size. In our baseball example, effects of random variation could be virtually eliminated if many thousands of at-bats were included in the analysis. We could therefore obtain a very precise answer to the question of whether runners tend to score less often after a bunt or not. But this information by itself would be of little practical value to a baseball manager, who wants to know when calling for a sacrifice will have a causal effect on the chances of scoring a run.

Throughout this book, the term *bias* will mean the extent to which a particular measure of a *causal effect* has been systematically distorted. Forms of bias that fall under this umbrella derive from shortcomings of research design, implementation, and analysis, and they can thus be considered *methodological biases*. To say that a particular study is biased is to assert that the research methods employed have resulted in systematic error in the estimation of a causal effect. Systematic error, or nonrandom error, is inherent in the research process itself. The magnitude and direction of bias do not depend on random variation across the particular sample of subjects included in the study.

When scientists refer to research bias, they generally mean methodological bias. However, discussions of bias are sometimes confusing because this term also has several other connotations. To a mathematical statistician, bias is a technical property of an estimate. An estimate of some parameter, such as the mean of a given population, is biased if the estimate “on average” deviates from the true value of the parameter. To a social scientist, bias may pertain to aspects of human behavior or psychology. Do certain individuals or groups tend to think or act in a predetermined manner in a specified situation? In addition, bias may suggest a negative or prejudicial attitude toward a particular group or ideology. As used throughout this book, bias is only incidentally related to any of these other interpretations.

Because they result from systematic and not random distortion, methodological biases are generally not amenable to correction by mathematical formulas. An understanding of potential biases in practice requires not only quantitative sophistication, but also a solid grounding in the relevant scientific context. The topic of bias resides in a kind of no-man’s-land between the discipline of statistics and the various scientific fields in which research takes place. This orphan status may help to explain why a comprehensive theory of bias has yet to emerge.

1.2 STATISTICS VS. CAUSATION

We have defined bias as a systematic error in estimating a causal effect based on statistical data. Attempts to estimate causal effects represent one of the most common, and arguably the most important, application of statistical methods. However, statistical theory, at least until quite recently, has been almost exclusively concerned with the implications of *random* error. As a result, classical statistical methods are applicable to a very narrow range of problems related to causal inference. Indeed, it is a universal mantra that statistical association, or correlation, does not necessarily imply causation. To the layperson, it must seem odd that statistics has so little to offer for learning about causal effects. To explain this irony, we must understand the primary problem that statistical methods were originally designed to address.

Classical statistical methods were devised primarily to deal with uncertainty that arises from the limited nature of the available data. Intuitively, it was recognized long ago that a small set of observations of some quantity was generally less reliable as a guide to action than a larger sample. For instance, a farmer might wish to learn how many apples he could expect to obtain from his orchard in a typical year, or perhaps in relation to factors such as rainfall and soil quality. Having data from many farms would provide much better information than relying on only a few. But better in what sense and by how much? The genius of modern statistical theory lies largely in its conceptual framework for formalizing and answering such questions.

Central to this conceptualization was the idea that the set of units in hand (e.g., apple orchards) could be imagined to comprise a representative "sample" randomly drawn from a much larger (virtually infinite) population of units. In principle, this hypothetical infinite population would include all of the units that could *potentially* have been observed, whether or not they were actually observed in the available sample. Furthermore, this population is assumed to possess a particular *distribution* of characteristics that can be described by the values of different variables (yield per acre, soil conditions, moisture, wind, etc.). This distribution essentially describes the proportions (or probabilities) of various possible values of the variables. The aim of statistical inference then becomes to describe the parameters (e.g., mean, median, variance, correlation) pertaining to this hypothetical population's distribution. For example, the farmer might wish to know the average yield of apples per acre and how this yield relates to the amount of rainfall during the growing season.

This statistical paradigm has by now become so familiar that it is hard to appreciate that it embodies certain *assumptions* about the world. First and foremost, there is the mental construct of a hypothetical infinite population. Moreover, the distribution of variables is often assumed to have a particular mathematical form, such as the (Gaussian) "normal" distribution. Buried even deeper, however, is another critical assumption: the probability distribution is regarded as stable, reflecting a fixed set of underlying conditions. Chance and uncertainty enter through the (assumed) process of randomly sampling from the population. However, because this variability is now subject to well-established mathematical rules of probability theory, a world of statistical inference opens up. For instance, the uncertainty associated with a small-sample estimate can be expressed as a confidence interval or a Bayesian posterior distribution. As long as the statistical model of the world remains fixed, inferences based on probability theory will be valid.

In particular, the implicit supposition of a stable universe allows the possibility of making accurate *predictions*. Our farmer may measure various conditions early in the growing season and then try to predict what his yield is likely to be. If relevant circumstances remain stable, and if he has a substantial database of prior observations, he can make a reliable forecast. This could be accomplished by effectively *conditioning* on the measured values he has observed. Suppose that the farmer magically knew the full distribution of yields per acre for the hypothetical apple-orchard population. Then he could identify all orchards that have approximately the same characteristics as his own. He could, for example, compute the average yield per acre *for this subgroup of the population*. That would be a logical value to predict for his current crop. In general, using these *conditional probabilities* is the basic idea underlying many sophisticated techniques for prediction. But the stability of the population distribution is what makes reliable prediction based on conditioning possible.

Now let us consider the problem of causal inference. Causal inference is also about making predictions. However, causation is not concerned primarily with random variation under a stable set of circumstances. Rather, causation pertains to what systematic alteration would occur if the circumstances were to change in a specified manner. For example, our farmer might be deciding whether to introduce an irrigation system. He wants to know what change in yield this innovation would *cause*. In effect, he envisions two hypothetical populations: one without irrigation and one with irrigation. The (causal) parameter of interest would then become the difference between the average yields produced in these two populations.

To answer causal questions, the classical statistical machinery just described is still necessary to cope with random variability. However, the strategy of conditioning is not adequate for causal inference. Conditioning for prediction depends on a stable set of circumstances, but analysis of causation entails consideration of a real or hypothetical modification of at least one important circumstance. Consequently, conditional probabilities within a fixed population cannot tell us what would happen under such alteration. For that, we must carry out (or at least envision) a *manipulation* of the circumstances. Classical statistical methods address random variation by invoking a stable hypothetical infinite population consisting of all units that *might have been* observed, whether or not they actually were observed. Similarly, causal inference requires a way to conceptualize what *might have been* observed under different specified circumstances. This central concept will be elaborated at length in Chapter 2. The key point for now is that the stable population assumed by traditional statistical methods can only reveal how various factors are *associated*, but it does not by itself disclose how a change in one factor would produce changes in some other factor of interest.

Bias in comparative studies has traditionally been either ignored by statisticians or addressed solely within the classical statistical framework. The result has been a failure to develop data-analytic methods capable of dealing appropriately with this pervasive methodological problem. Consequently, a conceptual framework for causal thinking that *extends* classical statistical theory is necessary to obtain a deeper understanding of bias. Such a causal framework has been evolving for roughly the past 30 years and has provided many of the building blocks needed for understanding the nature and sources of bias. A goal of this book is to draw together and elaborate those strands of this causal theory that pertain to the problem of bias in comparative studies.

Unlike some applications of this new theory, our primary goal is not to "solve" the problem of bias by offering more complicated or mathematically sophisticated statistical methods. Indeed, comprehending the nature and sources of bias can help to clarify why improved technology based on ever more complex mathematical analysis can be counterproductive. More mathematically "advanced" methods can even become an impediment to insight, because they remove the data analyst further from direct contact with the data. Consequently, the analyst may be forced to accept on faith that the assumptions underlying the statistical model are consistent with the data.

We will take the position that the sort of data-analytic tools required are those that will facilitate the exercise of logic and scientific judgment to reach conclusions that are supported by the weight of available evidence. Such methods typically cannot provide the degree of certainty or quantification familiar to us in managing random variability via standard statistical techniques. The successful development and application of causal knowledge ultimately depend on cultivation of sound scientific judgment, as well as basic mathematical facility, to discover what is likely, though by no means certain, to be true and real:

> A scientist's actions are *guided*, not determined, by what has been derived from theory or established by experiment, *as is his advice to others*. The judgment with which isolated results are put together to guide action or advice in the usual situation, which is too complex for guidance to be *deduced* from available knowledge, will often be a mixture of individual and collective judgments, but judgment will play a crucial role. Scientists know that they will sometimes be wrong; they try not to err too often, but they accept some insecurity as the price of wider scope. Data analysts must do the same. (Tukey, 1962, 9)

1.3 BIAS IN THE REAL WORLD

Statistics textbooks present mathematical techniques that can be applied in a variety of scientific areas. These statistical tools are almost exclusively devoted to the management of uncertainty attributable to random variability. It is therefore quite natural to consider these techniques in the abstract, with only minimal reference to the details of their application. A thorough knowledge of the context in which a particular procedure will be applied is not essential to understanding how and why it works. Training in the *application* of statistical methods is relegated largely to the various academic disciplines (epidemiology, economics, psychology, etc.) in which substantive scientific issues arise.

When dealing with the subject of bias, on the other hand, neatly severing theory from practice is not feasible. Certain general principles can be abstracted from the scientific context, but the motivation for these ideas cannot be grasped fully in the abstract. The relevant intellectual framework for thinking about bias has evolved primarily in the natural course of scientific research, and secondarily in generalizations made by philosophers of science observing this research. In particular, the concept of *causality* will be central to our discussion of bias throughout this book. Causal analysis would ideally be grounded in extensive background knowledge. For example, two anthropologists arguing about the effect of a certain cultural practice in various societies would share a foundation of theory and information necessary for a meaningful interchange. Obviously, we cannot hope to approach such a breadth and depth of contextual understanding. On the other hand, a theory of bias divorced completely from concrete scientific issues in the real world would be hopelessly sterile.

To partially address this conundrum, we present in this chapter a set of case histories. Each of these describes an actual study, or set of studies, to which we can refer later when discussing various sources of bias. The narratives offered here are necessarily somewhat sketchy. We attempt to highlight the main aspects of the research sufficiently to provide the reader with a tangible feel for the methodological challenges in making causal inferences. In selecting these case histories, several criteria have been taken into account. Each example pertains to an issue that was (or still is) considered to be important and subject to substantial uncertainty and disagreement. The range of these cases in terms of research design and subject matter is quite broad. Epidemiology, clinical trials, and social science research are all represented. Each of these narratives highlights a particular pivotal article or report that was central to the controversy. Most important, this set of studies allows us to illustrate a wide range of biases that either were or could have been considered by the investigators.

Throughout the book, we will draw upon these studies to provide context for various points, often introducing hypothetical elements. For instance, we may suggest a possible distorting influence that could theoretically have affected the study, even though there is no actual evidence to indicate that such a source of bias actually existed. When such liberties have been taken with the facts, the fictitious aspects will be noted. More substantive discussion of possible conclusions regarding the issues will be deferred to the final chapter.

The narratives presented here are intended in part to illustrate why statistical methods fail to address fully the range of methodological concerns related to bias and causation. Classical statistical methods are designed to provide answers to specific questions. Is this new medication superior to the standard treatment for a certain disease? Will this new educational approach improve academic performance? But the questions of practical interest are often much more particular, subtle, and complex. A practitioner may need to decide whether to try a new drug on her patient. A teacher may need to decide whether the new educational approach will work in her classroom. These practitioners may be interested in what is known about the possible causal effects (both beneficial and adverse) of these interventions for different kinds of individuals. Their decisions ultimately must be based on all the available evidence, both statistical and nonstatistical, that they can bring to bear, filtered through years of practical experience. For this purpose, they need information that is quantitative and rigorous, but also open-ended enough to connect with their richer base of qualitative knowledge in fruitful ways. To develop such information may call for a new analytic paradigm that is both more tolerant of ambiguity and more respectful of subject-matter knowledge.

Recently, there has been an explosion of interest in causal analysis within the field of statistics. This represents a very positive development, but there is a danger that causal inference will be reduced to just another mathematical–statistical technology. It would be regrettable if causal models were judged narrowly by their ability to solve statistical problems in the ways such problems have conventionally been formulated.

> But paradigm debates are not really about relative problem-solving ability, though for good reasons they are usually couched in those terms. Instead, the issue is which paradigm should in the future guide research on problems many of which neither competitor can yet claim to resolve completely. A decision between alternate ways of practicing science is called for, and in the circumstances that decision must be based less on past achievement than on future promise. The man who embraces a new paradigm at an early stage must often do so in defiance of the evidence provided by problem-solving. He must, that is, have faith that the new paradigm will succeed with the many large problems that confront it, knowing only that the older paradigm has failed with a few. A decision of that kind can only be made on faith. (Kuhn, 1962, 156–157)

This book was motivated by faith that a new paradigm for dealing with bias is possible, based on a deeper understanding of causality. This new paradigm will not reject the existing paradigm, but will define its limits, identifying the research questions for which it is not applicable. This new paradigm may provide some improved solutions to conventional problems, but its larger value will be in daring to pose and address novel questions. For example, traditional approaches concentrate almost exclusively on average effects, ignoring a largely unmet need to tailor effective interventions to individual characteristics and conditions. Causal theory has the potential to offer a proper language within which useful answers can be articulated, although the form of these answers may appear unfamiliar—and lacking the mathematical precision to which statisticians have become accustomed.

Evaluating the Efficacy of Antityphoid Vaccine

In a classic article, epidemiologist Mervyn Susser (1977) discussed the need for sound judgment grounded in subject-matter expertise to augment statistical analysis. The article referred to several historical examples, including one of the earliest uses of statistical analysis to evaluate the efficacy of a new medical treatment. The story begins in England in 1896, when Dr. Almroth Wright developed a vaccine to prevent typhoid fever. After several tests of the new vaccine among volunteers in the British Army, the Medical Advisory Office to the War Office was attempting to decide whether the army ought to adopt routine vaccination as a general policy. To aid in making this decision, the available data were submitted to Karl Pearson, the most eminent statistical expert of his day.

Pearson's analysis of the data led him to conclude that efficacy had not been firmly established; his published report suggested that more research was needed (Pearson, 1904). Wright took issue with Pearson's report in an accompanying editorial, and a heated debate in the pages of the *British Medical Journal* ensued. For a variety of reasons that went beyond the purely scientific issues, Wright "won" the debate and a policy of vaccination was adopted. However, a program of continued research was implemented, as Pearson had recommended. The results were summarized by Leishman, a colleague of Wright who directed the follow-up program (Leishman, 1910). The analysis appeared to provide strong support for the decision to implement inoculation in all units being sent overseas. Shortly after Leishman's data were released, the vaccine was adopted for routine use by the British, French, and American militaries. An article appearing in the *New York Times* hailed the success chronicled by Leishman as a triumph of modern medical science and its heroic practitioners:

> Trained scientists have labored weary hours without number in their laboratories bending over their microscopes and watching their test tubes to attain the golden truth. The result has been victory, a new triumph in the domain of medicine. It has not only been proved, say its champions that typhoid fever can be prevented by vaccination by anti-typhoid serum, but they claim immunity already has been conferred upon thousands and thousands of persons—soldiers chiefly—in this and other lands. (*New York Times*, June 5, 1910)

The data upon which the dispute between Pearson and Wright was based are presented in Tables 1.1 and 1.2, which we have adapted from Susser's summary (Susser, 1977). Table 1.1 speaks to the possible prophylactic effect of the vaccine. In each of the cohorts, the rate at which typhoid fever was contracted was lower among those inoculated with the vaccine than among those who were not. However, the magnitude of the rates and the difference between the two groups varied widely. Table 1.2 pertains to the question of whether the vaccine lowered mortality among those who contracted the disease. Here again, the mortality rates vary across cohorts. With one exception (Ladysmith garrison) the rates are lower in the inoculated group. Leishman's data are not presented here but show a similar and even stronger pattern of apparent effectiveness.

Table 1.1 Prophylactic Effect of Antityphoid Vaccine[a]

	Inoculated		Not Inoculated	
Cohort	*N*	Rate	*N*	Rate
Hospital staffs	297	10.8%	279	26.9%
Ladysmith garrison	1,705	2.1%	10,529	14.1%
Methuen's column	2,535	1.0%	10,981	2.3%
Single regiments	1,207	6.0%	1,285	6.4%
Army in India	15,389	0.8%	136,360	1.6%

[a]Adapted from Susser (1977).

Table 1.2 Effect on Mortality of Antityphoid Vaccine[a]

	Inoculated		Not Inoculated	
Cohort	*N*	Rate	*N*	Rate
Hospital staffs	32	6.3%	75	16.0%
Ladysmith garrison	35	22.9%	1489	22.1%
Single regiments	72	12.5%	82	25.6%
Special hospitals	1174	7.3%	4991	10.8%
Various military hospitals	764	8.2%	3374	10.8%
Army in India	84	13.1%	1475	28.7%

[a]Adapted from Susser (1977).

Our main purpose in presenting this data will be to consider possible sources of bias and the extent to which these biases may have compromised the studies. One obvious concern was that the various substudies all relied on data from soldiers who had volunteered for the experimental inoculation. Another pertained to the potential lack of reliability in diagnosis of typhoid fever at that time. Furthermore, the medical officers assigned to monitor the results were aware of whether a patient was or was not inoculated. In addition, the specificity of the treatment was subject to uncertainty about inoculation histories and lack of quality control in manufacturing the vaccine. Also, because soldiers were often transferred into or out of units, obtaining a valid count to use for a denominator in calculating the rates in particular units was complicated. Moreover, recording of the duration of exposure was also untrustworthy, because exposure status (inoculated or not) was recorded as of the end of the observation period. Finally, it is possible that the apparent effectiveness of the vaccine was attributable to other changes in personal hygiene or the water supply that were occurring at the same time (Cockburn, 1955).

With all these potential problems, most of which were recognized at the time, the true value of Wright's typhoid vaccine was far from certain. However, the public seemed to regard the matter as case closed. Major F. F. Russell of the Medical Corps of the U.S. Army speaking at Johns Hopkins was quoted at length in the *New York Times* article:

> Among the exposed regiments who had been inoculated with the vaccine in use at present there were 3.7 cases per 1,000 against 32.8 per 1,000 among the untreated. … The observation of this group of 12,000 men covers a period of over three years, and no more perfect or convincing statistics are needed to show the value of this method of prophylaxis. (As quoted in the *New York Times*, June 5, 1910)

A more sober and professional statistical analysis several years later came to a similar conclusion, while recognizing the methodological limitations of the existing data (Greenwood and Yule, 1915).

Despite whatever lingering doubts may have existed in the scientific community, Wright's antityphoid vaccine with various refinements remained in use without benefit of a controlled clinical trial for five decades. Considerable observational data accumulated attesting to reductions in typhoid incidence throughout the world that appeared to result from vaccination. Then in the 1950s the discovery of the antibiotic chloromycetin made possible a randomized test of typhoid vaccine, because those assigned to the control group who contracted typhoid fever could be cured. Fortunately, the vaccine was able to satisfy the more rigorous testing needed to receive the stamp of modern scientific validation (Cvjetanovic, 1957).

Racial Disparities in Death Sentencing

The death penalty is one of the most controversial issues related to the U.S. criminal justice system. In *Furman v. Georgia*, decided in 1972, the U.S. Supreme Court ruled essentially that the death penalty was being administered in a way that was arbitrary, capricious, and based on impermissible factors. Although not rejecting its use globally, the Court effectively set a higher standard for the manner in which death penalties could be imposed. The *Furman* decision led to reforms by many states aimed at avoiding the completely unstructured sentencing statutes that the Supreme Court had ruled unconstitutional (Baldus et al., 1990).

Since the *Furman* decision, lawyers seeking to overturn death penalty convictions have often argued that the decisions were disproportionate and/or discriminatory. Disproportionate would mean that the severity of the sentence was out of proportion to that received by other similarly situated defendants. Discriminatory would mean that the conviction and/or sentencing were tainted by impermissible factors, such as race or socioeconomic status. Many critics of the death penalty believed that judicial systems post-*Furman* remained permeated by racial discrimination and lack of proportionality in sentencing. Against this backdrop, David Baldus and his colleagues undertook two major interrelated studies of capital punishment in Georgia during the years 1973–1980.

Their main purposes were to estimate the extent of disproportionality and of racial discrimination in death-penalty decision making. These studies have been described in detail in a book titled *Equal Justice and the Death Penalty* (Baldus et al., 1990).

The second and larger of the studies was called the Charging and Sentencing Study (CSS). The CSS included 1066 cases from both the pre-*Furman* and post-*Furman* periods. These cases comprised a stratified random sample from a total of 2484 defendants "arrested and charged with homicide who were subsequently convicted of murder or voluntary manslaughter." Among the cases sampled, 127 resulted in a death penalty. For each of the 1066 cases, a wide array of variables was collected pertaining to five stages of the charging and sentencing process:

- Grand-jury indictment decisions
- Prosecutorial plea-bargaining decisions
- Jury guilt-trial decisions
- Prosecutorial decisions to seek a death penalty after conviction
- Jury penalty-trial sentencing decisions

Broadly speaking, the degree of discretion exercised by the decision-makers becomes more structured and constrained as a case moves through the process.

If the guilt trial results in a conviction for capital murder, the prosecutor must decide whether to seek the death penalty. Statutory criteria for potential death-eligibility are spelled out in general terms, but they must be interpreted by the prosecutor. If she believes that the death penalty is warranted, a second and entirely separate penalty trial will be held. The sole issue is to determine whether a death penalty should be imposed. To reach this decision, the penalty-trial jury is instructed to weigh specific aggravating and mitigating circumstances. To impose the death penalty, the jury must find at least one of the statutory aggravating factors. However, the jury is also permitted to consider any potentially mitigating factors.

One motive for undertaking the CSS was its potential use by attorneys representing convicted killer Warren McCleskey. McCleskey's death sentence had been imposed after his conviction for murdering a police officer named Frank Schlatt. In 1980, when the CSS study was first being considered, McCleskey's appeal was working its way through the Georgia legal system. An important basis for the appeal was McCleskey's assertion that the decision was tainted by racial discrimination; he was black and the victim white. Previous research in Georgia and elsewhere had suggested that both the race of the defendant and the race of the victim might play a role in death-sentencing decisions. A main goal of the CSS was to establish the extent to which death-sentencing decisions in Georgia had been influenced by race.

Based on their extensive database of cases in Georgia, Baldus and his team performed a variety of statistical analyses aimed at assessing possible

Table 1.3 Sentencing by Race of Defendant and Victim[a]

Category	Defendants	Death Sentences	Rate
Black on white	233	50	21.5%
White on white	748	58	7.8%
Black on black	1443	18	1.2%
White on black	60	2	3.3%
All cases	2484	128	5.2%

[a]Adapted from Baldus et al. (1990).

discrimination. Their book presents the data and analyses, along with details of their presentation in federal district court, and eventually to the U.S. Supreme Court. The book also deals extensively with various methodological issues raised during the appeals process and provides the authors' views on the validity of various criticisms. Rarely has a statistical study been subjected to so much scrutiny and with so much potentially at stake. Because issues of potential bias were dissected in great depth from both a statistical and a legal perspective, this case is highly instructive.

The basic data at issue can be summarized very simply. Table 1.3 shows the results of sentencing decisions in Georgia for cases in the post-*Furman* period (Baldus et al., 1990, 315). The unadjusted rates reveal some striking racial disparities, especially with respect to the victim's race. These rates were then adjusted in a variety of ways to account for the circumstances of the cases, especially as these pertained to the "moral culpability" of the defendant. One statistical model that was highlighted in court relied on a logistic regression that included 39 independent variables. This "core model" contained variables that both statistically and theoretically "appeared to exercise the greatest influence in determining which defendants indicted for murder would actually receive a death sentence." The coefficient (odds ratio) for the race-of-victim variable in this model was 4.3 and had a *p*-value of 0.005 (Baldus et al., 1990, 326).

As powerful as this statistical evidence appears to be, it did not carry the day. Eventually, the U.S. Supreme Court on April 22, 1987, in a 5–4 decision failed to overturn McCleskey's sentence. A part of the reasoning articulated by several of the justices was related to the appropriateness of *any* statistical argument. However, many specific criticisms of the methodology were also raised. In two long methodological appendices, the CSS investigators thoughtfully addressed these and other issues.

In terms of bias, there were three main areas of potential concern. One area related to the way that cases in the CSS had been selected. Only defendants convicted of voluntary murder were included, leaving out other potential death-penalty candidates whose cases reached other dispositions. Therefore, the potential for selection bias in estimating racial effects existed. A second problem related to the measurement of the culpability measures being used as covariates in the model. There were several aspects of the judicial system that made accurate and consistent data collection difficult. By far the most complex issue, however, related to the adequacy of the covariates collected, extensive as they were, to rule out other confounding factors. Was the observed difference in sentences truly the result of the victim's race, or alternatively of some other factors that were not measured but were correlated with race?

In Mr. McCleskey's case, further appeals on his behalf were put forward based on nonstatistical evidentiary grounds. To settle these, the case was eventually heard again by the Supreme Court. In the end, on September 26, 1991, Warren McCleskey was executed. In a *New York Times* editorial run 3 days later, the fourfold disparity estimated by David Baldus and his colleagues was prominently mentioned (*New York Times*, 1991).

Evaluation of Employment Training Programs

In the United States during the 1970s political support for government-sponsored social interventions to eliminate poverty and social inequity was strong. A number of major experimental educational and social programs were initiated, and the methodology of program evaluation became a major preoccupation of social scientists. The great majority of such government-sponsored efforts were observational (i.e., did not involve random assignment to different types of programs). Rather, subjects were assigned either to the innovative program being evaluated or to a more conventional control program according to some known criteria. For example, the program might be offered to those satisfying some needs-based eligibility criterion. Because this assignment mechanism was deliberate rather than random, the groups assigned to the different programs might be different in important respects.

In general, randomized experiments were not considered feasible in social program evaluation for a variety of ethical and practical reasons. However, a fortuitous exception to this limitation occurred in the area of worker training programs. The National Supported Work Demonstration (NSW) aimed to assist disadvantaged workers to enter the labor market successfully by providing work experience and counseling in a sheltered work environment (Dickinson and Maynard, 1981; Masters and Maynard, 1981). The target population for the NSW was composed of two main subgroups: women in the Aid to Families with Dependent Children (AFDC) program, and men who were high-school dropouts and often had a background that included drug addiction and criminal activity. Unlike most other government programs, individuals were selected for the available slots *randomly* from among a pool of qualified applicants, and the candidates who were not chosen became the controls.

Data related to annual income and various related socioeconomic and demographic individual characteristics were collected at baseline and at three follow-up points over 36 months. The postprogram income after 36 months was the primary outcome variable for the NSW. The official report of the study's findings found a very small impact of the NSW on the male participants and a fairly substantial improvement in earnings for the AFDC women (Manpower Demonstration Research Corporation, 1983). Because of the randomized design, these results were widely viewed as authoritative. Although there were some issues related to retention and compliance of participants that needed to be addressed, the statistical analysis was relatively straightforward. In contrast to the situation with other employment and training programs, no complex adjustments to deal with potential differences between the subjects who received the intervention and those who did not were necessary.

The existence of such a "gold standard" was viewed as a golden opportunity by econometrician Robert LaLonde. He wondered what would have been concluded if the NSW had relied on the more common quasi-experimental approach. To simulate such observational results, he created several different comparison groups based on available survey data. He then applied several alternative statistical techniques to obtain estimates based on these comparisons. Each of these statistical adjustments was based on a somewhat different mathematical model. LaLonde was primarily interested in whether any of the nonexperimental approaches could faithfully reproduce the "true" experimental findings. In addition, LaLonde asked whether it would be possible to discern bias in an observational study based on its own data, without reference to the gold standard. Specifically, would some violation of the adjustment model's assumptions be apparent to tip the researcher off that a problem existed? LaLonde's answers to these questions were published in an article that shook the econometric world:

> This study shows that many of the econometric procedures and comparison groups used to evaluate employment and training programs would not have yielded accurate or precise estimates of the impact of the National Supported Work Program. The econometric estimates often differ significantly from the experimental results. Moreover, even when the econometric estimates pass conventional specification tests, they still fail to replicate the experimentally determined results. Even though I was unable to evaluate all nonexperimental methods, this evidence suggests that policymakers should be aware that the available nonexperimental evaluations of employment and training programs may contain large and unknown biases from specification errors. (LaLonde, 1986, 617)

LaLonde's results, along with similar findings by Fraker and Maynard (1987), were hailed by methodologists who were strong advocates of randomized experiments (Burtless and Orr, 1986; Barnow, 1987). On the other hand, some econometricians and statisticians continued to defend nonexperimental studies as both necessary and viable (see Heckman and Hotz, 1989, with discussion; Heckman and Smith, 1995). In particular, James Heckman and his colleagues performed their own reanalyses of the NSW data and came up with estimates closer to those in the original randomized study. Their statistical models were selected based on the ability to pass certain tests of the model's assumptions. According to Heckman and Hotz, models that were not ruled out by these "specification tests" tended to perform quite well.

The debate over the ability of such specification tests to identify valid estimates in observational studies continues to this day. For example, Dehejia and Wahba (1999) have attacked the NSW problem from the perspective of propensity-score analysis (Rosenbaum and Rubin, 1983b). The merits of this relatively new approach have been debated (Smith and Todd, 2005a,b; Dehejia, 2005). How this controversy will ultimately play out is uncertain. What seems clear, however, is that proponents of observational research have been put on the defensive by those who argue that only true randomized experiments can yield reliable evidence of causality.

Phenylpropanolamine and Hemorrhagic Stroke

Phenylpropanolamine (PPA) was a component of many popular cold and cough medicines available over-the-counter prior to the year 2000. It was also used as an appetite suppressant. In several decades of use, few serious side effects had been observed. However, starting in the 1980s sporadic reports of hemorrhagic stroke (bleeding in the brain), particularly in young women, began to emerge. Epidemiologic evidence of a causal relationship was quite tenuous, but concerns persisted. By 1992, the Food and Drug Administration (FDA) decided to commission a large-scale case–control study to determine whether PPA was in fact implicated as a risk factor for hemorrhagic stroke.

The study, conducted by a team of researchers at Yale University, began in 1994 and was expected to require four years to complete. Results of the Yale Hemorrhagic Stroke Project (HSP) were reported to the FDA on May 10, 2000, and eventually published on December 21, 2000, in the *New England Journal of Medicine* (Kernan et al., 2000). The authors concluded that "phenylpropanolamine in appetite suppressants, and possibly in cold and cough remedies, is an independent risk factor for hemorrhagic stroke in women." The FDA indicated an intention to reclassify the drug as being unsafe for over-the-counter use and urged manufacturers to withdraw it from the market. All of the PPA producers agreed to this FDA request and made plans to transition to other medications for use in cold and cough products. Unfortunately for the manufacturers, the story did not end there, as a flood of lawsuits followed, brought by people who suffered strokes they believed were caused by PPA.

During the ensuing litigation, the HSP study's findings were hotly contested by the contending parties. Medical and epidemiological experts were retained by both sides. Experts brought in by the plaintiffs extolled the virtues of the HSP study as a model of scientific rigor. Experts for the defendant companies raised many serious methodological criticisms. As a result, the court records contain a wealth of information that sheds light on potential biases. To understand why the study was so controversial, it will be useful to describe briefly the study's design and the specific results obtained.

Because hemorrhagic stroke is such a rare event in young adults, it would be difficult to study its occurrence prospectively, waiting for a sufficient volume of cases to accumulate. So, the Yale investigators decided to conduct a "case–control" study. A case–control study is essentially run in reverse. The study starts by collecting a group of patients who have experienced the outcome event (e.g., hemorrhagic stroke) and then looks backward to identify factors that appear to be responsible. Loosely speaking, the method proceeds by comparing the rate of exposure to a particular risk factor (e.g., PPA) among the cases and among a group of noncases. If exposure is more prevalent among cases than among noncases, the risk factor may be a cause of the event. A more rigorous discussion of the case–control methodology will be presented in Chapter 4.

The HSP study enrolled 702 men and women 18–49 years of age who were recruited at 43 U.S. hospitals and had experienced a hemorrhagic stroke within 30 days prior to enrollment. For each of these cases, two control subjects were identified through the use of random-digit telephone dialing. Each of the two controls was matched with the corresponding case based on telephone exchange, race, sex, and age. For each case, the "focal time" was defined as the calendar day and time believed to mark the onset of the stroke-related symptoms. A focal time for each control was defined as the same day-of-week and time-of-day as the focal time of the matched case. Interviews of controls were conducted within 7 days of this focal time. Case and control interviews employed a structured questionnaire to obtain demographic, clinical, behavioral, and pharmaceutical data. Exposure to PPA was defined as use of a product containing PPA on the day of the stroke (prior to the event) or on any of the previous 3 days.

The analyses were performed using a technique called conditional logistic model for matched sets. This approach attempted to adjust for several other variables, in addition to the matching variables. The final model included an adjustment for hypertension, smoking status, and education. These factors were considered because each was believed to be associated with the occurrence of hemorrhagic stroke, either as a direct cause (hypertension, smoking) or indirectly as a surrogate for other unknown causal factors (education). Therefore, the statistical relationship between exposure and being a case of hemorrhagic stroke might be related to these "confounding factors" rather than a causal effect of PPA.

The analysis performed on the HSP data resulted in an estimated odds ratio (OR) of 1.49, with a p-value of 0.17. An odds ratio is a measure of effect that is roughly equivalent to the ratio of event rates with and without exposure. Thus, the study estimated a 49% increase in the frequency of hemorrhagic strokes attributable to PPA. However, the significance level of 0.17 was above the conventional 0.05 criterion commonly applied. The estimated odds ratio of 1.98 for women only was barely significant ($p = 0.05$), and the OR of 0.62 for men was not significant ($p = 0.41$).

For use of PPA in appetite suppressants, however, the HSP reported a whopping OR value of 16.58 ($p = 0.02$), among women (there was no male exposure to appetite suppressants), but based on only six exposed cases vs. one exposed control. Furthermore, a secondary analysis based on "first use" of PPA resulted in an odds ratio of 3.13 for women ($p = 0.08$). Here first use was defined as use of PPA within 24 hours of focal time, but no prior use within the past 2 weeks.

The HSP investigators read these results to suggest a causal association between PPA and hemorrhagic stroke in young women. Their report also acknowledged several possible sources of bias, and discussed measures taken that were believed to have minimized any problems. Confounding could have affected the estimated OR values, despite attempts to identify and correct for important confounding variables. Publicity about PPA might have influenced referral and diagnosis patterns: physicians could have preferentially identified as cases those who were thought to have consumed PPA (selection bias); case subjects could have had either clearer or less accurate memories of events just prior to the index date (recall bias). Finally, the report mentioned *temporal-precedence bias*, which can occur "when exposure is counted although the exposure occurs after the onset of the disease under study, often in response to disease symptoms" (Kernan et al., 2000). This concern was raised by awareness of a phenomenon known as *sentinel headaches* in which a transient headache may herald the onset of a stroke that is not recognized for hours, or even days. As a result, an individual who had used PPA after the sentinel headache but before the index date would be incorrectly regarded as exposed.

In the course of the litigation that followed, biostatistical experts retained by the PPA manufacturers raised these and several other potential biases as reasons to doubt the HSP conclusions (e.g., Weisberg, 2004). Some of their arguments are discussed later in this book to illustrate how various sources of bias can arise in a case–control study. In the majority of trials, the drug companies prevailed, deterring some plaintiffs from pursuing cases and motivating many others to settle for relatively modest amounts (Frankel, 2006). In the end, we will probably never know whether PPA really was responsible for causing strokes, as its removal from the market has made this question moot.

Postmenopausal Hormone Replacement Therapy and Cardiovascular Risk

Prior to 2002, estrogen supplementation was being used routinely by millions of postmenopausal women to control vasomotor symptoms (hot flashes, night sweats) and by many in the hope of reaping a variety of health benefits. The most common formulation of replacement hormones used in the United States consisted of conjugated equine estrogen, possibly in conjunction with progestin. Simplistically, because declining hormone levels were a natural concomitant of aging, replacement of the lost estrogen seemed to many women a logical step to help retain health and vitality. During the 1980s and 1990s a large number of observational studies appeared to confirm that hormone replacement therapy (HRT) did indeed provide a number of health benefits, in addition to generally effective relief of vasomotor symptoms. This "wonder drug" seemed to reduce the risk of osteoporosis, fractures, and cardiovascular disease and possibly even to slow progression toward dementia; the only known serious adverse effect was a possible slight increase in breast cancer. On balance, the profile of risks and benefits was generally considered quite favorable by hard-headed scientists as well as more subjective enthusiasts.

The only fly in this promising ointment was the lack of definitive evidence from randomized controlled trials. Skeptics argued that the observational data reflected potentially serious methodological weaknesses. Most significant was the suspected lack of comparability between women who were using HRT and those who were not; the HRT users appeared to be generally healthier and better educated than nonusers. Attempts were made in various ways to control for this healthy-user effect, but the success of these statistical adjustments was uncertain. To obtain more definitive answers, several large-scale randomized controlled trials were implemented during the 1990s. The focus of these efforts was on a range of health endpoints thought to be influenced by hormone levels. Of particular interest were cardiovascular outcomes, considered to be a major potential benefit of HRT.

In 1998, the results of a major clinical trial with a primary focus on cardiovascular disease were published (Hulley et al., 1998). The Heart and Estrogen/Progestin Replacement Study (HERS) was intended to evaluate HRT for secondary prevention in a cohort of postmenopausal women who had previous coronary heart disease (CHD). The main endpoint was occurrence of a serious CHD event (myocardial infarction or sudden death). The results were disappointing, as no overall difference between the treated and untreated groups emerged. Then in 2002, a much-anticipated randomized study of HRT for primary prevention of CHD yielded even more disturbing news. The Women's Health Initiative (WHI) study showed that in a large cohort of healthy postmenopausal women, HRT was associated with a modest *increase* in CHD events (relative risk of 1.29 overall), as well as elevated risk of breast cancer and stroke (Writing Group for the Women's Health Initiative Investigators, 2002). Although the study demonstrated that some beneficial effects accrued for other endpoints, the apparent harm caused by HRT for these serious adverse events clearly tilted the risk–benefit balance against routine use of HRT.

Other randomized studies, including a multiyear extension of HERS called HERS-II (Hulley et al., 2002) seemed generally to confirm the results of HERS and WHI that suggest either a neutral (e.g., Grady et al., 2002; Pentti et al., 2006) or harmful (Vickers et al., 2007) effect of HRT on CHD. The 180-degree turn between the observational studies and randomized experiments has created confusion among researchers and distress among women and their physicians. Hormone replacement therapy is still recommended for short-term relief of vasomotor symptoms, but not as a long-term regimen to promote good health. For the scientific community, it has been especially unsettling that observational studies seemed so convincing a few years ago, but apparently got the story completely wrong! Or did they? A variety of possible explanations have been offered by biostatisticians and clinical researchers.

Most of this methodological soul-searching accepts that the observational designs were flawed and tries to understand exactly why in order to avoid similar mistakes in the future. However, a substantial minority of methodologists refuse to accept the results of the randomized trials as gospel (e.g., Machens and Schmidt-Gollwitzer, 2003; Naftolin et al., 2004). These skeptics point to a number of methodological problems with the clinical trials. In particular, the possibility of selection bias has been suggested, based on the inclusion in both HERS and WHI of primarily older women who had not previously used HRT and were many years beyond menopause (Naftolin et al., 2004; van der Schouw and Grobbee, 2005). Other aspects of the eligibility criteria and screening process for entry into the trial may also have resulted in an unusual study population (Michels, 2003). Furthermore, it has been suggested that the particular hormone regimen (type and dose) utilized in HERS and WHI may not have been optimal, at least not for all women, and possibly different from that usually received in routine practice (Grodstein et al., 2003; Hoffman and Zup, 2003; Garbe and Suissa, 2004). So, it is possible that the restrictions imposed by the trials in order to enhance internal validity may have engendered a lack of external validity.

Although at present the pendulum has swung strongly away from long-term use of HRT, especially for cardioprotective purposes, there remains much uncertainty. Many still believe that HRT in some form can play a valuable role for some women under certain circumstances. The biological processes leading to CHD are complex, and the impact of hormonal supplementation may be highly variable across different individuals. If so, the challenge is not to determine simply whether or not to use HRT but when and for whom, and in what manner to apply this approach. A recent reanalysis of the WHI represents a potentially important step in this direction. This study found that, for relatively younger women within 10 years of menopause, the risk of CHD events was actually reduced (Roussouw et al., 2007). This finding reinforces the idea that HRT is safe for short-term use by newly menopausal women to relieve vasomotor symptoms.

Much of the recent research and controversy about HRT concerns the extent to which specific HRT formulations can safely provide health benefits, as well as alleviation of discomfort, to specific subgroups of women. For example, it has been hypothesized that women with more severe menopausal complaints may be those for whom HRT would tend to be most beneficial (van der Schouw and Grobbee, 2005). Identifying this or other markers of substantial benefit and low risk would be extremely helpful in practice and might reconcile apparently conflicting results of observational and controlled studies. Clinical research is also progressing with respect to various novel preparations that may provide the benefits of HRT without the alleged side effects of the conventional estrogen and estrogen/progestin regimens. For example, the synthetic steroid tibolone (Tib) that is used in Europe but not approved in the United States has shown promise in small-scale trials (Koh et al., 2005).

Finally, the idea of so-called *bioidentical* hormones has great appeal to many women. Bioidentical preparations are derived from plant extracts that have been chemically modified to be indistinguishable from hormones produced naturally in the body. These products are generally compounded by pharmacists who are not subject to FDA manufacturing regulations. Therefore, practices employed in compounding can vary widely. Some pharmacists customize prescriptions based on saliva tests or blood serum levels. The use of bioidentical hormone therapy is controversial. Advocates are swayed by the rationale that these products are "natural" and tuned to individual characteristics. They push for additional research, while being encouraged by the very limited scientific evidence available (e.g., Moskowitz, 2006). The medical research community, on the other hand, seems generally much more skeptical. Scientists tend to emphasize the lack of controlled trials, as well as the essential similarity of bioidentical and "synthetic" hormonal products (Fugh-Berman and Bythrow, 2007).

The hormonal changes that occur during and after menopause have profound and complex implications, but it has become clear that modifying or regulating these changes safely is not a simple matter. For biostatisticians, the efforts to understand when, how, and for whom HRT can be beneficial will continue to shed valuable light on the relative strengths and weaknesses of controlled trials and observational studies. From a methodological perspective, the impact of the WHI study on epidemiologists was similar to the impact of LaLonde's study on econometricians and other social scientists. The apparent reversal of what seemed a well-established body of knowledge shook the faith of many in the reliability of observational studies.

Antidepressants and Adolescent Suicide

During the 1990s a new generation of medications became widely available to treat major depression (MD) and anxiety disorders. Most of these new antidepressant drugs were in a class known as selective serotonin reuptake inhibitors (SSRIs). Initially, these new drugs were considered effective and safe, leading to rapidly expanding use that was thought by some to have played a role in observed decreases in population suicide rates (Olfson et al., 2003). However, some concerns began to surface in case reports and one clinical trial that these drugs might actually prompt suicidal thoughts and behavior in some patients, particularly adolescents. The existence of such an effect would be ironic, as suicidal tendencies can be a concomitant of MD that antidepressants are intended to treat. However, the emerging evidence was deemed sufficient by late 2003 to result in warnings by several European regulatory agencies. Then in October 2004 the U.S. Food and Drug Administration (FDA) delivered a *coup de grace* by ordering pharmaceutical companies to add a "black box" warning regarding possible risk of suicidality to the labeling of all antidepressants prescribed for pediatric use.

The FDA action was based primarily on a meta-analysis conducted to summarize the available evidence from randomized placebo-controlled trials on the risk of suicidality in adolescents who used modern antidepressants (U.S.

Food and Drug Administration, 2006; Hammad et al. 2006a,b). A meta-analysis is a type of study that produces an overall estimate of treatment effect by combining the results of several individual studies. Based on 24 clinical trials, the FDA meta-analysis found an approximate doubling of risk apparently attributable to the use of new-generation antidepressants. The regulatory actions by both U.S. and European authorities during 2003 and 2004 precipitated a rapid decline in prescriptions for these medications (Wheeler et al., 2009). The impact of this decrease on suicide rates is not yet clear, and the meaning of the available data is being debated (e.g., Leslie et al., 2005; Dubicka et al., 2006; Bridge et al., 2007). Psychiatric professionals are uncertain about the true balance of risks and benefits associated with the use of antidepressants. All agree that additional research is needed to better understand the circumstances, if any, when antidepressants may do more harm than good.

Much of the uncertainty derives from the limitations of the studies upon which the regulatory bodies based their decisions. In particular, the FDA meta-analysis has had great influence and been subject to much discussion. The major strength of this study is that it is based on placebo-controlled randomized trials, generally considered to be the gold standard of clinical research. However, a number of methodological problems have been pointed out by those who remain unconvinced that SSRIs and other antidepressants increase suicide risk. Some of these problems pertain to the paucity of relevant data in the clinical trials. Fortunately, there were no actual suicides in any of these study populations. So, the analyses of "suicidality" were based on indirect measures of relatively rare serious adverse events that reflected "suicidal behavior or ideation" as judged by a panel of experts (Hammad et al., 2006a,b). However, most of the trials were of short duration (4–16 weeks), so that even using this indirect proxy endpoint, very few events occurred.

Besides the small numbers of events in the clinical trials, there were two major potential sources of bias. First, the relationship between suicidality as measured in the studies and actual potential for self-harm is unclear. There was potential for inter-rater disagreement among the expert ratings of the adverse events. Second, the evidence contained in adverse-event reports might not have been adequate to allow accurate prediction of real suicidal intent.

An even more vexing issue pertains to the selection of study samples. Nearly all of the clinical trials on antidepressant use for adolescents attempted to exclude individuals who appeared at high prior risk for suicide. Several reviewers have noted that such screening could have affected the generalizability of the results (Dubicka and Goodyer, 2005; Greenhouse et al., 2008; Weisberg et al., 2009). Specifically, it is plausible that the observed relative risk was inflated by excluding some of those most likely to benefit from treatment with antidepressants. If so, it is conceivable that the regulatory actions may have been counterproductive by discouraging use of products that, properly monitored, could exert a net beneficial effect. From a methodological perspective, this situation highlights the difficult realities that arise in evaluating many complex interventions. Randomized controlled trials are geared primarily to establishing an overall or average treatment effect. When the effect on individuals can vary, perhaps even in direction as well as size, this overall effect can be misleading. Understanding when, how, and for whom antidepressants should be prescribed will require years to unravel. As in the HRT situation, there is much grist here for the methodological mills to grind.

GUIDEPOST 1

This chapter has introduced the topic of bias in comparative studies and presented several case studies that illustrate both the importance and the difficulties inherent in causal inference. These examples were presented against a backdrop of introductory ideas that emphasized the limitations of classical statistical theory for causal inference. We suggested that methods for dealing with bias must be built upon a deep understanding of the real problems posed by attempting to estimate causal effects. Out of such understanding, a new paradigm may emerge that draws on statistical theory but expands beyond its borders to better connect with subject-matter knowledge and clinical insight.

In the next chapter, we explain in more detail the theoretical basis for recent developments related to analysis of causation in comparative studies. The basic concepts of counterfactuals and potential outcomes are defined. The central idea of viewing human populations as collections of "response patterns" is introduced and illustrated with a simple hypothetical example. This idea then leads to a notion of "exchangeability" that, at least conceptually, solves the "apples and oranges" dilemma.