Statistical Science
1999, Vol. 14, No. 3, 259–304

Choice as an Alternative to Control in Observational Studies

Paul R. Rosenbaum

Abstract. In a randomized experiment, the investigator creates a clear and relatively unambiguous comparison of treatment groups by exerting tight control over the assignment of treatments to experimental subjects, ensuring that comparable subjects receive alternative treatments. In an observational study, the investigator lacks control of treatment assignments and must seek a clear comparison in other ways. Care in the choice of circumstances in which the study is conducted can greatly influence the quality of the evidence about treatment effects. This is illustrated in detail using three observational studies that use choice effectively, one each from economics, clinical psychology and epidemiology. Other studies are discussed more briefly to illustrate specific points. The design choices include (i) the choice of research hypothesis, (ii) the choice of treated and control groups, (iii) the explicit use of competing theories, rather than merely null and alternative hypotheses, (iv) the use of internal replication in the form of multiple manipulations of a single dose of treatment, (v) the use of undelivered doses in control groups, (vi) design choices to minimize the need for stability analyses, (vii) the duration of treatment and (viii) the use of natural blocks.

Key words and phrases: Causal effects, control groups, internal replication, observational studies, sensitivity analysis, stability analysis, treatment effects, undelivered doses.

1. INTRODUCTION AND SOME GROUND RULES

1.1 Active Observation: Control and Choice

Many observational studies do not succeed in providing tangible, enduring and convincing evidence about the effects caused by treatments, and those that do succeed often exhibit great care in their design. Particularly at the early stages of design, this care consists of choices that determine the circumstances of the study and the data to be collected. A convincing observational study is the result of active observation, an active search for those rare circumstances in which tangible evidence may be obtained to distinguish treatment effects from the most plausible biases. In an experiment, treatment effects are seen clearly because the environment is tightly controlled, whereas in a compelling observational study, control is, to a large extent, replaced by choice—the environment is carefully chosen. Studies of samples that are representative of populations may be quite useful in describing those populations, but may be ill suited to inferences about treatment effects. For instance, describing early work in public program evaluation, Donald Campbell (1988, page 324) wrote:

> There was gross overvaluing of, and financial investment in, external validity, in the sense of representative samples at the nationwide level. In contrast, the physical sciences are so provincial that they have established major discoveries like the hydrolysis of water... by a single water sample.

Passive observation of a natural population followed by regression analysis is often unsuccessful as an approach to inference about treatment effects; see, in particular, Box (1966) and Freedman (1997). A recent report of the National Academy of Sciences

Paul R. Rosenbaum is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6302.

(Meyer and Fienberg, 1992, page 106) concerning evaluation studies of bilingual education makes a similar point:

> In comparative studies, comparability of students in different programs is more important than having students who are representative of the nation as a whole. Elaborate analytic methods will not salvage poor design or implementation of a study.

Here, several principles are discussed which guide the choices made in the early planning and design of an observational study. The principles are illustrated using three observational studies that use choice effectively, one each from economics, clinical psychology and epidemiology. The studies are briefly described in Section 2 and the principles in Section 3. In Section 4, certain issues are restated in formal terms. Section 1.2 mentions certain ground rules in discussions of the quality of evidence.

1.2 The Quality of Evidence: Some Ground Rules

Most empirical disciplines make use of some form of mathematical reasoning. To discuss the quality of evidence provided by an empirical study one must first recognize that evidence is not proof, can never be proof, and is in no way inferior to proof. It is never reasonable to object that an empirical study has failed to prove its conclusions, though it may be reasonable, perhaps necessary, to raise doubts about the quality of its evidence. As expressed by Sir Karl Popper (1968, page 50): "If you insist on strict proof (or strict disproof) in the empirical sciences, you will never benefit from experience, and never learn from it how wrong you are."

One expects that a proof will be correct and that evidence will be candidly presented, despite occasional disappointments. One is often concerned about the relevance of a proof and the quality of evidence. If it is proved that a certain bridge can withstand certain forces, and yet the bridge collapses, it is not the correctness but the relevance of the proof that is likely to be called into question.

The distinction between the mathematical correctness and the relevance of a proof is familiar in most empirical sciences that use mathematical methods, for instance, economics, where Samuelson in a technical treatise (Samuelson, 1947), *The Foundations of Economic Analysis*, wrote:

> [O]ur theory is meaningless in the operational sense unless it does imply some restrictions upon empirically observable quantities, by which it could conceivably be refuted [page 7]. ......By a meaningful theorem I mean simply a hypothesis about empirical data which could be refuted, if only under ideal conditions. A meaningful theorem may be false [page 4].

Evidence may refute a theorem, not the theorem's logic, but its relevance. A related point is made by Quine (1951).

Evidence, unlike proof, is both a matter of degree and multifaceted. Useful evidence may resolve or shed light on certain issues while leaving other equally important issues entirely unresolved. This is but one of many ways that evidence differs from proof.

Evidence, even extensive evidence, does not compel belief. Rather than being forced to a conclusion by evidence, a scientist is responsible and answerable for conclusions reached in light of evidence, responsible to his conscience and answerable to the community of scientists. Of this, Michael Polanyi (1964) writes:

> [O]ur decision what to accept as firmly established cannot be wholly derived from any explicit rules but must be taken in the light of our own personal judgment of the evidence.
>
> Nor am I saying that there are no rules to guide verification, but only that there are none which can be relied on in the last resort....
>
> We may conclude that just as there is no proof of a proposition in natural science which cannot conceivably turn out to be incomplete, so also there is no refutation which cannot conceivably turn out to have been unfounded. There is a residue of personal judgment required in deciding—as the scientist eventually must—what weight to attach to any particular set of evidence in regard to the validity of a particular proposition [pages 30–31].
>
> The scientist takes complete responsibility for every one of these actions [page 40] ...for the process as a whole—he will assume full responsibility before his own conscience [page 46].

Evidence may be convincing beyond reasonable doubt; yet being convinced or remaining skeptical is a judgment for which a scientist is responsible. The desire to be compelled rather than convinced by evidence is the desire to evade responsibility for judging the evidence.

Aesthetics matter in proof, but also in evidence. A beautiful proof is simple, illuminating its conclusion and rendering it obvious. The same aesthetic applies to evidence. In both cases, the simple and the obvious are the product of care, effort and skill.

2. THREE OBSERVATIONAL STUDIES

2.1 Basis for Selecting These Studies, and a Limited Warranty

The following several issues guided the choice of studies.

(i) The studies employ choice in their design in an instructive manner which meaningfully strengthens the evidence, but of course does not prove, that the treatment caused its ostensible effects.

(ii) They come from three fields, economics, clinical psychology and epidemiology, that conduct many observational studies.

(iii) The treatments and outcomes under investigation in these studies are immediately intelligible and interesting to individuals outside these three particular fields.

A limited warranty and a few caveats are in order. In selecting these studies, I am not asserting that they have reached true conclusions about the effects of the treatments they study. Such an assertion is never warranted based on the results of a single observational study. Indeed, the economics example is currently the focus of fierce debate—that is part of the charm of this particular example. I am asserting simply that, in these studies, certain choices in design led to stronger evidence than would have been available under other circumstances, and that the same principles may be applied in other studies.

The studies have various implications for public policy, but it is their methodologies, specifically their use of choice in design, and not their policy implications that are under examination. This deserves emphasis, for it is easily misunderstood. Commonly, when evidence from varied studies is weighed, the goal is to examine the conclusions of those studies. That common goal is not the goal here. Although studies of the minimum wage, bereavement and occupational hazards will be discussed, no conclusions are reached about these subjects. *This paper concerns the relationship between choices in research design and the quality of the resulting evidence about effects caused by treatments.*

The discussion here of these studies is based on their published reports, so issues of data quality are not explored. Inspection of the data file for the minimum wage study hints that interviewers and respondents may, on occasion, have misunderstood one another, as happens in many if not all surveys. Issues of this sort may or may not be relevant to the conclusions of the three studies, but they are not relevant to the methodological issues under discussion here. Section 4.3 briefly discusses some reanalysis of data from two of the studies, comparing their sensitivity to hidden bias.

These three studies are useful for teaching about observational studies, where issues not addressed in this paper would also be discussed. In particular, adjustments for overt biases and sensitivity analyses for hidden biases are mentioned only briefly in Section 4.3 of the current paper, because these methods are discussed in detail elsewhere; however, the three studies can also be used to illustrate these techniques, which are central to the later, analytical parts of an observational study. Two of the studies are matched, and two use regression adjustments. The original published account of the epidemiology example includes its small data set together with a nonparametric analysis. The data for the larger economics example is publicly available electronically by FTP, or File Transfer Protocol; see Card and Krueger (1995, page 18). A sensitivity analysis for the epidemiology example is given in Rosenbaum (1993; see also 1995, Section 4) and requires only elementary calculations. A related analysis is given in Rosenbaum (1991). Parallel sensitivity analyses for both the epidemiology and economics examples are given in Rosenbaum (1999a); these are briefly discussed in Section 4.3 below. For various methods of sensitivity analysis, see Angrist, Imbens and Rubin (1996), Copas and Li (1997), Cornfield et al. (1959), Gastwirth (1992), Gastwirth, Krieger and Rosenbaum (1998), Greenhouse (1982), Manski (1990,1995), Marcus (1997), Rosenbaum and Rubin (1983b) and Rosenbaum (1986; 1987a; 1995a, Section 4; 1999a, b).

2.2 The Economics Example: The Effects of Increasing the Minimum Wage

The most familiar of all arguments of economic analysis are those of comparative statics, as discussed, for example, in the early chapters of Samuelson's (1947) *Foundations of Economic Analysis*. Often motivated with the aid of diagrams in elementary economics courses, such an argument pictures the world as determined by an equilibrium of forces, and it describes how that equilibrium changes as the forces are changed. Arguments of this form (cf. "Illustrative Market Case" in Samuelson, 1947, page 17) suggest that increasing the price of a commodity will decrease the demand for that commodity, other things "being equal." Viewing labor as a commodity and applying such considerations leads many economists to expect that an increase in the minimum wage will cause a decline in employment among workers receiving the minimum wage. Developed in slightly different terms, such arguments suggest that an increase in the minimum wage shifts production to use less labor and more capital equipment, or to substitute goods produced outside the reach of the minimum wage. For this reason, many economists argue that increasing the minimum wage hurts the very individuals it is intended to benefit. Inspection of the details of such a theorem suggest that, as a theoretical argument removed from data, this expectation seems logical.

Writing in the *American Economic Review*, David Card and Alan Krueger (1994; hereafter CK) attempted to estimate the effects of an increase in New Jersey's minimum wage that occurred on April 1, 1992. They looked at employment in the fast food industry (Burger King, Kentucky Fried Chicken, Wendy's and Roy Rogers) in New Jersey before and after the increase in the minimum wage and compared this to a control group consisting of fast food restaurants in adjacent, eastern Pennsylvania. They found "no evidence that the rise in New Jersey's minimum wage reduced employment at fast-food restaurants in the state." Methodological aspects of CK are nicely discussed by Meyer (1995).

The CK study is highly controversial in its conclusions and somewhat unusual in its design. To shed light on its design, it is useful to describe a related study conducted in the more traditional manner, which reached very different conclusions. The traditional study used here for comparison actually came in response to Card and Krueger and was conducted by Deere, Murphy and Welch (1995; hereafter DMW). The DMW study attempted to estimate the effects of the increases in the Federal minimum wage that took place from $3.35 to $3.80 on April 1, 1990, and from $3.80 to $4.25 on April 1, 1991, by applying regression to national, monthly data from the Current Population Survey from 1985 to 1993. Deere, Murphy and Welch conclude: "The regression estimates have no surprises. When the cost of employing low-wage laborers is increased, fewer low-wage laborers are employed." Without taking sides on the substantive conclusion, Section 3 will, at several places, compare the designs of these two studies.

Before leaving this introduction to the studies by CK and DMW, one should note that their conclusions do not, strictly speaking, contradict one another. First, CK discuss a particular change in a state minimum wage while DMW discuss a particular change in the Federal minimum wage, and it is possible in principle that these different interventions had different effects. Second, DMW examine changes in employment in certain demographic groups with varied percentages of minimum wage earners, while CK discuss employment in particular fast food restaurants. Note, however, that the simple comparative statics argument above would suggest that these distinctions should not matter and that the same general pattern of effects should be seen in both cases.

2.3 The Psychology Example: The Effects of Loss of a Spouse or Child in a Car Crash

Lehman, Wortman and Williams (1987; hereafter LWW) attempted to estimate the long-term psychological effects of a sudden and unexpected death of a spouse or a child in a motor vehicle crash. The study identified 39 individuals who had lost a spouse and 41 individuals who had lost a child in a motor vehicle crash four to seven years prior to the study. This group of exposed subjects was the result of a selection process that applied various criteria and sampling to a record of motor vehicle fatalities in Michigan between 1976 and 1979, with some nonresponse that is discussed in the LWW paper.

Noting that "many previous studies on the impact of bereavement have not included control or comparison groups," so that these studies were "difficult to interpret," LWW constructed a control group in the following way. From a reservoir of 7,581 individuals who came to renew their licenses, one control was matched to each exposed subject based on gender, age, family income in 1976 (i.e., before the crash), education level, number of children and ages of children. The outcomes included measures of depression and various psychiatric symptoms. Bereaved spouses and parents both were depressed substantially and significantly more than matched controls between four and seven years after the loss, and bereaved spouses exhibited higher rates of several other psychiatric symptoms.

Lehman, Wortman, and Williams concluded:

> The results presented here suggest that the current theoretical approaches to bereavement may need to be reexamined [page 228].... [The discussion goes on to contrast the study's results with the views of Bowlby and Freud, among others.] From 4 to 7 years after the sudden loss of a spouse or child, bereaved re-

> spondents showed significantly greater distress than did matched controls, suggesting little evidence of timely resolution. Contrary to what some early writers have suggested about the duration of the major symptoms of bereavement . . . both spouses and parents in our study showed clear evidence of depression and lack of resolution at the time of the interview, which was 4 to 7 years after the loss occurred . . . . The present study suggests that exposure to stress can trigger enduring changes in mental health and functioning [page 229].

2.4 The Epidemiology Example: Lead in Children of Workers Subject to Occupational Exposures to Lead

Morton et al. (1982; hereafter MSSORS) examined lead levels in the blood of children whose parents worked in a factory that used lead in the manufacture of batteries. They suspected that parents brought lead home in their clothes and hair, thereby exposing their children. Thirty-three children from different families with a parent at the battery factory were matched to 33 unexposed control children, the matching being based on age, neighborhood and exposure to traffic.

They found that exposed children had substantially higher levels of lead in their blood than did matched control children. The exposed children were also classified in two ways, namely, parental exposure to lead on the job (low, medium, high) and parental hygiene upon leaving the factory (good, moderately good, poor). They found lower levels of lead in the blood of exposed children whose parents had lower exposures to lead on the job, and lower levels of lead among children whose parents had better hygiene. Morton et al. (1982, page 555) concluded: “the data presented justify more stringent enforcement of lead containment practices.”

3. CHOICE IN THE DESIGN OF OBSERVATIONAL STUDIES

3.1 The Choice of Research Hypothesis: Narrow, Focused, Controlled Examination of a Broad Theory

Consider, first, a laboratory experiment in the physical or biochemical sciences. It begins with a broad theory that makes assertions about the effects of a treatment. Specifically, such a broad theory makes innumerable predictions about what would be observed in the innumerable circumstances in which the particular treatment might be applied, now and in the future, in different locations and so on. A laboratory experiment makes no effort to draw up a frame comprising all circumstances, locations and times when the treatment might be applied and to draw a representative sample of such circumstances. Rather, the laboratory experiment examines the theory under highly unrepresentative circumstances, namely, circumstances in which sensitive, calibrated measuring instruments are used in an environment carefully freed of forces that might intrude on the experiment, in which the treatment is delivered at doses sufficient to produce dramatic effects if the theory is correct or to produce an equally dramatic absence of effect if the theory is incorrect. In a way, it is part of the essence of a laboratory experiment that its circumstances are unrepresentative. In a well-conducted laboratory experiment, one of the rarest of things happens: the effects caused by treatments are seen with clarity.

Observational studies of the effects of treatments on human populations lack this level of control, but the goal is the same. Broad theories are examined in narrow, focused, controlled circumstances.

Broad theories are desired because they predict more, and in consequence are both more useful and can be more thoroughly scrutinized. Quoting Popper (1968) again:

> [T]hose theories should be given preference which can be most severely tested [page 121] . . . if the class of potential falsifiers of one theory is “larger” than that of another, there will be more opportunities for the first theory to be refuted by experience [. . . and . . .] the first theory says more about the world of experience than the second theory, for it rules out a larger class of basic statements . . . . Thus it can be said that the amount of empirical information conveyed by a theory, or its empirical content, increases with its degree of falsifiability [page 112—113) [which] explains why simplicity is so highly desirable . . . . [Simple theories] are to be prized more highly than less simple ones because they tell us more; because their empirical content is greater; and because they are better testable [page 142] . . . . Theories are not verifiable, but they can be “corroborated,” . . . we should try to assess what tests, what trials, [the theory] has withstood [page 251] . . . it is not so much the number of corroborating instances

> which determines the degree of corroboration as the severity of the various tests to which the hypothesis can be, and has been, subjected. But the severity of the tests, in its turn, depends upon the degree of testability, and thus upon the simplicity of the hypothesis: the hypothesis which is falsifiable in a higher degree, or the simpler hypothesis, is also the one which is corroborable in a higher degree [page 267].

Similar claims about simplicity, falsification and corroboration are made by Milton Friedman (1953, pages 8–9), who adds: "A hypothesis is important if it 'explains' much by little, that is, if it abstracts the common and crucial elements ... and permits valid predictions on the basis of them alone" (page 14) and "... the only relevant test of the validity of a hypothesis is comparison of its predictions with experience" (page 9). As noted by Putnam (1995, page 71), a similar point had been made earlier by Charles Sanders Peirce (1903, page 418–419):

> But if I had the choice between two hypotheses ... I should prefer ... [the one which] would predict more, and could be put more thoroughly to the test .... It is a very grave mistake to attach much importance to the antecedent likelihood of hypotheses .... Every hypothesis should be put to the test by forcing it to make verifiable predictions.

The terms "falsify," "reject" or "refute," when applied to scientific theories, are not quite accurate. Lakatos (1981, page 117) describes such a theory as "shelved," the implication being that the theory is stored along with the evidence against it. Under rare circumstances, a shelved theory might be reconsidered; see the quote from Polanyi in Section 1.2.

Broad theories permit close scrutiny in numerous particular cases, and a research hypothesis is intended to focus attention on one such case in which close scrutiny—a severe test—is possible. Passing such a severe test corroborates but does not prove the theory. Platt (1964), Meehl (1978) and Dawes (1996) make similar points.

In the economics example, broad arguments from comparative statics predict that increases in the cost of labor will diminish employment. If true, this theory operates in innumerable instances in the U.S. economy on a daily basis; however, in almost all such instances, it operates amid numerous other forces that obscure its effects. For instance, if wages rise faster in one company than in a second company making a different brand of a similar product, then the theory may be true even though employment does not decline in the first company, precisely because increasing demand for its brand of product may have led the first company to raise wages in an effort to increase output by expanding its workforce. Even if the theory is correct, it rarely operates in isolation. Card and Krueger attempted to find one of those rare instances in which the theory might be hoped to operate in relative isolation. Specific aspects of this isolation are discussed later, but the point here concerns their choice of research hypothesis. The effect of the increase in the New Jersey minimum wage on fast food chains in 1992 is, in itself, at most a very minor footnote to economic history. However, as an opportunity to scrutinize closely and thereby possibly to corroborate or refute the broad theory that forcible increases in the minimum wage cause declines in employment, the 1992 increase in New Jersey's minimum wage becomes much more important. In observational studies, one chooses a research hypothesis that permits a broad theory to operate dramatically and in relative isolation.

Some accounts in the popular press have emphasized, perhaps even slightly exaggerated, the breadth of the theory challenged by Card and Krueger's results. Writing in *Forbes*, the Stanford economist Thomas Sowell (1995) said:

> [I]f true, these results challenge the very foundations of economics. If rises in the price of labor do not reduce employment, why should we expect that a rise in the price of anything else affects the quantity purchased? This is to economics what disproving the law of gravity would be to physics.

Whether or not this is what CK did, whether or not they were really operating on this scale, nonetheless, the spirit of Sowell's remark is right: one seeks broad and consequential theories exposed to decisive challenges in focused, clear circumstances.

In the psychology example, the belief that bereavement should have short lived effects on mental functioning stems from a much broader theory, the dominant but not universally accepted theory in clinical psychology, that holds that the structure of mental functioning is largely shaped by a mixture of biology and experiences as a young child in relation to caretakers, usually parents. In particular, that theory suggests that bereavement should not greatly alter mental functioning over long periods of time. As in the economics example, if the broad theory is correct, then it is in constant operation in innumerable lives, but in almost all cases its operation is obscured by limited knowledge of the biology of mental functioning, by the difficulty in accurately measuring the experiences of early childhood and by the difficulty in distinguishing what is cause and what is consequence in the mental life and experiences of an adult. Sudden deaths from car crashes are a situation in which the broad theory operates or fails to operate with clarity.

Is it really true that a broad theory will be refuted or shelved based on a decisive challenge in a single, focused circumstance? Probably not. As argued by Lakatos (1970), a single, focused challenge may knock the wind out of a research program, may quicken interest in competing research programs, may stimulate further challenges in other focused circumstances, but it is unlikely to cause the immediate abandonment of a research program that has enjoyed some success—and this is for the best. Empirical studies, particularly observational studies, are attended by various uncertainties and ambiguities, some of which are difficult to quantify or even to identify, so consistent results in several studies are typically needed to force changes in the direction of a research program. Nonetheless, each such influential study is likely to examine the same broad theory in different focused circumstances, attended by different uncertainties.

In short, the choice of a research hypothesis focuses a broad theory on a narrow instance in which the theory's operation may be viewed clearly. Often, this setting permits the theory to operate in relative isolation from other forces, or to operate on a dramatic scale due to concentrated exposures, or else the treatment is imposed suddenly, at a discrete moment, in a manner not influenced by the individuals under study.

3.2 A Control Group

In designing an observational study, a key step is to choose a situation in which a control group can be constructed. A control group consists of subjects or units which did not receive the treatment. The control groups used in the three examples were described in Section 2.

Some observational studies do not have a control group. For instance, in the study by DMW, everyone covered by the Current Population Survey was in a region affected by the increase in the Federal minimum wage, so there is no control group—all subjects received the treatment. In contrast, the CK study of New Jersey's minimum wage had a control group consisting of branches of the same fast food chains across the Delaware river in eastern Pennsylvania where the minimum wage had not been increased.

Lacking a control group, DMW estimate the effect of the increase in the Federal minimum wage using regression in the following way. The Federal minimum wage increases went into effect on April 1, 1990, and April 1, 1991, so DMW define years which begin on April 1 and end on March 31. They focus on two groups of individuals, namely, teenagers aged 15–19 and adult high school dropouts aged 20–54, reasoning that these groups contain a disproportionate number of individuals earning the minimum wage, so they should be most affected by changes in the minimum wage. These two groups are examined in parallel but separate analyses, but, for brevity, only the analysis for teenagers is described here. Within these groups, men, women and blacks are examined in similar but not exactly parallel analyses; to permit a brief discussion here, the focus will be on the analysis for male teenagers. For each year from 1985 to 1992, for each state, there is an outcome measure, namely, the log of the fraction of male teenagers in the United States who were employed, as estimated by the current population survey. Deere, Murphy and Welch regress this outcome on the following predictors: (i) the log of the fraction of 15 to 64-year-old men who were employed in the same state and year, (ii) state indicator variables and (iii) indicator variables identifying the level of the Federal minimum wage. Concerning male teenagers, they write: "Compared to the employment level projected from the movement in aggregate employment with the $3.35 minimum wage, teenage employment was 4.8 percent . . . lower in 1990 . . . and 7.3 percent . . . lower in 1991–1992." In other words, employment among male teenagers in 1990 and 1991–1992 fell more sharply than employment among all 15 to 64-year-old men in the same period, adjusting for state-to-state differences that are constant through time. Qualitatively similar though numerically different results were obtained for females, blacks and adult high school dropouts. Their conclusion from these regressions was quoted in Section 2.2.

Conclusions reached in the absence of a control group are not necessarily wrong, but they are typically open to plausible objections and legitimate skepticism of types that are inapplicable with a control group. The assumption implicit in DMW's method of estimation is that changes in the log employment fraction for male teenagers would have been linearly related to changes in the log employment fraction among all 15 to 64-year-old men if the minimum wage had not been increased, and any departure from such a linear change is an effect caused by the change in the minimum wage. Is this assumption self-evidently true? Might not employment among teenagers and high school dropouts fall disproportionately more than employment in the general workforce during times of generally declining employment even without an increase in the minimum wage? Alternatively, a rise in the minimum wage might increase employment in certain demographic groups and decrease it in others because employers find that they can pull into the workforce better educated or more experienced workers who, at lower wages, would not work or who would work fewer hours. For instance, an anecdotal account in the popular press of changes in employment practices in response to the most recent increase in the Federal minimum wage describes an employer as introducing special hiring practices to avoid "wasting the extra 50 cents on unreliable help" (Duff, 1996). Even a bit of nonlinearity on the log scale might be mistaken for an effect of changing the minimum wage. Objections of this sort may well be incorrect and unfounded, but what is important here is that they can reasonably be raised for a study which lacks a control group.

If raising the minimum wage decreases employment, then, DMW reasoned, the larger decreases should occur in groups with more minimum wage earners. Although DMW focused on comparisons involving teenagers and adult dropouts, other groups with a disproportionate number of minimum wage earners were also examined briefly. For instance, low, medium and high wage states were compared, the first group being thought to be most impacted by increases in the minimum wage. Men and women were compared, where women as a group receive lower wages. These comparisons pointed in the opposite direction from the comparisons discussed in earlier paragraphs; see DMW's Table 3. Employment in low wage states declined less than employment in high wage states and employment among women declined less than employment among men after increases in the minimum wage. Women in low wage states experienced no decline in employment. If increasing the minimum wage decreased employment, the reasoning DMW applied to teenagers and dropouts would have predicted larger employment declines in low wage states and among women. Deere, Murphy and Welch (1995, page 234) write: "The latter fact is easily dismissed based on long-standing trends." Evidently, certain interactions do exist in which different demographic groups experience larger or smaller declines in employment, although DMW believe they can distinguish which effects are caused by the minimum wage and which are irrelevant demographic trends, and perhaps others will agree with them. Nonetheless, looking at the methodology, inconsistencies of this sort can arise in studies in which treated and control groups are replaced by groups with higher and lower exposure to the treatment.

An interesting paper by Holland and Rubin (1983) looks at various "paradoxes" that arise when there is no control group and a model or a calculation is used to estimate the treatment effect. They interpret the paradoxes as consequences of diverging, uncheckable and unarticulated assumptions about what the control group would look like.

If a study lacks a control group, a minimal requirement is the articulation of the assumptions about what the control group would look like, together with a discussion of the tangible evidence in support of those assumptions and the sensitivity of conclusions to violations of the assumptions. Articulation and evaluation of assumptions about the absent control group aid in judging the roles evidence and assumptions play in the study's conclusions.

3.3 Defining Treated and Control Groups: Sharply Distinct Treatments That Could Happen to Anyone

In a randomized experiment, treated and control groups are defined by, first, defining the treatments themselves and, second, defining and implementing a mechanism for random assignment of treatments. Typically, in prolonged experiments with human populations, two or at most a few quite distinct treatments are compared (Peto et al., 1976).

The situation in observational studies is different. The investigator does not control the assignment of treatments to subjects and so must define a treated and a control group using available subjects who have already received treatment or control. The goal in defining the treatment groups should be to produce a situation that resembles, to the extent possible, the situation in a randomized experiment: markedly distinct treatments that could happen to anyone. Consider the two parts separately.

In discussing the design of controlled trials, Peto et al. (1976, page 590) say:

> A positive result is more likely, and a null result is more informative, if the main comparison is of only 2 treatments, these being as different as possible.... [I]t is the mark of good trial design that a null result, if it occurs, will be of interest.

Treatment groups that are distinct in concept but not in actual implementation may not differ in their outcomes even if the conceived but unimplemented distinction would have an important effect on outcomes. For instance, this was a concern in a National Academy of Sciences report on studies of bilingual education which found that: "...Immersion and Early-exit Programs were in some instances indistinguishable from one another" (Meyer and Fienberg, 1992, page 102).

To say that the distinct treatments "could happen to anyone" is shorthand. One wishes to define the treated and control groups in such a way that the treatment could easily have happened to the controls, and the treated subjects could easily have been spared the treatment, the actual assignment of subjects to treatments being determined by haphazard or ostensibly irrelevant circumstances. Haphazard is not random, and haphazard treatment assignments can produce severe, consequential and undetected biases that would not be present with random assignment of treatments. Still, haphazard or ostensibly irrelevant assignments are to be preferred to assignments which are known to be biased in ways that cannot be measured and removed analytically.

An example of careful definition of treated and control groups comes from the study by LWW of the effects of a traumatic loss of a spouse or child. First, the treated and control conditions are markedly distinct. The loss is confined to a spouse or a child less than 18 years of age living at home, and the loss was produced suddenly by a car crash. One could study the effects of a loss of other relatives, such as an adult's parent or an adult's adult sibling, or gradual losses due to chronic disease, but these might have smaller psychological effects. As in experiments, effects should be demonstrated for markedly distinct treatments before refined studies of smaller effects are undertaken.

Second, LWW took care to define the treated group so it "could happen to anyone." In particular, they used published, relatively objective criteria to appraise probable responsibility or fault in the car crashes, and then insisted that the treated group consist of individuals from cars that were not at fault. For instance, if one car crossed a center dividing line and collided with an oncoming car, the occupants of the first car would be ineligible for the study while the occupants of the second car would be eligible. Their reasoning was that fault in car crashes is related to alcohol and drug use and to certain forms of psychopathology, all of which would be studied later as outcomes for survivors. In contrast, a car crash for which one is not responsible could happen to any driver. No matter what the particulars, car crashes are a far cry from random numbers, but car crashes for which one is not responsible are plausibly a limited but meaningful step closer to random.

In studying the effect of class size on academic achievement, Angrist (1997) gives another interesting example of sharply distinct treatments that could happen to anyone. In the United States, a class of size 40 will often be located in a very different school district than a class of size 20, so it is difficult to distinguish the effects of class size from the other consequences of the differences between school districts. In contrast, Israeli public schools implement the following version of a rule due to Maimonides: when class size exceeds 40 students, the class must be divided. In this case, a class of size 41 becomes two much smaller classes. In that setting, one might compare certain classes of size near 40 to split classes of size near 20, knowing that these are very different class sizes, and yet the rather minor event of enrolling one or two more students determined this dramatic change in class size. Again, this is not a random assignment of class sizes to students, but it is a meaningful step closer to random. Angrist and Krueger (1998) discuss additional examples.

Still another example of sharply distinct treatments that could happen to anyone is found in Bronars and Grogger's (1994) study of the economic consequences of unwed motherhood. Here, too, as a group, women who bear children prior to marriage differ from unmarried women who do not bear children, and it is important to avoid mistaking these differences from economic effects of an additional unplanned child. Instead of comparing these two groups, Bronars and Grogger compared unwed mothers who had twins to unwed mothers who had singletons, reasoning that having twins rather than a single child is a comparatively haphazard event, one that could happen to anyone. See also Rosenzweig and Wolpin (1980).

3.4 Competing Theories, Not Just Null and Alternative Hypotheses

In his essay, "How to be a good empiricist—a plea for tolerance in matters epistemological," Paul Feyerabend (1968, pages 14–15) writes:

> You can be a good empiricist only if you are prepared to work with many alternative theories rather than with a single point of view and "experience." This plurality of theories must not be regarded as a preliminary stage of knowledge which will at some time in the future be replaced by the One True Theory. Theoretical pluralism is assumed to be an

> essential feature of all knowledge that claims to be objective.... The function of such concrete alternatives is, however, this: They provide means of criticizing the accepted theory in a manner which goes beyond the criticism provided by a comparison of that theory "with the facts."... This, then, is the methodological justification of a plurality of theories: Such a plurality allows for a much sharper criticism of accepted ideas than does the comparison with a domain of "facts" which are supposed to sit there independently of theoretical considerations.

Elsewhere, Feyerabend (1975) argues that alternative theories are needed to unearth new facts, advising one to:

> Introduce and elaborate hypotheses which are inconsistent with well-established theories and/or well-established facts.... [T]he evidence that might refute a theory can often be unearthed only with the help of an incompatible alternative [page 29] many facts become available only with the help of alternatives, [so] the refusal to consider [alternative theories] will result in the elimination of refuting facts as well [page 42].

Popper (1965, page 112) makes a closely related point:

> A theory is tested not merely by applying it, or trying it out, but by applying it to very special cases—cases for which it yields results different from those we would have expected without that theory, or in the light of other theories. In other words we try to select for our tests those crucial cases in which we should expect the theory to fail if it is not true. Such cases are "crucial" in Bacon's sense; they indicate the cross-roads between two (or more) theories.

Lakatos (1981, pages 114–115) distinguishes the "internal" testing of theories from the "external" competition between theories, suggesting the latter is an aid to the former:

> [F]acts are only noticed if they conflict with some previous expectation. [This is a] cornerstone of Popper's psychology of discovery. Feyerabend developed another interesting psychological thesis of Popper's, namely, that proliferation of theories may—externally—speed up internal Popperian falsification.

A scientific theory that has been fairly successful—a theory discussed in journals and textbooks, explained to undergraduates, offered to graduate students as an area for research, and so on—is likely to agree with empirical observations at many points and to seem helpful in interpreting those observations. Certainly, sufficiently large, externally imposed increases in the price of a commodity *may* cause demand to decrease, and both biology and the events of early childhood *may* have enduring effects on mental functioning. Both of these theories offer coherent interpretations of frequent observations. A successful theory is likely to work in many situations—this is an aspect of its success. Unaided by a competing theory, an empirical investigation may do no more than rediscover why the successful theory achieved success in the first place. Such an investigation may not place the successful theory at much risk of refutation, and since it was never at much risk, failing to refute the theory provides little corroborating evidence in support of the successful theory.

A competing theory focuses attention on certain observable events about which the successful theory and its competitor make very different predictions. The competing theory directs attention to places where the successful theory might fail, placing the successful theory at severe risk of refutation. The competing theory anticipates a particular refutation of the successful theory, making refutation of the successful theory more likely and more decisive, and providing stronger corroboration of the successful theory if its predictions turn out to be correct. Card and Krueger quote the following remark of Paul Samuelson: "In economics it takes a theory to kill a theory; facts can only dent a theorist's hide" (Card and Krueger, 1995, page 355).

As discussed in Section 2.2, the conventional argument anticipates that an increase in the minimum wage in New Jersey drives up the cost of some labor in New Jersey, with two consequences for employment: first, an increase in prices of final products resulting in reduced demand; second, a tendency to substitute capital equipment, whose cost has not increased, for labor, whose cost has increased. The increase in the minimum wage affects all New Jersey firms, so the price increases may have smaller effects on a single firm's business than would be the case if that one firm raised prices while other firms did not. Card and Krueger (1995, page 359) express this same idea more formally as follows:

> For purposes of modeling the effect of an industry-wide wage increase, however, the relevant product–demand elasticity is one that takes into consideration simultaneous price adjustments at all firms. This elasticity will tend to be smaller (in absolute value) than the elasticity of demand for a firm's output with respect to its own price. In the case of the restaurant industry, for example, any individual restaurant presumably faces a relatively elastic demand for its product, holding constant prices at nearby restaurants. When the minimum wage increases, however, prices will tend to rise at all restaurants, resulting in a smaller net reduction in demand at any particular firm.

In other words, if Roy Rogers alone raised the price of hamburgers, it might face a substantial decline in business as customers switched to Burger King, but if all restaurants raise prices at the same time, the decline in business might be much smaller. Notice that this is particularly true of the fast food industry, because it is not practical for a New Jersey restaurant to import cooked food from, say, Hong Kong or Pennsylvania, to escape the effects of the minimum wage increase in New Jersey. If the conventional argument were incorrect, if increases in the minimum wage had only slight effects when simultaneous price adjustments occur in all firms, then the fast food industry, though unrepresentative of all industries, is one place to see this. Both theories accept that increases in the minimum wage might result in higher prices for fast food, but only one theory predicts a dramatic decline in business and employment in the fast food industry. In fact, Card and Krueger (1994, Sections 3E and 5) found no association between increases in the minimum wage and (i) the number of hours a restaurant is open on a weekday, (ii) the number of cash registers and (iii) the number of cash registers in operation at 11:00 am, but they did find some evidence of slightly higher increases in prices in New Jersey than Pennsylvania during this period. During the period of the minimum wage increases, CK (1994, Table 2) found that the average price of a specifically defined "full meal" changed from $3.04 to $3.03 in the Pennsylvania restaurants and from $3.35 to $3.41 in the New Jersey restaurants. In other words, the structure of a competing theory suggests where to look to test a standard theory, for instance, which industry to study and which outcome measures to examine.

Card and Krueger (1995) write: "We suspect that the standard model . . . does correctly predict the effect of the minimum wage on some firms" (page 355), but they say their results are "inconsistent with the proposition that the standard model is always correct" (page 383). Whether or not one agrees with CK about the minimum wage, the methodological issue is clear: a study of a narrow and unrepresentative corner of a population cannot, by itself, be the sole basis for policy for the whole population, but it may provide a severe test of a theory that purports to apply throughout the population. Failing such a test raises doubts about using that theory as the sole basis for policy, whereas passing such a severe test corroborates the theory and tends to strengthen the case for using it as a basis for policy.

Similar, if more direct, considerations apply in the LWW study. A theory which suggests that mental functioning is largely shaped by biology and early childhood experiences might reasonably be contrasted with a theory which holds that events later in life shape mental functioning over long periods. Sudden deaths of close relatives in car crashes are instances in which these theories make markedly different predictions.

In short, scrutiny of a theory is aided by a competing theory. The competition between theories suggests circumstances in which the theories make sharply different predictions about particular observable quantities.

3.5 Internal Replication: Multiple Treatment Assignment Mechanisms

Replication is important in observational studies, as it is in experiments. Susser (1987, page 88) writes:

> The epidemiologist's alternative to exact replication is the consistency of a result in a variety of repeated tests . . . . Consistency is present if the result is not dislodged in the face of diversity in times, places, circumstances, and people, as well as of research design.

In part, biases that are peculiar to the circumstances of one study may not replicate, whereas an effective treatment is expected to produce similar results in studies of varied circumstance and design.

Some observational studies incorporate a form of internal replication. These studies replicate the treatment assignment mechanism, so that essentially the same treatment is assigned to subjects by more than one process. As will be discussed in a moment, this is true of two of the studies in Section 2. If two treatment assignment mechanisms produce a pattern of associations consistent with an actual treatment effect, then to explain the pattern as a hidden bias, one must attribute a bias to both assignment mechanisms, and moreover a bias yielding the pattern anticipated from an actual effect. In Popper's terms, each mechanism provides a check on the theory that the treatment is the cause of its ostensible effects, so if several assignment mechanisms produce compatible estimates of effect, then this theory receives greater corroboration.

In the MSSORS study, children were exposed to low levels of lead in three different ways. First, the parents of control children did not work in the battery factory. Second, among exposed children whose parents did work in the battery factory, some were believed to have received lower doses of lead because their parents worked in jobs in the factory that provided little exposure to lead. Third, some parents exposed to high levels of lead practiced good hygiene. In fact, no matter which device assigned a child to low exposures to lead, the results were similar—children with lower exposures tended to have less lead in their blood. This pattern could, conceivably, be the result of hidden bias, but it is somewhat more difficult, though of course not impossible, to imagine biases that would produce all three associations.

In the CK study of the increase in the New Jersey minimum wage, a fast food restaurant could escape a legal requirement to increase wages in either of two ways. Restaurants in Pennsylvania were not required to increase wages. Restaurants in New Jersey whose lowest wage was above the new minimum wage were not required to increase wages. In connection with their Table 3, CK (1994, page 778) write: "Within New Jersey, employment expanded at the low-wage stores . . . and contracted at the high-wage stores . . . . Indeed, the average change in employment at the high-wage stores . . . is almost identical to the change among Pennsylvania stores . . . ." In other words, restaurants placed under no new legal requirement by the minimum wage saw similar employment changes, whether they were Pennsylvania restaurants or high wage New Jersey restaurants.

In short, the central concern in an observational study is that treatments may be assigned to subjects in a biased manner. The choice of a setting in which essentially the same treatment is assigned to subjects by several very different processes provides a partial check of this central concern. More precisely, the theory that the treatment is the cause of its ostensible effects predicts similar effects no matter which mechanism delivered the treatment, and this prediction offers an opportunity to refute or corroborate the theory.

3.6 Nondose Nonresponse: An Absent Association

A causal theory not only predicts the presence of certain associations, but also the absence of certain others. In some studies, it is possible to determine the dose of treatment that a control would have received had this control received the treatment. Call this the potential dose. While it is often reasonable to expect higher responses from treated subjects who received higher doses, the same pattern is not expected among controls—higher potential doses that were never received should not predict higher responses if higher responses are being caused by the treatment.

The minimum wage study contains an example. Card and Krueger (1994) define a measure GAP_i of the impact of the minimum wage legislation on restaurant i, as the percentage increase in the starting wage in restaurant i needed to achieve the new New Jersey minimum wage. A restaurant that already paid more that the new minimum would have $GAP_i = 0$. A restaurant that paid the old minimum wage would have $GAP_i = 18.8\%$. This variable GAP_i resembles a dose of treatment in that the law has greater impact on the starting wage when GAP_i is larger. A concern, however, is that GAP_i is not only a dose of treatment, but also a variable describing local labor market conditions. Presumably, some restaurants must pay higher starting wages than others to attract employees, so GAP_i is confounded with labor market conditions. For instance, the labor market in the poor city of Camden is different from the labor market in the relatively affluent suburb of Princeton. Still, it is not unreasonable to think that this confounding operates in a similar way in New Jersey and Pennsylvania. For the control restaurants in Pennsylvania, the undelivered potential dose of treatment is the percentage change in the starting wage needed to achieve New Jersey's new minimum wage. If these undelivered potential doses predicted employment changes in Pennsylvania, then that could not be an effect of New Jersey's minimum wage legislation and would strongly suggest confounding. Card and Krueger (1994, page 784) write:

> [W]e exclude stores in New Jersey and (incorrectly) define the variable for Pennsylvania stores as the proportional

> increase in wages necessary to raise the wage to \$5.05 per hour. In principle the size of the wage gap for stores in Pennsylvania should have no systematic relation with employment growth. In practice, this is the case. There is no indication that the wage gap is spuriously related to employment growth.

In short, there is often the concern that not only the assignment to treatment or control but also the dose of treatment is confounded with unobserved covariates. When potential but undelivered doses are known for controls, the theory that the treatment is the cause of its ostensible effects predicts that delivered doses in the treated group should be related to the magnitudes of responses, but undelivered doses in the control group should be unrelated to responses. This prediction provides an additional check of the theory that the treatment is the cause of its ostensible effects.

3.7 Stability Analyses, and Minimizing the Need for Stability Analyses

A complex analysis involves numerous implementation or analytical decisions. The audience for such an analysis typically wishes to be assured that conclusions are not artifacts of such decisions, but rather are stable over analyses that differ in apparently innocuous ways. All three examples in Section 2 include stability analyses for certain descisions, as described below.

A sensitivity analysis asks to what extent plausible changes in assumptions change conclusions. In contrast, a stability analysis asks how ostensibly innocuous changes in analytical decisions change conclusions. A sensitivity analysis typically examines a continuous family of departures from a critical assumption, in which the magnitude of the departure and the magnitude of the change in conclusions are the focus of attention. In contrast, a stability analysis typically examines a discrete decision, and the hope and expectation is that the conclusions are largely unaltered by changing this decision. Stability analyses are necessary in most complex analyses; however, the extent to which they are needed varies markedly from study to study.

As an example of a stability analysis, consider the MSSORS study of lead exposures, which needed to address the possibility that workers exposed to lead were more likely to have lead-related hobbies. Morton et al. (1982, pages 552–553) write:

> [Eleven] pairs were found in which the study children had potential for lead exposure other than that due to the father's occupation, while their matched controls had no such exposures. Exogenous sources of lead found in the children's environment were automobile body painting, casting of lead and playing with spent gun shell casings in the home. Six children in the study group had fathers whose hobby was casting lead into fish sinkers; none of the control children's fathers did this. It was speculated that those who work with lead on the job are more accustomed to handling lead, thereby promoting its use in the home environment.... [A]ny study/control matched pair in which one of these hobbies was present for the study child and not present for the control was eliminated from the analysis.... When these 11 children and their controls were eliminated, and the remaining 22 pairs were analyzed, the study and control groups continued to be statistically different ($p < .001$).

Morton et al. go on to show that the estimates of lead levels do not change much when the 11 pairs are excluded. Notice that a discrete decision—whether or not to include the 11 pairs—is investigated by carrying out the analysis both ways and comparing the results, a process that is very different from sensitivity analysis.

Similarly, in the LWW study of the effects of the death of a spouse or child in a car crash, there was concern that the death of a spouse might alter family income and, possibly, that it was this sustained loss of income and not the death itself that has psychological effects. This was investigated by adjusting for income by regression, concluding that most of the findings remained stable (LWW, 1987, page 226). The minimum wage study included several stability analyses, for instance, various ways of handling temporarily closed restaurants; see CK (1994, Table 5).

The examples mentioned above have a common feature that arises frequently in observational studies. Because it is clear that one wishes to compare treated and control subjects who were comparable prior to treatment, it is clear that visible pretreatment differences need to be removed by adjustments of one kind or another. It may happen, however, that treated and control groups are noted to differ after the start of treatment. In this case, the posttreatment difference may reflect unobserved pretreatment differences or effects caused by the treatment or the effects of other treatments occurring at the same time. Lead hobbies are another treatment coexisting with occupational lead exposure. Income loss is an outcome affected by the loss of a spouse. The closing of restaurants may be related to business conditions, the level of the minimum wage being one such condition. Adjustments for posttreatment differences may remove part of the actual treatment effect, and they may either remove bias or introduce bias into comparisons (Rosenbaum, 1984). Stability analyses are common in this setting, and they seek to demonstrate that results are stable whether or not adjustments are made for a posttreatment difference. An alternative approach is to be explicit about the effect of the treatment on the posttreatment variable, replacing the discrete choice of a stability analysis by the continuous variation in assumptions of a sensitivity analysis; see Rosenbaum (1984, Sections 4.3, 4.4) for detailed discussion of this alternative.

As another example with a different result, the minimum wage debate between Neumark and Wascher (1992) and Card, Katz and Krueger turns in part on a stability analysis. Neumark and Wascher had conducted a panel study of changes in state minimum wages between 1973 and 1989 in relation to unemployment among teenagers and young adults, reaching the conclusion that increases in minimum wages depress employment. Card, Katz and Krueger then commented that the results were unstable in the following sense. Neumark and Wascher had made adjustments for the "proportion of the age group enrolled in school." Card, Katz and Krueger argued that, first, the conclusions about the minimum wage change dramatically if adjustments are not made for this variable and, second, that the definition of this variable is such that it is "mechanically" related to the response variable, namely employment, because anyone working even part time was counted as not enrolled in school. In their response, Neumark and Wascher disagreed.

The purpose here is to examine methodology and not the minimum wage debate. From a methodological view, the study by Neumark and Wascher (1992) relies heavily on analytical models in comparisons, whereas the study by Card and Krueger (1994) relies more on the design of the study and the choice of circumstances in which fairly comparable units, here the restaurants, are being compared. A study which relies heavily on analytical models and adjustments will make many more implementation decisions in analysis, and so will need to conduct more extensive stability analyses to be convincing. This is an argument in favor of simple study designs that compare ostensibly comparable units under alternative treatments.

3.8 The Role of Time: Abrupt, Short-Lived Treatments

The concept that a treatment must precede its effects influences the study of treatments in many ways, with similar issues arising in experiments and observational studies. A treatment may be delivered over a prolonged period of time, so the distinction between what precedes a treatment and what follows it may not be sharp. Subjects may switch from one treatment to another, possibly in part in response to an earlier failure of the first treatment. Other extraneous treatments may intervene, and these interventions may themselves be, in part, effects stimulated by the treatments under study. Subjects may know about the treatment before they receive it, and part of the ultimate effect of the treatment may begin to materialize before the treatment is delivered. For instance, the increases in New Jersey's minimum wage were public legislation well before the increases actually took place. For discussion of various aspects of the role of time in intervention studies, see Campbell and Stanley (1963, pages 5–6), Diggle, Liang and Zeger (1994), Holland (1993), Joffe et al. (1997), Peto et al. (1976), Robins (1989a, 1992, 1998), Robins, Rotnitzky and Zhao (1995), Rosenbaum (1984), Rubin (1991, Section 5.2), Shafer (1996), Sobel (1995) and Susser (1987, page 86).

In research design, given the choice, one would prefer a single, abrupt, unexpected, short-lived treatment of dramatic proportions. With such a treatment, the line between what precedes treatment and what follows it is sharply drawn. This is true of only one of the three studies in Section 2, namely, the LWW study of the psychological effects of the death of a spouse or a child in a car crash.

Abrupt, unexpected, short-lived treatments of dramatic proportions are sometimes informally associated with the term "exogenous" as it is used in econometrics. However, formal discussions of "exogeneity" actually define the matter rather differently (e.g., Engle, Hendry and Richard, 1983).

Another example of an abrupt, short-lived treatment is found in a study of the effects of immigration on labor markets. This is generally a difficult topic because immigration occurs gradually and immigrants may favor labor markets where jobs are available and attractive. It is usually difficult to disentangle the effects of immigration on labor markets from the effects of labor markets on immigration. Card (1990) exploited a rare exception, in which immigration was abrupt, short-lived, unexpected and of dramatic proportions. In the Mariel boatlift, about 125,000 Cubans immigrated to Miami between May and September 1980, increasing Miami's labor force by 7%. Card compared changes in Miami's labor market following the Mariel boatlift to the concurrent changes in four unaffected cities, Atlanta, Houston, Los Angeles and Tampa–St. Petersberg.

3.9 Natural Blocks

A final opportunity to use choice in place of control in the design of an observational study involves natural blocks. Whereas matching is used to pair unrelated individuals having similar values of measured covariates, natural blocks create pairs or groups of individuals who are related in ways judged to be important but difficult to measure explicitly. Twins, siblings, neighbors and schools are familiar examples of natural blocks.

Twins, for example, resemble one another in terms of genetics and childhood environment in many ways that cannot practically be described in measured covariates. The LWW study of the psychological effects of the loss of a spouse or of a child could not adjust for genetic differences between exposed subjects and controls. In related work, Lichtenstein et al. (1996) examined the psychological effects of widowhood by comparing twins, one bereaved and one still married. In partial corroboration of the LWW study, they also found long-term psychological effects of the loss of a spouse, suggesting that genetic differences are not a likely explanation of the psychological outcomes. Behrman, Rosenzweig and Taubman (1996) use twins in an economic observational study of the effect of college quality on subsequent earnings. Ashenfelter and Krueger (1994) make a similar use of twins, while Altonji and Dunn (1996) use siblings instead.

It is possible to combine matching for covariates and pairing using natural blocks. In the MSSORS study of lead exposure, exposed and control children were compared to a neighbor's child of about the same age. Here, age is a covariate whereas neighborhood is a block. Similarly, in Rosenbaum (1986), high school dropouts were matched to students with similar grades, test scores and behavior who remained in the same high school. Here, the high school is the block. In both cases, matching controlled for blocks and began the adjustment for covariates.

4. A FORMAL RESTATEMENT OF SELECTED ISSUES

4.1 Formal Restatement: Outline

In Section 4, a few of the ideas of Section 3 are restated in slightly more formal terms. The goals of this restatement are two: (i) to discuss, in simple or special cases, certain ideas in more precise terms; (ii) to link the ideas of Section 3 more closely with the statistical literature on observational studies. After briefly reviewing a notation for causal effects in Section 4.2 and the role of adjustments and sensitivity analyses in Section 4.3, a simple formal illustration of the role of broad theories, from Section 3.1, is given in Section 4.4. Then Section 4.5 restates the issues in Section 3.3 about the choice of treated and control groups in terms of reducing sensitivity to hidden bias.

4.2 Review of Notation for Causal Effects

This section briefly reviews for later use a notation for causal effects that was introduced by Neyman (1923) in the context of experiments and developed by Rubin (1974, 1977) for use in observational studies. Suppose there is a population Ω of units, and a collection T of treatments, where one or more of the treatments $t \in T$ may be "control" treatments. Each unit $\omega \in \Omega$ may be subjected to any one treatment $t \in T$. If unit $\omega \in \Omega$ were to receive treatment $t \in T$ then this unit would exhibit the vector response $\mathbf{r}_{t\omega}$, and the effect on unit ω of applying, say, treatment 1 rather than treatment 0 is a comparison of $\mathbf{r}_{1\omega}$ and $\mathbf{r}_{0\omega}$ such as $\mathbf{r}_{1\omega} - \mathbf{r}_{0\omega}$.

This notation implicitly embodies an important assumption, called "no interference between units" by Cox (1958, Section 2.4), namely, that "the observation on one unit should be unaffected by the particular assignment of treatments to other units." Rubin refers to this assumption as SUTVA, or the stable-unit-treatment-value-assumption. Instead of assuming that units do not interfere with one another, one may define the "unit of analysis" as that set of units Ω such that units do not interfere with one another. For instance, in an educational intervention, the treatment applied to one student in a class might affect other students in the same class, but not affect students in other classes, so if Ω is the set of classes, not the set of students, and each $t \in T$ specifies treatments for all students in the class, then there would be no interference between units; see the discussion of "unit of analysis" in Meyer and Fienberg (1992, pages 81–84). Robins (1992) and Joffe et al. (1997) discuss important extensions of this notation to the case of different treatments applied at different times to the same subjects, so that interference is inevitable.

If one knew $E = \{\mathbf{r}_{t\omega}, t \in T, \omega \in \Omega\}$, then calculating treatment effects would require only arithmetic. Of course, one does not know E. One observes a response only for some subjects $\omega \in \Omega$ and then only the response $\mathbf{r}_{t\omega}$ for the treatment t that ω actually received. If subject $\omega \in \Omega$ is observed and receives treatment t, write $Z_\omega = t$, and then $\mathbf{r}_{t\omega}$ is observed, but $\mathbf{r}_{t'\omega}$ is not observed for $t' \neq t$, so $\mathbf{r}_{t\omega} - \mathbf{r}_{t'\omega}$ cannot be calculated. In a randomized clinical trial, a subset—not typically a sample—of Ω is entered in the trial, and for this subset Z_ω is determined by a random number generator, yielding randomization inferences about treatment effects and unbiased estimates of average treatment effects over the subset of units who are in the trial, though not necessarily unbiased estimates over all of Ω; see Fisher (1935), Kempthorne (1952, Sections 7, 8) and Lehmann (1975, page 5). An observational study is similar except that treatment assignment is not controlled by the investigator and hence is not determined by a random number generator, so randomization inferences and unbiased estimates are no longer available.

Unlike an outcome $\mathbf{r}_{t\omega}$, a covariate is, by definition, a variable measured prior to treatment, so it is unaffected by the treatment, and a covariate exists in a single version not varying with t. Write $\mathbf{x}_\omega$ for an observed covariate describing unit $\omega \in \Omega$, and write u_ω for an unobserved covariate.

4.3 Overt and Hidden Biases: Adjustments and Sensitivity Analysis

Randomization permits valid comparisons ignoring both observed and unobserved covariates, which tend to be balanced across different treatment groups, although adjustments are sometimes made for observed covariates to increase the precision of estimated treatment effects; see Fisher (1935) for the theory of randomization inference. In contrast, in observational studies, it is often found that treatment groups differ markedly with respect to observed covariates, and there is often concern that the groups may also differ with respect to unobserved covariates. Differences in observed covariates are controlled by adjustments, for example, by matching subjects ω, $\omega \in \Omega$ with the same value of observed covariates, $\mathbf{x}_\omega = \mathbf{x}_{\omega'}$, receiving different treatments, $Z_\omega = t$, $Z_{\omega'} = t'$; for example, Smith (1997). As seen from Theorem 4 of Rosenbaum and Rubin (1983a), matching and other similar methods of adjustment permit estimation of average treatment effects when treatment assignment is strongly ignorable given the observed covariates $\mathbf{x}_\omega$ in the sense that

$$\begin{aligned} &Z_\omega \perp\!\!\!\perp \{\mathbf{r}_{t\omega}, t \in T\} \mid \mathbf{x}_\omega \\ &\quad 0 < \operatorname{prob}(Z_\omega = t \mid \mathbf{x}_\omega) < 1, \quad t \in T, \end{aligned}$$

where $A \perp\!\!\!\perp B \mid C$ is Dawid's (1979) notation for A is conditionally independent of B given C. Ignorability given $\mathbf{x}_\omega$ implies that treatment assignments are unrelated to potential outcomes within strata defined by $\mathbf{x}_\omega$.

Often, there is concern that units which appear similar in terms of observed covariates $\mathbf{x}$ may differ in terms of a relevant unobserved covariate u, so that treatment assignment is not ignorable given $\mathbf{x}$ alone. A model that expresses this possibility asserts that treatment assignment is ignorable given $(\mathbf{x}, u)$ and that, for two subjects ω, $\omega' \in \Omega$ with the same observed covariates $\mathbf{x}_\omega = \mathbf{x}_{\omega'}$, the odds of receiving one treatment rather than another differ by at most a factor of $\Gamma \geq 1$,

$$\begin{aligned} \Gamma &\geq \frac{\operatorname{prob}(Z_\omega = t \mid \mathbf{x}_\omega, u_\omega) \cdot \operatorname{prob}(Z_{\omega'} = t' \mid \mathbf{x}_{\omega'}, u_{\omega'})}{\operatorname{prob}(Z_\omega = t' \mid \mathbf{x}_\omega, u_\omega) \cdot \operatorname{prob}(Z_{\omega'} = t \mid \mathbf{x}_{\omega'}, u_{\omega'})} \\ &\geq \frac{1}{\Gamma}. \end{aligned}$$

For each fixed $\Gamma \geq 1$, bounds on permutation significance levels, confidence intervals and point estimates are available, and, by calculating these bounds for a range of values of Γ, one obtains a sensitivity analysis which displays how hidden biases of various magnitudes might alter the conclusions of the study (Rosenbaum, 1995a, Section 4).

Sensitivity analyses for the lead exposure example are given in Rosenbaum (1993, 1995a, 1999a), and for the minimum wage example, with and without an instrumental variable, in Rosenbaum (1999a). The lead exposure example is much less sensitive to hidden bias than the minimum wage example; see Rosenbaum (1999a). Specifically, in the lead exposure example, an unobserved covariate u would need to be a near perfect predictor of children's lead levels and associated with at least a $\Gamma = 4$-fold increase in the odds of exposure to lead to explain the elevated levels of lead in the blood of children of lead workers. In contrast, in the minimum wage example, a much smaller hidden bias of $\Gamma = 1.5$ would suffice to render plausible a 12% decline in employment or a 27% increase in employment caused by increasing the minimum wage. In short, much smaller hidden biases could alter the conclusions of the minimum wage example compared to the lead exposure example. For extensive details, see Rosenbaum (1999a).

4.4 Empirical Content of Theories: An Explicit Example

This section elaborates the point in Section 3.1 in a trivial but precise illustration. The issue concerned a preference for theories which are, in Popper's phrase, "better testable" because their "empirical content is greater." The illustration also concerns the issue raised in the quote from Campbell in Section 1.1. Write $r_{t1\omega}$ for the first coordinate of $\mathbf{r}_{t\omega}$. Suppose that there are three treatments, 0, 1 and 2, and consider two possible claims about the effects of these treatments:

Claim 1.

$$\frac{1}{|\Omega|}\sum_{\omega\in\Omega} r_{01\omega} < \frac{1}{|\Omega|}\sum_{\omega\in\Omega} r_{11\omega} < \frac{1}{|\Omega|}\sum_{\omega\in\Omega} r_{21\omega},$$

where $|\Omega|$ is the number of units in the population Ω

Claim 2. $r_{01\omega} < r_{11\omega} < r_{21\omega}$, $\forall\, \omega \in \Omega$.

Here, Claim 1 asserts that the average over the entire population Ω of the first response is lowest under treatment zero and highest under treatment 2, whereas Claim 2 asserts that every single unit in Ω has lowest response under treatment 0 and highest response under treatment 2. For instance, Claims 1 and 2 would both be true if the treatments had constant, additive effects, with the smallest constant for treatment $t = 0$ and the largest for treatment $t = 2$. Note carefully that neither Claim 1 nor Claim 2 can be determined to be true or false by direct inspection of units, because each unit is observed under only one treatment; that is, $r_{t1\omega}$ is observed from unit $\omega \in \Omega$ only if $Z_\omega = t$. To determine by inspection whether Claim 1 or 2 is true would require observing $r_{01\omega}$, $r_{11\omega}$ and $r_{21\omega}$, simultaneously from the same unit ω, and this is not possible; hence, one can draw inferences relevant to Claims 1 and 2 but one cannot determine whether either is true by inspection of individual units.

Now Claim 2 implies Claim 1, but not conversely. This has two consequences: first, Claim 1 is at least as plausible as Claim 2, but, second, Claim 2 is "better testable" and has greater "empirical content" in the sense of the quote from Popper in Section 3.1.

Strong evidence against Claim 1 might be obtained by drawing a sufficiently large random sample of units $\omega \in \Omega$ and conducting a randomized experiment for these units. In contrast, strong evidence against Claim 2, but not against Claim 1, might be obtained by selecting any sufficiently large group of units $\omega \in \Omega$, not necessarily a sample, and conducting a randomized experiment for these units. If the treatments are three new drugs and the relevant population Ω is the set of all patients who might eventually be treated with these drugs, then Claim 2 might be tested in a randomized clinical trial at a single medical center using patients who volunteered for the trial by providing informed consent—something that happens every day—but Claim 1 could not be tested in any practical sense. Because Claim 2 says more than Claim 1, because it has greater empirical content, Claim 2 is at greater risk of refutation and can be subjected to greater scrutiny.

Because Claim 2 says more than Claim 1, if Claim 2 is false, we are more likely to discover that it is false. Discovering that Claim 2 is false is progress. For instance, in a randomized experiment, we might refute Claim 2 by learning that treatment 0 is better for an identifiable group of patients; that is, we might find that for units ω with certain values of the observed covariates $\mathbf{x}_\omega$ higher responses are typically observed from treatment group 0 than from treatment group 2. (Obviously, such a "refutation" would need to make due allowance for both chance and any consequences of multiple comparisons.)

Because Claim 2 says more than Claim 1, if Claim 2 is true, then corroborating evidence supporting Claim 2 is easier to obtain. That is, Claim 2 implies more than Claim 1, so there are more ways to refute Claim 2, and each serious but unsuccessful attempt to refute Claim 2 provides corroboration in the sense discussed by Popper.

A theorem of statistics deduces conclusions from assumptions. Quite often, however, assumptions which are used in similar ways in a proof, nonetheless, have very different scientific status. Assumptions are of three kinds: (i) the scientific model or hypothesis, which is the focus of scientific interest, (ii) incidental assumptions, needed for statistical inference but of little or no scientific interest and (iii) pivotal assumptions which rival the scientific hypothesis and are the focus of scientific controversy. Claims 1 and 2 are both scientific hypotheses, and, as such, greater simplicity and greater empirical content are desirable. A better hypothesis says more, and in consequence is less likely to be true, is more likely to be discovered to be false and is more easily and strongly corroborated when efforts to refute it are unsuccessful. An incidental assumption is an unnecessary convenience, for instance, the assumption of Normal errors in a completely randomized experiment. The assumption is unneeded, since a nonparametric procedure may be used in a completely randomized experiment. For incidental assumptions, it is preferable to assume less or to use methods that are robust to violations of incidental assumptions. An incidental assumption, unlike a scientific hypothesis, is not of scientific interest, so one would prefer that incidental assumptions had little impact on the final conclusion, and one would prefer incidental assumptions that intruded as little as possible on the scientific questions of interest. A better hypothesis says more, but a better incidental assumption says less. A pivotal assumption is a rival to the scientific hypothesis, and the dependence of the inference on a pivotal assumption cannot be removed by employing a different statistical method, for instance, by switching from the t-test to the Wilcoxon rank sum test. For instance, in an observational study, the assumption that, after adjustment for observed covariates, the different treatment groups are comparable—that is, the assumption that treatment assignment is strongly ignorable given x—is a pivotal assumption. Pivotal assumptions and scientific hypotheses are rivals: if a pivotal assumption is false, our inference about the scientific hypothesis using this assumption may be wrong, and refinements of analytical technique cannot resolve this rivalry. Pivotal assumptions are often a focus of attention in the discussion sections of scientific papers, a focus of concern for referees and a topic of discussion in subsequent letters to the editor. Pivotal assumptions are addressed, in part, through sensitivity analyses and through separate efforts to corroborate or refute these assumptions (Rosenbaum, 1995a, Section 4–8). Pivotal assumptions are not synonymous with identifying assumptions, for reasons outlined in Rosenbaum (1997a).

To recap, this section has discussed a formal illustration of Popper's general principle that broader, simpler, scientific models and hypotheses are preferred because they have greater empirical content and therefore are more easily refuted or corroborated. This statement has been distinguished from the preference for weaker, more innocuous incidental assumptions, and both preferences have been distinguished from issues surrounding pivotal assumptions that rival the scientific model.

4.5 Defining Treated and Control Groups to Reduce Sensitivity to Hidden Biases

In Section 3.3, it was suggested that treated and control groups should be chosen in such a way that they are sharply distinct and yet could happen to anyone. This may be expressed in the formalism of the sensitivity analysis mentioned in Section 4.3. Specifically, a study is insensitive to bias when two conditions hold simultaneously: the treatment effect is large, both practically and relative to the variability in outcomes; and the magnitude of plausible hidden biases, as measured by Γ in Section 4.3, is not large.

In the psychology example, the restriction of attention to a treated group that had lost either a spouse or a young child made the treated and control conditions more sharply distinct than would have been the case had losses of more distant relatives been included. This, in turn, probably produced a larger, and hence less sensitive, difference in outcomes in treated and control groups. Again, in the psychology example, the restriction of attention to drivers who were not at fault reduced somewhat the magnitude of plausible hidden biases, since, for example, alcohol abuse is related to accidents in which one is at fault and also related to psychiatric outcomes. In other words, this restriction reduced somewhat the range of plausible values of Γ.

Decisions made during the design of an observational study, such as the definition of treated and control groups, affect the sensitivity of the conclusions to hidden biases, but decisions made during analysis can have this effect as well. For instance, the sensitivity to hidden bias may be different at different quantiles of the control responses when the treatment has a nonconstant, dilated effect, and this may be clarified during the analysis; see Rosenbaum (1999b) for discussion and illustration. Similarly, a test of no treatment effect against the alternative of a coherent treatment effect may be less sensitive to hidden bias than a test against a less focused alternative (Rosenbaum, 1997b).

In short, in seeking "sharply distinct treatments that could happen to anyone" one is seeking a study that is insensitive to hidden biases of plausible size.

5. SUMMARY

In many well-conducted observational studies, clarity about treatment effects is enhanced by the choice of circumstances in which the study is conducted. Several principles guiding choice have been suggested. The use of choice in place of control has been illustrated with examples from economics, clinical psychology and epidemiology. Because these principles apply in diverse disciplines, the principles are properly viewed as part of the subject of statistics.

ACKNOWLEDGMENTS

This work was supported by a grant from the Methodology, Measurement and Statistics Program and the Statistics and Probability Program of the NSF and by the University of Pennsylvania Research Foundation. The author thanks Leon Gleser, an Associate Editor and referees for helpful comments.

REFERENCES

ALTONJI, J. G. and DUNN, T. A. (1996). Using siblings to estimate the effect of school quality on wages. *Review of Economics and Statistics* **77** 665–671.

ANGRIST, J. D. (1997). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Working Paper 5888, National Bureau of Economic Research, Cambridge, MA.

ANGRIST, J. D., IMBENS, G. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Amer. Statist. Assoc.* **91** 444–469.

ANGRIST, J. D. and KRUEGER, A. B. (1998). Empirical strategies in labor economics. *Handbook of Labor Economics.* (Working paper 401, Industrial Relations Section, Princeton Univ.) To appear.

ASHENFELTER, O. A. and KRUEGER, A. B. (1994). Estimates of the economic return to schooling from a new sample of twins. *American Economic Review* **84** 1157–1173.

BEHRMAN, J., ROSENZWEIG, M. and TAUBMAN, P. (1996). College choice and wages: estimates using data on female twins. *The Review of Economics and Statistics* **78** 672–685.

BORUCH, R. (1997). *Randomized Experiments for Planning and Evaluation*. Sage, Thousand Oaks, CA.

BOX, G. E. P. (1966). The use and abuse of regression. *Technometrics* **8** 625–629.

BRONARS, S. G. and GROGGER, J. (1994). The economic consequences of unwed motherhood: using twin births as a natural experiment. *American Economic Review* **84** 1141–1156.

CAMPBELL, D. T. (1988). Can we be scientific in applied social science? In *Methodology and Epistemology for Social Science: Selected Papers* [Originally published in *Evaluation Studies Review Annual* **9** (1984) 26–48.] 315–333. Univ. Chicago Press.

CAMPBELL, D. and STANLEY, R. (1963). *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, Chicago.

CARD, D. (1990). The impact of the Mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review* **43** 245–257.

CARD, D. and KRUEGER, A. (1994). Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* **84** 772–793.

CARD, D. and KRUEGER, A. (1995). *Myth and Measurement: The New Economics of the Minimum Wage*. Princeton Univ. Press.

COPAS, J. B. and LI, H. G. (1997). Inference for non-random samples (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 55–96.

COX, D. R. (1958). *The Planning of Experiments*. Wiley, New York.

CORNFIELD, J., HAENSZEL, W., HAMMOND, E., LILIENFELD, A., SHIMKIN, M. and WYNDER, E. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* **22** 173–203.

DAWES, R. (1996). The purpose of experiments: ecological validity versus comparing hypotheses. *Behavioral and Brain Sciences* **19** 20.

DAWID, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 1–31.

DEERE, D., MURPHY, K. and WELCH, F. (1995). Employment and the 1990–1991 minimum-wage hike. *American Economic Review* **85** 232–237.

DIGGLE, P. J., LIANG, K. Y. and ZEGER, S. L. (1994). *Analysis of Longitudinal Data.* Oxford Univ. Press.

DUFF, C. (1996). New minimum wage makes few waves: employers offset 50-cent raise with minor shifts. *Wall Street Journal* **20** November 1996 2–4.

ENGLE, R., HENDRY, D. and RICHARD, J. (1983). Exogeneity. *Econometrica* **51** 277–304.

FEYERABEND, P. (1968). How to be a good empiricist—a plea for tolerance in matters epistemological. In *The Philosophy of Science* (P. H. Nidditch, ed.) 12–39. Oxford Univ. Press.

FEYERABEND, P. (1975). *Against Method*. Verso, London.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

FREEDMAN, D. (1997). From association to causation via regression. *Adv. in Appl. Math.* **18** 59–110.

FRIEDMAN, M. (1953). The methodology of positive economics. In *Essays in Positive Economics* 3–43. Univ. Chicago Press.

GASTWIRTH, J. L. (1992). Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics* **33** 19–34.

GASTWIRTH, J. L., KRIEGER, A. M. and ROSENBAUM, P. R. (1998). Cornfield's inequality. In *Encyclopedia of Biostatistics* 952–955. Wiley, New York.

GREENHOUSE, S. (1982). Jerome Cornfield's contributions to epidemiology. *Biometrics* (*Suppl.*) **38** 33–45.

HEDGES, L. and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, New York.

HOLLAND, P. (1993). Which comes first, cause or effect? In *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues* (G. Keren and C. Lewis, eds.) 273–282. Erlbaum, Hillsdale, NJ.

HOLLAND, P. and RUBIN, D. (1983). On Lord's paradox. In *Principles of Psychological Measurement: A Festschrift for Frederic Lord* (H. Wainer and S. Messick, eds.) 3–25. Erlbaum, Hillsdale, NJ.

JOFFE, M., HOOVER, D., JACOBSON, L., KINGSLEY, L., CHMIEL, J., VISSCHER, B. and ROBINS, J. (1997). Estimating the effect of zidovudine on Kaposi's sarcoma from observational data using a rank preserving structural failure-time model. *Statistics in Medicine*. To appear.

KEMPTHORNE, O. (1952). *Design and Analysis of Experiments*. Wiley, New York. [Reprinted (1973)] by Krieger, (Malabar, FL).

LAKATOS, I. (1970). Falsification and the methodology of scientific research programs. In *Criticism and the Growth of Knowledge* (I. Lakatos and A. Musgrave, eds.) 91–196. Cambridge Univ. Press. [Reprinted in I. Lakatos, *Philosophical Papers* **1** 8–101. Cambridge Univ. Press 1978.]

LAKATOS, I. (1981). History of science and its rational reconstructions. In *Scientific Revolutions* (I. Hacking, ed.) 107–127. Oxford Univ. Press. [Reprinted from *Boston Studies in the Philosophy of Science* **8** (1970).

LEHMAN, D., WORTMAN, C. and WILLIAMS, A. (1987). Long-term effects of losing a spouse or a child in a motor vehicle crash. *Journal of Personality and Social Psychology* **52** 218–231.

LEHMANN, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

LICHTENSTEIN, P., GATZ, M., PEDERSEN, N., BERG, S. and MCCLEARN, G. (1996). A co-twin-control study of response to widowhood. *Journal of Gerontology: Psychological Sciences* **51B** 279–289.

MANSKI, C. (1990). Nonparametric bounds on treatment effects. *American Economic Review* **80** 319–323.

MANSKI, C. (1995). *Identification Problems in the Social Sciences*. Harvard Univ. Press.

MARCUS, S. (1997). Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *Journal of Educational and Behavioral Statistics* **22** 193–202.

MEEHL, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology* **46** 806–834. [Reprinted in P. Meehl, *Selected Philosophical and Methodological Papers* 1–42. Univ. Minnesota Press (1991).]

MEYER, B. D. (1995). Natural and quasi-experiments in economics. *J. Bus. Econom. Statist.* **13** 151–161.

MEYER, M. and FIENBERG, S. eds. (1992). *Assessing Evaluation Studies: The Case of Bilingual Education Strategies*. National Academy Press, Washington, DC.

MORTON, D., SAAH, A., SILBERG, S., OWENS, W., ROBERTS, M. and SAAH, M. (1982). Lead absorption in children of employees in a lead-related industry. *American Journal of Epidemiology* **115** 549–555.

NEUMARK, D. and WASCHER, W. (1992). Employment effects of minimum and subminimum wages: lanel data on state minimum wage laws. *Industrial and Labor Relations Review* **46** 55–81. [See also (1993) **47**, 487–512 for discussion by Card, Katz and Krueger and a reply by Neumark and Wascher.]

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Roczniki Nauk Roiniczych* **10** 1–51. (In Polish.) [Reprinted in English (1990) *Statist. Sci.* **5** 463–480, with discussion by T. Speed and D. Rubin.]

PEIRCE, C. S. (1903). On selecting hypotheses. [Reprinted (1960) *Collected Papers of Charles Sanders Peirce* (C. Hartshorne and P. Weiss, eds.) **5** 413–422. Harvard Univ. Press.]

PETO, R., PIKE, M., ARMITAGE, P., BRESLOW, N., COX, D., HOWARD, S., MANTEL, N., MCPHERSON, K., PETO, J. and SMITH, P. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer* **34** 585–612.

PLATT, J. (1964). Strong inference. *Science* **146** 347–352.

POLANYI, M. (1964). *Science, Faith and Society*. Univ. Chicago Press. [Originally published (1946) by Oxford Univ. Press.]

POPPER, K. (1968). *The Logic of Scientific Discovery*. Harper and Row, New York. (English translation of Popper's 1934 *Logik der Forschung*.)

POPPER, K. (1965). *Conjectures and Refutations*. Harper and Row, New York.

PUTNAM, H. (1995). *Pragmatism*. Blackwell, Oxford, UK.

QUINE, W. (1951). Two dogmas of empiricism. *Philosophical Review*. [Reprinted in W. Quine, *From a Logical Point of View* 20–46. Harvard Univ. Press (1980).]

ROBINS, J. (1989). The control of confounding by intermediate variables. *Statistics in Medicine* **8** 679–701.

ROBINS, J. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* **79** 321–334.

ROBINS, J., ROTNITZKY, A. and ZHAO, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90** 106–121.

ROBINS, J. (1998). Correction for non-compliance in equivalence trials. *Statistics in Medicine* **17** 269–302.

ROSENBAUM, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. Roy. Statist. Soc. Ser. A* **147** 656–666.

ROSENBAUM, P. R. (1986). Dropping out of high school in the United States: an observational study. *Journal of Educational Statistics* **11** 207–224.

ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26.

ROSENBAUM, P. R. (1991). Some poset statistics. *Ann. Statist.* **19** 1091–1097.

ROSENBAUM, P. R. (1993). Hodges–Lehmann point estimates of treatment effect in observational studies. *J. Amer. Statist. Assoc.* **88** 1250–1253.

ROSENBAUM, P. R. (1995a). *Observational Studies*. Springer, New York.

ROSENBAUM, P. R. (1997a). Discussion of a paper by Copas and Li. *J. Roy. Statist. Soc. Ser. B* **59** 90.

ROSENBAUM, P. R. (1997b). Signed rank statistics for coherent predictions. *Biometrics* **53** 556–566.

ROSENBAUM, P. R. (1999a). Using combined quantile averages in matched observational studies. *J. Roy. Statist. Soc. Ser. C* **48** 63–78.

ROSENBAUM, P. R. (1999b). Reduced sensitivity to hidden bias at upper quantiles in observational studies with dilated effects. *Biometrics*, **55** 560–564.

ROSENBAUM, P. R. and RUBIN, D. B. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

ROSENBAUM, P. and RUBIN, D. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Roy. Statist. Soc. Ser. B* **45** 212–218.

ROSENZWEIG, M. and WOLPIN, K. (1980). Testing the quantity–quality fertility model: the use of twins as a natural experiment. *Econometrica* **48** 227–240.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688–701.

RUBIN, D. B. (1977). Randomization on the basis of a covariate. *Journal of Educational Statistics* **2** 1–26.

RUBIN, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.* **6** 34–48.

RUBIN, D. B. (1991). Practical implications of modes of statistical inference for causal inference and the critical role of the assignment mechanism. *Biometrics* **47** 1213–1234.

SAMUELSON, P. (1947). *Foundations of Economic Analysis*. Harvard Univ. Press. [Reprinted (1983) by Harvard Univ. Press.]

SHAFER, G. (1996). *The Art of Causal Conjecture*. MIT Press.

SMITH, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology* **27** 325–353.

SOBEL, M. (1995). Causal inference in the social and behavioral sciences. In *Handbook of Statistical Modelling for the Social and Behavioral Sciences* (G. Arminger, C. Clogg and M. Sobel, eds.) 1–38. Plenum, New York.

SOWELL, T. (1995). Repealing the law of gravity. *Forbes* **22** May 1995, 82.

SUSSER, M. (1987). Falsification, verification and causal inference in epidemiology: reconsideration in the light of Sir Karl Popper's philosophy. In *Epidemiology, Health and Society: Selected Papers* (M. Susser, ed.) 82-93. New York: Oxford.

Comment

Charles F. Manski

In his lucid and engaging article, Rosenbaum develops a coherent perspective on observational studies. The logical beginning of his argument is his assertion that the objective of studies of treatment effects, whether experimental or observational, should be to assess the soundness of broad theories. The next step is his citation of a basic principle of scientific inference, that evidence on broad theories may be amassed by examining predictions in particular empirical settings. He then points to laboratory experiments as ideal for evaluation. He concludes that observational studies should seek to emulate such experiments. Rosenbaum presents his argument concisely and elegantly in Section 3.1:

> In a well-conducted laboratory experiment one of the rarest of things happens: the effects caused by treatments are seen with clarity.
>
> Observational studies of the effects of treatments on human populations lack this level of control but the goal is the same. Broad theories are examined in narrow, focused, controlled circumstances.

Although I respect Rosenbaum's perspective on observational studies, I do not share it. I approach the study of treatment effects with a different objective in mind, and consequently reach different conclusions for the conduct of observational studies. I am glad to have this opportunity to juxtapose Rosenbaum's world view and my own.

TREATMENT CHOICE IN HETEROGENEOUS POPULATIONS

I begin from the premise that studies of treatment effects should aim to improve the treatment choices made in a population of interest. In particular, I am concerned with treatment choice in heterogeneous populations, where treatment response may vary from person to person. I judge studies, whether experimental or observational, by their usefulness in improving treatment choices. This perspective has motivated my research on nonparametric bounds on treatment effects (Manski, 1990, 1994, 1995, 1997a, 1997b; Manski and Pepper, 2000) and has become explicit in my recent work connecting the empirical analysis of treatment response with the normative analysis of treatment choice (Manski, 1998, 1999).

I find it useful to suppose that a planner must choose a *treatment rule* which assigns a treatment to each member of a heterogeneous population of interest. The planner might, for example, be a physician choosing medical treatments for each member of a population of patients or a judge deciding sentences for each member of a population of convicted offenders. The planner observes certain covariates for each person: demographic attributes, medical or criminal records and so on. Each member of the population has a *response function* which maps treatments into a real- valued outcome of interest, perhaps a measure of health status or recidivism in the examples above.

I suppose that the planner wants to choose a treatment rule that maximizes the population mean outcome. (In economic terms, the planner wants to maximize a utilitarian social welfare function.) An optimal treatment rule assigns to each member of the population a treatment that maximizes mean outcome conditional on the person's observed covariates (Manski, 1998, 1999). This motivates interest in the empirical study of treatment effects. We want to infer treatment effects to help the planner select a good treatment rule.

THE SAMPLE DATA AND THE POPULATION OF INTEREST

My concern with the problem of treatment choice makes me part ways with Rosenbaum in the very first paragraph of his article. Here Rosenbaum downplays the importance of having sample data that is representative of the population of interest: "Studies of samples that are representative of populations may be quite useful in describing those populations, but may be ill suited to inferences about treatment effects." He cites with favor a statement by Campbell asserting that external validity has been overvalued in the literature on program evaluation. Rosenbaum accepts the idea, sometimes associated with Campbell, that studies

Charles F. Manski is Professor, Department of Economics and Institute for Policy Research, Northwestern University, 2003 Sheridan Road, Evanston, Illinois 60208-2600 (e-mail: cfmanski@nwu.edu).

of treatment effects should be judged primarily by their internal validity and only secondarily by their external validity.

A planner may find the Campbell–Rosenbaum position congenial if the planner has a strong prior belief that treatment response is homogeneous across the population of interest. If so, the planner may determine the best treatment in some easily studied subpopulation and then extrapolate to its complement. In human populations, however, homogeneity of treatment response seems the exception rather than the rule. Whether the context be medical or educational or social, the norm seems to be that people vary in their response to treatment, with the optimal treatment varying from person to person.

A planner choosing a treatment rule for a heterogeneous population cannot easily extrapolate findings on treatment effects from an easily studied subpopulation to its complement, as optimal treatments in the two may differ. Hence the planner should value sample data that is representative of the entire population of interest. Indeed, the planner may judge an observational study of the entire population to be more useful than an ideally executed experiment performed on a subpopulation.

PRIOR INFORMATION AND BOUNDS ON TREATMENT EFFECTS

I imagine that Rosenbaum, and some others, may react negatively to the assertion above. I shall therefore elaborate, using notation similar to Rosenbaum's.

Let T denote the feasible treatments. Let ω denote a person in the population Ω. Let $r_{t\omega}$ denote a real-valued outcome that ω would experience if this person were to receive treatment t. Let $x_\omega \in X$ denote covariates of ω that the planner can observe. Let $E(r_t \mid x)$ be the mean outcome under treatment t for persons with covariates x. Assume that the planner wants to maximize the mean outcome in Ω. It can be shown that an optimal treatment rule assigns to each person a treatment that maximizes the mean outcome conditional on the person's observed covariates. Thus, an optimal treatment for persons with covariates x maximizes $E(r_t \mid x)$ on $t \in T$.

A study of treatment effects is useful to the planner to the extent that it reveals the ordering of the conditional mean outcomes $[E(r_t \mid x), t \in T]$ for each $x \in X$. Consider an experiment performed on a subpopulation $X_0 \subset X$. The experiment identifies $[E(r_t \mid x), t \in T]$ for persons with covariates in X_0, but reveals nothing about the outcomes of persons with covariates in the complement subpopulation $X - X_0$. The usefulness of the experiment to the planner depends on the fraction of the population who have covariates in X_0, and on the credible prior information the planner can use to extrapolate from X_0 to its complement (Manski, 1996).

Now consider an observational study performed on the entire population. It is well known that an observational study can be used to identify treatment effects if sufficiently strong prior information is available. Unfortunately, the prior information needed to achieve identification is so strong that it is rarely credible in practice (Manski, 1995, Chapter 2). This being the case, my research program has sought to determine what may be learned about treatment response when observational studies are combined with weak prior information that may be deemed credible in practice. The findings generally take the form of bounds on the conditional mean outcomes $[E(r_t \mid x), t \in T]$, $x \in X$.

The starting point is the *worst-case analysis* of Manski (1990); see Manski, Sandefur, McLanahan and Powers (1992) for an empirical application. This shows that if the outcome variable r is itself bounded, then an observational study reveals informative bounds on $[E(r_t \mid x), t \in T]$ for $x \in X$, even in the absence of prior information. However, the bounds for different treatments necessarily intersect, implying that prior information is necessary if the planner is to rank treatments. Consequently, in Manski (1990, 1994, 1995, 1997a) and Manski and Pepper (2000), I have investigated the identifying power of various nonparametric restrictions on the distribution of treatment response or on the process of treatment selection in an observational study. Such prior restrictions enable the planner to tighten the worst-case bounds on $[E(r_t \mid x), t \in T]$ for $x \in X$.

Some forms of prior information may yield nonintersecting bounds on mean outcomes under different treatments. When this happens, the planner can use an observational study to rank treatments. In particular, the *instrumental variable* assumptions used by economists in observational studies for over 50 years have this important property. See Manski (1990, 1994), Manski and Pepper (2000) and related work on experiments with imperfect compliance by Robins (1989b), Robins and Greenland (1996) and Balke and Pearl (1997). See Manski and Nagin (1998) for an empirical application.

ROSENBAUM'S SENSITIVITY ANALYSIS

Another form of prior information that may yield nonintersecting bounds is proposed by Rosenbaum in his Section 4.3 discussion of *sensitivity analysis*. Here Rosenbaum, who views the assumption of ignorable treatment assignment as critical to the interpretation of observational studies, considers weakening this assumption in a particular manner. His approach has elements in common with research on robust statistics, which begins from some central model and examines how the possibilities for inference degrade as one moves away from that model in specified ways. To Rosenbaum, the central model is ignorable treatment selection conditional on x.

Where Rosenbaum and I differ is that I do not view the assumption of ignorable treatment selection to have a special status in observational studies of treatment effects. As an economist, I usually am inclined to think that treatments are purposefully selected and that comparison of outcomes plays an important role in the selection process. Perhaps the departures from ignorable treatment selection that Rosenbaum entertains in his sensitivity analysis can be interpreted behaviorally in terms of some model of purposeful treatment selection, but for now I do not see how.

Comment

James M. Robins

Rosenbaum provides a nice discussion of the role of design choice as an alternative to analytic control using three actual observational studies as examples. My discussion will cover three distinct areas. First, I will comment on particular points made by Rosenbaum. Second, I will complement Rosenbaum's discussion by presenting a thought experiment that illustrates that the choice of an appropriate statistical analysis depends as much on the design of the study and background subject-matter knowledge as on the data.

In the third part of my discussion, I will review special difficulties that arise in drawing causal inferences from randomized or observational data in the presence of time-varying or sequential treatments or exposures and show how these difficulties can impact on the choice of design. In particular, I will focus on testing for a direct effect of a treatment on a disease outcome controlling for the effects of a second later treatment. I will show that, in the absence of unmeasured confounders, one can construct valid tests of the null hypothesis of no direct treatment effect in a prospective cohort study, but not in a case–control study in which the control sampling fraction is unknown. In actual case–control studies, the control sampling fraction is often unknown, as when controls are selected either by random digit dialing or from the case's nearest geographical neighbors. However, I will show that when the disease under study is rare in the population, as is often the case in case–control studies, approximately valid tests of the direct effect null hypothesis can be constructed.

1. COMMENTS ON PARTICULAR POINTS

Rosenbaum somewhat privileges studies based on comparisons between groups at a given time rather than within group or within subject comparisons over time. Although I would often agree with this choice, I would not always. Indeed, single subject randomized repeated crossover trials of a particular intervention interspersed by washout periods seem quite reasonable, provided one believes that the washout period is sufficiently long and the number of repeats is large. In observational studies, an analogous design is the so-called case-crossover study of short-acting exposures (MacClure, 1991). Such designs have been rather successful in confirming that unusually vigorous physical exertion and sexual intercourse are triggers for myocardial infarction, by demonstrating a higher than normal incidence of the hypothesized triggering activity in the hour or two before onset of chest pain. The bouts of "treatment" are not assigned at random, and so confounding factors, such as time of day, day of the week, recency of a major meal, and so on, need to be controlled for in the analysis.

In the minimum wage study of Rosenbaum's Section 2.2, I would have some concern that between-state differences in employment could fluctuate

James M. Robins is Professor, Departments of Epidemiology and Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115 (e-mail: robins@hsph.harvard.edu).

by the amount theoretically attributable to the minimum wage treatment for unrecorded reasons, such as closure of a large manufacturing plant or the election of a fiscally conservative governor. Thus, I would be interested in seeing data on the between-state differences in employment rates in restaurants for many pairs of neighboring states, neither of which changed their minimum wage. This criticism is connected with the idea that it may be inappropriate to view each restaurant's employment as independent of one another. Further, a state's decision to pass a change in minimum wage may be a consequence of poorly measured economic and social factors that affect employment.

Although I agree with Rosenbaum that it is easier to get a clear idea about causal effects in studies of abrupt short-lived treatments (provided the effect of the brief treatment can be large), the policy question at issue may concern long-term, low-dose treatments such as determining the age-specific dose of exogenous estrogens that would be optimal for postmenopausal women or the risk associated with occupational exposure to various levels of low-dose radiation. In medical and epidemiological contexts, most treatments are given over prolonged time intervals.

Finally, in studies with time-varying or sequential treatments, one no longer has the option of following Rosenbaum in defining a covariate as a variable measured prior to treatment, because time-varying covariates are prior to later treatments but affected by earlier treatments. Often, the solution to this conundrum is to base estimation on the g-computation algorithm formula given in Section 3 below.

2. A THOUGHT EXPERIMENT

Consider the data given in Table 1; E is a correctly classified exposure of interest whose causal effect on an outcome D we would like to ascertain; E^* is a possibly misclassified version of E. We are interested in the effect of E on D. Data on E, E^* and D are available on all study subjects. Sampling variability can be ignored. I will now describe the designs of three different studies. For each study, the data are the same. Only the designs are different. We wish to answer the following question for each of the studies.

QUESTION. What association measure is most likely to have a causal interpretation?

As a guide, we present some candidate association measures. In Table 2, we calculate the exposure–disease odds ratio $OR_{ED} = 2.33$. We can also calculate the conditional ED odds ratio within strata of E^*, that is, $OR_{ED\,|\,E^*=1} = OR_{ED\,|\,E^*=0} = 3$. Similarly, we calculate that $OR_{E^*D} = 1$ and $OR_{E^*D\,|\,E=1} = OR_{E^*D\,|\,E=0} = 0.6$. We will report all associations on an odds ratio scale, although this is by no means the only or even the best scale on which to measure the effect. This choice is dictated by the fact that in study (a) below, the only population association measures that are identified from the data are odds ratios.

(a) CASE–CONTROL STUDY. Suppose the data arose from a case–control study of the effect of a particular nonsteroidal antiinflammatory drug (E) on a congenital defect (D) that arises in the second trimester of pregnancy. Cases ($D = 1$) are infants with the congenital defect. Controls ($D = 0$) are infants without the defect. The control sampling fraction is unknown. Note that in case–control studies the term "control" has a meaning different from that in Rosenbaum's paper. The data E^* were obtained one month postpartum by asking each mother whether she had taken drug E during the first trimester. The data E were obtained from a comprehensive accurate HMO record of first trimester medications. All relevant preconception confounders and other drug exposures were controlled by stratification. The data in Table 1 are taken from a particular stratum.

(b) PROSPECTIVE COHORT STUDY. Suppose the data were obtained from a follow-up study of total mortality (D) in a cohort of short-term uranium miners, all of whom only worked underground in 1967. The follow-up is complete through 1997. Suppose, for simplicity only, there is a biological threshold dose below which exposure to radon is known to have no effect on mortality. Let $E = 1$ denote above-threshold exposure to radon as mea-

TABLE 1

	$D = 1$			**$D = 0$**	
	$E^* = 1$	**$E^* = 0$**		**$E^* = 1$**	**$E^* = 0$**
$E = 1$	180	100	$E = 1$	600	200
$E = 0$	20	100	$E = 0$	200	600
	$OR = 9$			$OR = 9$	

TABLE 2

	$E = 1$	**$E = 0$**
$D = 1$	280	120
$D = 0$	800	800
	$OR = 2.33$	

sured with a perfectly accurate dosimeter. Similarly, let $E = 0$ denote exposure to below-threshold levels of radon. Each miner was also assigned an estimated radon exposure based on area sampling conducted in the particular mine in which he was employed. Let $E^* = 1$ denote an estimated above-threshold radon exposure and $E^* = 0$ denote an estimated below-threshold radon exposure. The investigators have stratified on the usual demographic factors and life-style risk factors (measured in 1967) such as cigarette smoking and blood pressure. The data come from a single joint level of all these potential confounders. The original assignment in 1967 of miners to particular mines was unrelated to their underlying health status. Further, a subject's actual exposure E depends not only on the level of radon E^* in the mine but also on the particular demands of the subject's job, such as the amount of exertion and thus the minute ventilation required to perform the requisite work. Thus, a subject's actual radon exposure E can differ from the estimated exposure E^*.

(c) RANDOMIZED CLINICAL TRIAL. Suppose the data were obtained from a randomized follow-up study of the effect of low fat diet on death (D) over a 15-year follow-up period. Study subjects were randomly assigned to either a low fat diet educational and motivational intervention arm $(E^* = 1)$ or to a standard care arm $(E^* = 0)$. Investigators were able to obtain accurate measures of the actual diet followed by the study subjects: $E = 1$ if a study subject followed a low fat diet, and $E = 0$ otherwise. Assume E^* has no direct effect on death (D) except through its effect on actual fat consumption E.

2.1 Answers

We now provide the appropriate answers followed by a discussion. In the case–control study (a), the best choice is the marginal odds ratio $OR_{DE} = 2.33$. The other measures are biased. In particular, the conditional odds ratio $OR_{ED\,|\,E^*} = 3$ is biased in the sense that it fails to equal the causal effect of exposure on disease among subjects within a particular stratum of E^*. That is, $OR_{ED\,|\,E^*}$ does not equal the conditional causal odds ratio

$$OR_{\text{causal},\,ED\,|\,E^*} = \frac{\{P[D(1) = 1\,|\,E^*]P[D(0) = 0\,|\,E^*]\}}{\{P[D(1) = 0\,|\,E^*]P[D(0) = 1\,|\,E^*]\}},$$

where $D(j)$ is a subject's potential or counterfactual outcome under exposure level j.

In the prospective cohort study (b), the best choice would be the conditional odds ratio $OR_{DE\,|\,E^*} = 3$, although, as explained below, even this measure is probably somewhat biased in the sense that it differs from the E^*-stratum-specific causal effect of exposure on disease.

In the randomized trial (c), the best choice is the marginal E^*D association $OR_{E^*D} = 1$, suggesting that the exposure E has no causal effect on the outcome D. In this case, both the marginal association $OR_{ED} = 2.33$ and the conditional association $OR_{ED\,|\,E^*} = 3$ are biased estimates of the causal effect of E on D. These answers clearly show that the appropriate statistical analysis depends on the design.

2.2 Justification of the Answers

2.2.1 *Causal graphs.* To justify the answers, we first digress and describe causal directed acyclic graphs (DAGS) as discussed by Pearl and Verma (1991), Spirtes, Glymour and Scheines (1993), Pearl (1995), Pearl and Robins (1995) and Greenland, Pearl and Robins (1999). We first provide an informal discussion. Formal justification will be given in the Appendix.

Informally, a causal graph for the measured variables in a study is a directed acyclic graph (DAG) in which the vertices (nodes) of the graph represent variables measured at specific times, the directed edges (arrows) represent direct causal relations and there are no directed cycles, because no variable can cause itself. The variables represented on the graph include the measured variables and additional unmeasured variables, such that if any two variables on the graph have a cause in common, that common cause is itself included as a variable on the graph. For example, in Figure 1, E and D are the measured variables; U represents all unmeasured common causes of E and D. We have made the arrow from E to D dotted to represent the fact that the purpose of data collection is to determine whether E causes D (i.e., whether the arrow from E to D is actually present).

Suppose our goal is to use the assumptions encoded in our causal graph to determine whether the association OR_{ED} between E and D represents the causal effect of E on D as measured on an odds ratio scale. To do so, we proceed as follows. We begin by pretending that we know that the null hypothesis of no causal effect of E on D is true by removing the arrow from E to D. If, under this null hypothe-

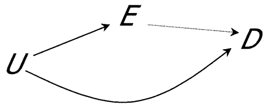

FIG. 1.

sis, E and D are still associated, then obviously the association does not reflect causation, and we say that the association is confounded. The existence of a common cause U of E and D will make E and D associated even under the causal null. If data on U have not been recorded for data analysis, confounding is intractable and we cannot identify the causal effect of E on D. However, if data on U are available, the conditional associations $OR_{ED\,|\,U}$ will represent the causal effect of E on D within strata of U. This reflects the fact that, under the causal null hypothesis of no arrow from E to D, if we condition on all common causes U, then E and D will be conditionally independent. Intuitively, among subjects with identical values of U, E and D can have no common cause and thus should be independent. Furthermore, it is a general result that if E and D are (conditionally) independent under the causal null, then, under the causal alternative, the (conditional) association between E and D will reflect the (conditional) causal effect of E on D. Next we consider the graphs in Figures 2 and 3, where, in addition to E and D, the variable C has been measured. We say a variable U is a cause of another variable, say E, if there is a directed path (sequence of directed arrows) from U to E. Thus, in Figure 3, U remains a common cause of E and D although it is not a direct cause of E. It follows that, in both Figures 2 and 3, the marginal association OR_{ED} is confounded because E and D will be marginally associated even under the causal null. However, the unmeasured variable U will not function as a common cause of E and D within strata of C. Thus $OR_{ED\,|\,C=1}$ and $OR_{ED\,|\,C=0}$ will represent the causal effect of E on D within strata of C.

Consider next Figure 4. There are no unmeasured common causes of E and D. Hence, under the causal null hypothesis of no arrow from E to D, E and D will be uncorrelated. It follows that the marginal association OR_{ED} represents the causal effect of E on D. In contrast, the conditional association $OR_{ED\,|\,C}$ will not be equal to the causal effect of E on D within strata of C, because, under the causal null, E and D will be conditionally dependent within strata of C. To see why, note that the measured variable C is a common effect of E and D. Under the causal null, the common causes E and D of C are independent. However, when one conditions on the common effect of independent common causes, the common causes will be conditionally dependent. To see why, consider the following example due to Pearl (1988). Suppose E encodes whether a sprinkler is on, D encodes whether it is raining and C is the indicator of whether the grass is wet. Then if it rains at random times of the day and the sprinkler is set to go on at times that do not depend on whether it is raining, clearly E and D will be independent, even though they both cause the grass to be wet (C). If we condition on the fact that the grass is wet ($C = 1$), and I tell you that it is not raining ($E = 0$), then you will know for certain that the sprinkler is on ($D = 1$). But if I tell you that it is raining, the probability that the sprinkler is on will not be increased above its marginal probability.

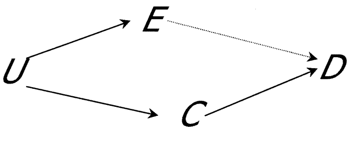

FIG. 2.

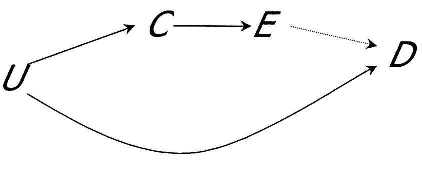

FIG. 3.

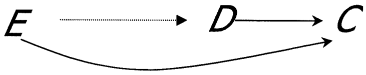

FIG. 4.

An extension of this last example provides an explanation of the well-known adage that one must not adjust for variables affected by treatment. To see why, consider the graph in Figure 5, in which the exposure E has a direct causal effect on C, and C and D have an unmeasured common cause U. Under the causal null with the arrow from E to D removed, E and D will be unassociated because they do not have an unmeasured common cause. Thus, the marginal association OR_{ED} will represent causation. However, the conditional associations $OR_{ED\,|\,C=1}$ and $OR_{ED\,|\,C=0}$ will be biased for the conditional causal effect within levels of C. This reflects the fact that, under the causal null, E and U will be associated once we condition on their common effect C. Thus because U itself is correlated

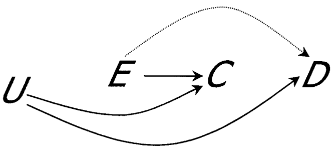

FIG. 5.

with D, E and D will be conditionally associated within levels of C.

With this background, we are ready to justify the answers given above.

2.2.2 *Justifications*

JUSTIFICATION FOR (a). We first argue that the causal graph representing our case–control study is as given in Figure 6. By assumption, we need not worry about unmeasured preconception confounders U. Further, we know that if there is an arrow between E and D, it goes from E to D because the HMO records were created in the first trimester, prior to the development of the second-trimester congenital defect. Also actually taking a medicine will be a cause of a woman reporting that she took a medicine. Hence the arrow from E to E^*. Finally, because a woman's self-report E^* is obtained after the woman discovers that her child has a congenital defect (D), it is at least conceivable that D is a cause of E^*. In fact, it is suspected that women whose children have a congenital defect do a much more thorough job of searching their memory for potential causes of that defect and thus are more likely to recall that they actually took a particular medicine than are women whose children are normal. Furthermore, women of children with a defect may falsely recall having taken the drug in an attempt to come up with some explanation for the defect. Thus, D may well be a cause of E^*. Indeed, if, as we have assumed, the only open question is whether there is an arrow from D to E^*, we can use the data to confirm that indeed such an arrow exists. For if it did not, E^* and D would be independent within levels of E, because conditioning on all common causes of causally unconnected variables renders them independent. But one can check from Table 1 that, among subjects with $E = 1$, D and E^* are correlated. Now Figure 6 is isomorphic to Figure 4 with E^* playing the role of C. Thus, as in Figure 4, we conclude that the marginal association OR_{ED} is causal but the conditional association $OR_{ED\,|\,E^*}$ will differ from the conditional causal effect of exposure and disease within strata of E^*. Mistakenly interpreting $OR_{ED\,|\,E^*} = 3$ as causal could in principle lead to poor public health decisions, as would occur if a cost–benefit analysis determines that a conditional causal odds ratio of 2.9 is the cutoff point above which the risks of congenital malformation outweigh the benefits to the mother of treatment with E. Finally, a possibility that we have not considered is that those mothers who develop, say, a subclinical infection in the first trimester are both at increased risk of a second trimester congenital malformation and of worsening arthritis, which they may then treat with the drug E. In that case, we would need to add to our causal graph an unmeasured common cause U of both E and D that represents subclinical first-trimester infection, in which case OR_{ED} would be confounded.

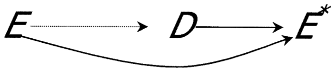

FIG. 6.

JUSTIFICATION FOR (b). In the prospective cohort study, sufficient information is given so that we know there is no confounding by unmeasured preemployment factors. Yet, as noted above, E^* is associated with D given E. Now E^*, which is a measure of the level of radon in mines, cannot itself directly cause death other than through its effect on a subject's actual radon exposure E, so that there cannot be a direct arrow from E^* to D. However, because E^* was measured prior to death, D cannot be a cause of E^* either. The most reasonable explanation for these facts is that E^* is a surrogate for some other unmeasured adverse causal exposure in the mine (say silica). Thus we might consider the causal graph shown in Figure 7. In this figure, *Mine* represents the particular mine in which the subject works. It is plausible that mines with high levels of radon may have low levels of silica-bearing rock (because silica-bearing rock is not radioactive). Therefore, E^* and *silica* will be negatively correlated. If Figure 7 is the true causal graph (with *Mine* and *silica* being unmeasured variables), then, under the causal null hypothesis in which the arrow from E to D is removed, E and D will still remain correlated because *Mine* is an unmeasured common cause of E and D. However, within levels of E^*, *Mine* no longer can act as a common cause. Hence, under the causal null, there will be no conditional association between E and D given E^*, which would imply that $OR_{DE\,|\,E^*}$ has a causal interpretation. In contrast, the conditional association $OR_{E^*D\,|\,E} = 0.6$ represents not a protective

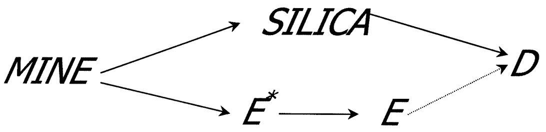

FIG. 7.

effect of E^* on D, but rather the negative correlation between E^* and *silica* conjoined with the adverse causal effect of *silica* on D.

However, Figure 7 probably does not tell the whole story. Recall that the radon level in the mine E^* can differ from a worker's actual level E because the demands of the worker's job also determine E. Similarly, one would expect that the demands of the job also determine a worker's actual silica exposure in conjunction with the air levels of silica associated with the mine. Hence, a more realistic causal graph would probably be Figure 8. On this graph, under the causal null in which the arrow from E to D has been removed, job demands are an unmeasured common cause of both E and D even when we condition on E^*, precluding unbiased estimation of the causal effects of E on D.

JUSTIFICATION FOR (c). The study is a typical randomized trial with noncompliance and is represented by the causal graph in Figure 9 (Balke and Pearl, 1997). Because E^* was randomly assigned, it has no arrows into it. However, given assignment, both the decision to comply and the outcome D may well depend on underlying health status U; E^* has no direct arrow to D, because, by assumption, E^* causally influences D only through its effect on E. We observe that, under the causal null in which the arrow from E to D is removed, E and D will be associated due to their common cause U both marginally and within levels of E^*. Hence, neither OR_{ED} nor $OR_{ED\,|\,E^*}$ will have a causal interpretation. However, under the causal null, E^* and D will be independent, because they have no unmeasured common cause. Hence we can test for the absence of an arrow between E and D (i.e., lack of causality) by testing whether E^* and D are independent. But this, of course, is just the standard intent-to-treat analysis of a randomized trial. Thus, even in the presence of nonrandom noncompliance, an intent-to-treat analysis provides for a valid test of the causal null hypothesis that E does not cause D. Since $OR_{E^*D} = 1$, we conclude it is likely that E does not cause D. Had there been an E^*D association, then we would in fact know that E caused D but we would not be able to determine its magnitude by standard methods, that is, by computing OR_{ED} or $OR_{ED\,|\,E^*}$. In fact, the magnitude of the causal effect on the study population is not identified, and one can only compute the bounds for it. Under further assumptions, Angrist, Imbens and Rubin (1996) show how to compute the magnitude of the effect on the subset of the study population who complied with their assigned treatment. Note that it is logically possible that even though the E^*D association is absent, nonetheless, E does cause D in some individuals and/or protects against D in others. However, this scenario is probably less likely. Finally, note that the conditional association $OR_{E^*D\,|\,E} = 0.6$ fails to have a causal interpretation. This reflects the fact that, under the causal null of no arrow from E to D, E^* and D will be conditionally associated within levels of E, because E is a common effect of both E^* and U and U is a cause of D.

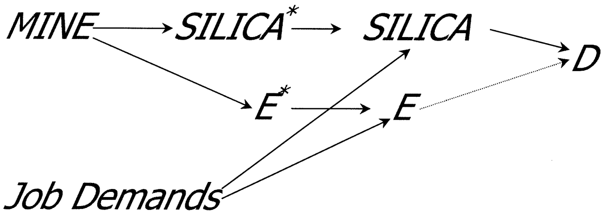

FIG. 8.

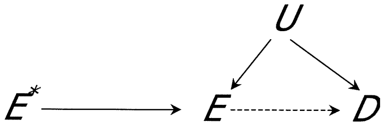

FIG. 9.

COMMENTS. I would not be surprised if many readers are somewhat taken aback by my reduction of complex phenomena to simple graphs that I then blithely endow with causal interpretations, so here is a defense. Yes, the world is much more complex than I have made out but if we do not learn how to reason correctly in simple causal Gedanken-experiments like that above, we have no chance of success in realistic situations. Indeed, the history of epidemiologic methods can be read as an increasingly systematic approach to recognize and classify settings where association is not causation. In 1980, before I had heard of either counterfactuals or causal graphs, I would have answered the questions above successfully, but my answers would have required much more story telling and appeal to heuristic study-specific arguments. As a result, these earlier explanations did not easily generalize and were often difficult for students unfamiliar with epidemiological studies to grasp. The development of formal counterfactual causal models for sequential and time-varying treatments (Robins, 1986, 1987) in the 1980s helped to codify the underlying common principles, but the insights gleaned from these models were often hard to communicate in the "vernacular" to mathematically untrained epidemiologists. The recognition by Pearl (1995) and Spirtes, Glymour and Scheines (1993) that these formal causal models could be encoded in causal DAGS and that complex statistical relations between variables could also be encoded in the simple graphical representations and algorithms described above is an advance, with positive effects on both clarity of thought and ease of communication.

Having said this, I admit I remain nervous about the ease with which users can be convinced they understand a causal process by reifying it in graphs. Part of the difficulty is that unsophisticated users do not really appreciate all the assumptions encoded in a causal graph, and thus treat the world as if it were no more complex than our Gedankenexperiment. Indeed, the informal explanation of the graphs I gave above glosses over certain subtleties that are discussed in the Appendix. One important point is that identification of causal effects from observational data is only possible if one can assume that there are certain pairs of variables with no important unmeasured common causes. As emphasized by Rosenbaum, good design choices can make such an assumption more plausible. However, often the more knowledge one has about the substantive area under investigation, the less plausible such assumptions may seem. For example, in the above case–control study, non-subject-matter experts may not have had the background needed to recognize the possibility that a subclinical first-trimester infection might be a common cause of exposure to E and the outcome D. But, of course, whether or not one uses graphs as aids to causal reasoning does not change the fact that it takes highly skeptical subject-matter experts to elaborate the rich, complex causal stories that comprise the alternative causal theories whose importance Rosenbaum rightly emphasized.

3. SEQUENTIAL TREATMENTS

In Section 3.1, I will illustrate in the simplest possible setting the difficulties that can arise in drawing causal inferences from randomized or observational studies with time-varying or sequential treatments. In Section 3.2, I will discuss the implications of these difficulties for the design of case–control studies.

3.1 A Sequential Randomized Trial

3.1.1 *Description of the trial.* The event tree in Figure 10 represents the data obtained from a hypothetical (oversimplified) sequential randomized trial of the joint effects of AZT (A_0) and aerosolized pentamidine (A_1) on the survival of AIDS patients (Robins, 1997). AZT inhibits the AIDS virus. Aerosolized pentamidine prevents pneumocystis pneumonia (PCP), a common opportunistic infection of AIDS patients. The trial was conducted as follows. Each of 32,000 subjects was randomized with probability 0.5 to AZT ($A_0 = 1$) or placebo ($A_0 = 0$) at time t_0. All subjects survived to time t_1. At time t_1, it was determined whether a subject had had an episode of PCP ($L_1 = 1$) or had been free of PCP ($L_1 = 0$) in the interval $(t_0, t_1]$. Because PCP is a potential life-threatening illness, all subjects with $L_1 = 1$ were treated with aerosolized pentamidine (AP) therapy ($A_1 = 1$) at time t_1. Among subjects who were free of PCP ($L_1 = 0$), one-half were randomized to receive AP at t_1 and one-half were randomized to placebo ($A_1 = 0$). At time t_2, the vital status was recorded for each subject with $Y = 1$ if alive and $Y = 0$ if deceased. We view A_0, L_1, A_1, Y as random variables with realizations a_0, l_1, a_1, y. All investigators agreed that the data supported a beneficial effect of treatment with AP ($A_1 = 1$) because among the 8,000 subjects with $A_0 = 1$ and $L_1 = 0$, AP was assigned at random and the survival rates were greater among those given AP:

$$\begin{aligned} &P[Y = 1 \mid A_1 = 1, L_1 = 0, A_0 = 1] \\ &\quad - P[Y = 1 \mid A_1 = 0, L_1 = 0, A_0 = 1] \\ &= 3/4 - 1/4 = 1/2. \end{aligned} \tag{3.1}$$

The remaining question was whether subjects, given that they were to be treated with AP, should also be treated with AZT. That is, we wish to determine whether the direct effect of AZT on survival controlling for (the potential intermediate variable) AP is beneficial or harmful (when all subjects receive AP). The most straightforward way to examine this question is to compare the survival rates in groups with a common AP treatment who differ on their AZT treatment. Reading from Figure 10 we observe, after collapsing over the data on L_1-status, that

$$\begin{aligned} &P[Y = 1 \mid A_0 = 1, A_1 = 1] \\ &\quad - P[Y = 1 \mid A_0 = 0, A_1 = 1] \\ &= 7/12 - 10/16 = -1/24, \end{aligned} \tag{3.2}$$

suggesting a harmful effect of AZT. However, the analysis in (3.2) fails to account for the possible confounding effects of the extraneous variable PCP (L_1). (We refer to PCP here as an "extraneous variable" because the causal question of interest, i.e., the question of whether AZT has a direct effect on survival controlling for AP, makes no reference to PCP. Thus adjustment for PCP is necessary only insofar as PCP is a confounding factor.) It is commonly

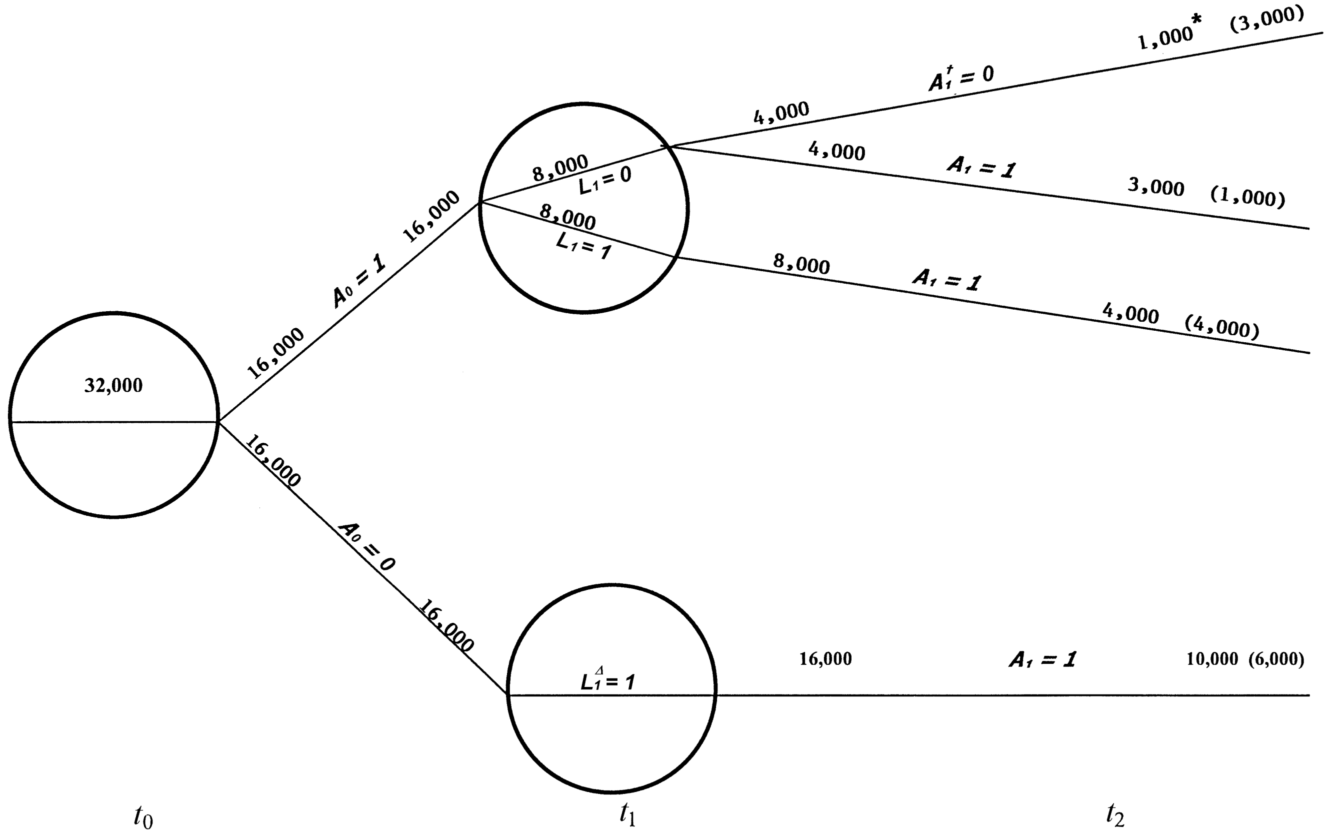

FIG. 10. *Date from a hypothetical study:* (∗) *survivors* ($Y = 1$) *at* t_2 [*Deaths* ($Y = 0$) *at* t_2 *in parentheses*]; (†) A_k *measured just after time* t_k, $k = 0$, *1*; (Δ) L_1 *measured at time* t_1.

accepted that PCP is a confounding factor and must be adjusted for in the analysis if PCP is (a) an independent risk (i.e., prognostic) factor for the outcome and (b) an independent risk factor for (predictor of) future treatment. By "independent" risk factor in (a) and (b) above, we mean a variable that is a predictor conditional upon all other measured variables occurring earlier than the event being predicted. Hence, to check condition (a), we must adjust for A_0 and A_1; to check condition (b), we must adjust for A_0.

Reading from Figure 10, we find that conditions (a) and (b) are both true:

$$\begin{aligned} 0.5 &= P\big[Y = 1 \mid L_1 = 1, A_0 = 1, A_1 = 1\big] \\ &\neq P\big[Y = 1 \mid L_1 = 0, A_0 = 1, A_1 = 1\big] \\ &= 0.75 \end{aligned} \tag{3.3}$$

and

$$\begin{aligned} 1 &= P\Big[A_1 = 1 \mid L_1 = 1, A_0 = 1\Big] \\ &\neq P\big[A_1 = 1 \mid L_1 = 0, A_0 = 1\big] = 0.5. \end{aligned} \tag{3.4}$$

The standard approach to the estimation of the direct effect of AZT controlling for AP in the presence of a confounding factor (PCP) is to compare survival rates among groups with common AP and confounder history (e.g., $L_1 = 1$, $A_1 = 1$) but who differ in AZT treatment. Reading from Figure 10, we obtain

$$\begin{aligned} &P\big[Y = 1 \mid A_0 = 1, L_1 = 1, A_1 = 1\big] \\ &\quad - P\big[Y = 1 \mid A_0 = 0, L_1 = 1, A_1 = 1\big] \\ &\qquad = 4{,}000/8{,}000 - 10{,}000/16{,}000 \\ &\qquad = -1/8. \end{aligned} \tag{3.5}$$

Hence the analysis adjusted for PCP also suggests an adverse direct effect of AZT on survival controlling for AP.

However, the analysis adjusted for PCP is also problematic, because, as discussed in Section 2 and in Rosenbaum (1984) and Robins (1986, 1987), it is inappropriate to adjust (by stratification) for an extraneous risk factor for the outcome that is itself affected by treatment. Reading from Figure 10, we observe that PCP is affected by previous treatment, that is,

$$\begin{aligned} 0.5 &= P\big[L_1 = 1 \mid A_0 = 1\big] \\ &\neq P\big[L_1 = 1 \mid A_0 = 0\big] = 1. \end{aligned} \tag{3.6}$$

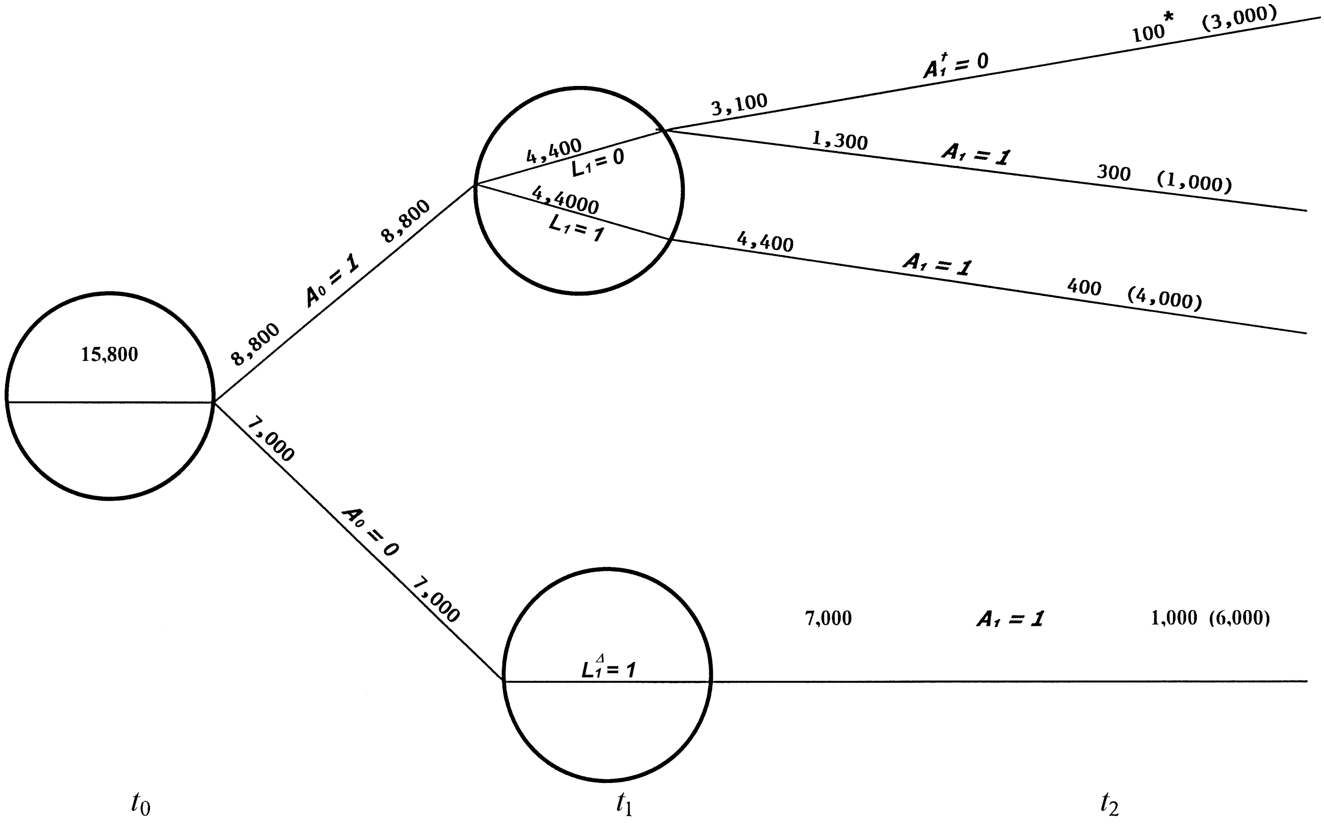

FIG. 11. *Date from a hypothetical study*: (∗) *survivors* ($Y = 1$) *at* t_2 [*Deaths* ($Y = 0$) *at* t_2 *in parentheses*]; (†) A_k *measured just after time* t_k, $k = 0, 1$; (Δ) L_1 *measured at time* t_1.

Thus, according to standard rules for the estimation of causal effects, one cannot adjust for the extraneous risk factor PCP, because it is affected by a previous treatment (AZT); yet one must adjust for PCP because it is a confounder for a later treatment (AP). Thus it may be that, in line with the adage that association need not be causation, neither (3.2) nor (3.5) may represent the direct causal effect of AZT controlling for AP. Because both treatments (AZT and AP) were randomized, one would expect that there should exist a "correct" analysis of the data such that the association observed in the data under that analysis has a causal interpretation as the direct effect of AZT controlling for AP. In the next subsection, we derive such a "correct" analysis based on the G-computation algorithm of Robins (1986). We show that there is no direct causal effect of AZT controlling for AP. That is, given that all subjects take AP, whether or not AZT is also taken is immaterial to the survival rate in the study population.

Suppose, however, the data from our trial were as in Figure 11. We shall show in the next subsection that when the data in Figure 11 are appropriately analyzed using the G-computation algorithm, the analysis reveals a direct AZT effect.

3.1.2 *The* G-*computation algorithm.* In this subsection, we describe how to analyze the data from a simple sequential randomized trial. In a study with sequential treatments, let $\overline{A}_K = (A_0, A_1, \dots, A_K)$ be the temporally ordered $(K + 1)$-vector consisting of the treatment variables of interest. Denote by t_k the time at which treatment A_k is received. Let L_k be the vector of all variables whose temporal occurrence is between treatments A_{k-1} and A_k, with L_0 being the variables preceding A_0 and L_{K+1} being the variables succeeding A_K. Hence, $\overline{L}_{K+1} = (L_0, \dots, L_{K+1})$ is the vector of all nontreatment variables. For notational convenience, define $\overline{A}_{-1}, \overline{L}_{-1}, \overline{V}_{-1}$ to be identically 0 for all subjects. We view $\overline{A}_K$ as a sequence of treatment (control, exposure) variables whose causal effect on the assumed univariate outcome $Y \equiv L_{K+1}$, measured at the end of the study, we wish to evaluate. Let $\overline{a} = \overline{a_K} = (a_0, \dots, a_K)$ denote possible realizations of the random vector $\overline{A}_K$. Let $Y(\overline{a})$ be the counterfactual variable representing a subject's outcome if, possibly contrary to fact, the subject had received the treatment $\bar a$ rather than the observed treatment $\bar A_K$.

Sequential randomization guarantees that, for any $\bar a$, treatment A_k received at t_k is conditionally independent of $Y(\bar a)$ given the observed past $\bar L_k$, $\bar A_{k-1} = \bar a_{k-1}$:

$$Y(\bar a) \coprod A_k \,\Big|\, \bar L_k, \bar A_{k-1} = \bar a_{k-1}, \tag{3.7}$$

which we refer to as the assumption of no unmeasured confounders. It is a sequential version of Rosenbaum and Rubin's (1983) strong ignorability assumption. Under natural consistency and positivity assumptions, Robins (1987) proves the following.

THEOREM 3.1. *Equation* (3.7) *implies*

$$\begin{aligned} f_{Y(\bar a)}&(y \mid \bar L_k = \bar\ell_k, \bar A_{k-1} = \bar a_{k-1}) \\ &= \int \cdots \iint f(y \mid \bar\ell_K, \bar a_K) \\ &\qquad \cdot \prod_{j=k+1}^{K} f(\ell_j \mid \bar\ell_{j-1}, \bar a_{j-1})\, d\mu(\ell_j), \end{aligned} \tag{3.8}$$

$$\begin{aligned} f_{Y(\bar a)}(y) &= \int \cdots \iint f(y \mid \bar\ell_K, \bar a_K) \\ &\qquad \cdot \prod_{j=0}^{K} f(\ell_j \mid \bar\ell_{j-1}, \bar a_{j-1})\, d\mu(\ell_j), \end{aligned} \tag{3.9}$$

where μ is a dominating measure and the unsubscripted densities refer to densities of the non-counterfactual random variables, for example, $f(y \mid \bar\ell_K, \bar a_K) = f_Y(y \mid \bar L_K = \bar\ell_K, \bar A_K = \bar a_K)$. The RHS of (3.9) *is known as the* g-computation *algorithm formula or functional* (*Robins,* 1986). *Equation* (3.9) *states that the marginal density of $Y(\bar a)$ is obtained from the joint distribution of the observables by taking a weighted average of the $f(y \mid \bar\ell_K, \bar a_K)$ with weights proportional to*

$$\omega(\bar\ell_K) \equiv \prod_{j=0}^{K} f[\ell_j \mid \bar\ell_{j-1}, \bar a_{j-1}].$$

Equation (3.8) *has a similar interpretation except that it conditions on the covariate history $\bar\ell_k, \bar a_{k-1}$.*

EXAMPLE. *A correct analysis of the trial of Section* 3.1.1. We can use (3.9) to calculate causal effects. For example, suppose the data in the trial were as in Figure 10, and we let $K = 1, L_0 \equiv 0, Y = L_2$. Then the probability a subject would survive to $t_2 (Y = 1)$ if all subjects were treated with AZT at t_0 and aerosolized pentamidine at t_1 is $f_{Y(\bar a)}(y = 1)$ with $\bar a = (1, 1)$ and equals, by (3.9),

$$\begin{aligned} \sum_{\bar\ell_1} & f(y = 1 \mid \bar\ell_1, \bar a_1) f(\bar\ell_1 \mid \ell_0, \bar a_0) \\ &= f(y = 1 \mid \ell_1 = 1, a_1 = 1, a_0 = 1) \\ &\qquad \cdot f(\ell_1 = 1 \mid a_0 = 1) \\ &\quad + f(y = 1 \mid \ell_1 = 0, a_0 = 1, a_1 = 1) \\ &\qquad \cdot f(\ell_1 = 0 \mid a_0 = 1) \\ &= \left(\frac{4{,}000}{8{,}000}\right)\left(\frac{8{,}000}{16{,}000}\right) \\ &\quad + \left(\frac{3{,}000}{4{,}000}\right)\left(\frac{8{,}000}{16{,}000}\right) \\ &= \frac{10{,}000}{16{,}000}. \end{aligned}$$

Similarly, $f_{Y(\bar a)}(y = 1)$ for $\bar a = (0, 1)$ is 10,000/16,000. Hence, there is, by definition, no direct effect of AZT on survival controlling for AP (when all subjects take AP).

In contrast, if the data arose from Figure 11, then

$$\begin{aligned} f_{Y(\bar a = (1,1))}&(y = 1) \\ &= \left(\frac{300}{1{,}300}\right)\left(\frac{4{,}400}{8{,}800}\right) \\ &\quad + \left(\frac{400}{4{,}400}\right)\left(\frac{4{,}400}{8{,}800}\right) = 0.12 \end{aligned}$$

and

$$f_{Y(\bar a = (0,1))}(y = 1) = 1{,}000/7{,}000 = 0.14,$$

indicating an adverse direct effect of AZT on survival.

REMARK. In observational studies with sequential treatments, it is a primary goal of an epidemiologist to collect in L_k data on a sufficient number of covariates to try to make assumption (3.7) at least approximately true. However, (3.7) cannot be guaranteed to hold even approximately and is not subject to empirical testing.

3.2 Implications for the Design and Analysis of Case–Control Studies

In an observational case–control study one collects treatment and covariate data on all cases (deaths) and a random sample (the controls) of the survivors. If the cohort data is as in Figure 10 and we use a control sampling fraction of 0.1, the case–control data will be precisely the data in Figure 11. If the cohort data is as in Figure 11 and we use a control sampling fraction of 1.0, the case–control data will again be the data in Figure 11. Thus, if we have collected the data in Figure 11 in a case–control study with an unknown sampling fraction, we cannot determine if the full cohort data is as in Figure 10 or as in Figure 11, and thus we cannot deduce whether or not AZT has a direct effect on survival. That is, even given (3.7), the causal null hypothesis of no direct AZT effect is not identified from case–control data with an unknown control sampling fraction.

In contrast, given that (3.7) holds, one can test the sharp null hypothesis

$$Y(\bar{a}) = Y \quad \text{with probability 1 for all } \bar{a}$$

of no joint effect of the treatments A_0 and A_1 on survival from case–control data with an unknown sampling fraction based on the following g-null theorem proved in Robins (1986).

THEOREM 3.2. *The right-hand side of* (3.8) *is the same for all* $\bar{a}$ *and* k *if and only if, for each* k,

$$(3.10) \qquad Y \coprod A_k \mid \bar{L}_k, \bar{A}_{k-1}.$$

Theorem 3.2 implies that, in the trial in Section 3.1.1, if the sharp null hypothesis of no joint treatment effect were true, then it would be the case that $\text{pr}(A_0 = 1 \mid Y = 1) = \text{pr}(A_0 = 1 \mid Y = 0)$ and $\text{pr}(A_1 = 1 \mid Y = 1, L_1, A_0) = \text{pr}(A_1 = 1 \mid Y = 0, L_1, A_0)$. Note that each of these conditional probabilities can be identified from case–control data; thus, their values are the same in Figure 10 as in Figure 11. Further, calculating from either figure, we find that both equalities are false so the joint null hypothesis is rejected. Given (3.7), with nonsequential treatments (i.e., $K = 0$, so $\bar{A}_K = A_0$) as in Section 2, the casual marginal odds ratio

$$OR_{\text{causal}} = \frac{\{\text{pr}[Y(1) = 1]\text{pr}[Y(0) = 0]\}}{\{\text{pr}[Y(1) = 0]\text{pr}[Y(0) = 1]\}}$$

can be identified from case–control data, when $L_0 \equiv 0$. Under the these same conditions, with sequential treatments, the causal marginal odds ratios

$$OR_{\text{causal}}(a_0, a_1) = \frac{\{\text{pr}[Y(a_0, a_1) = 1]\text{pr}[Y(0,0) = 0]\}}{\{\text{pr}[Y(a_0, a_1) = 0]\text{pr}[Y(0,0) = 1]\}}$$

cannot be identified from case–control data with unknown sampling fraction, whenever the joint causal null is false. To prove this result, one simply evaluates the causal marginal odds ratios using the g-computation algorithm formula separately in Figures 10 and 11, and notes that two different answers are obtained.

APPROXIMATELY VALID INFERENCE FOR RARE OUTCOMES. Define $w(a_1, l_1, a_0) = 1/P(A_1 = a_1 \mid L_1 = l_1, A_0 = a_0)$ and $W = w(A_1, L_1, A_0)$. Results on inverse probability of treatment weighted estimators of marginal structural models in Robins (1999) imply that if we obtain an estimator $\widehat{\beta}$ of the parameter β of the logistic regression model $\log it\, P(D = 1 \mid A_0, A_1) = \beta_0 + \beta_1 A_0 + \beta_2 A_1 + \beta_3 A_0 A_1$ by fitting our case–control data by weighted logistic regression with subject-specific weights W, then $\exp(\widehat{\beta}_1 a_0 + \widehat{\beta}_2 a_1 + \widehat{\beta}_3 a_0 a_1)$ will, under (3.7) with $L_0 \equiv 0$, converge to $OR_{\text{causal}}(a_0, a_1)$. In particular $\widehat{\beta}_1$ and $\widehat{\beta}_3$ will converge to zero if AZT (A_0) has no direct effect controlling for A_1. Now in an observational case–control study with unknown control sampling fraction, W is not identified. However, if the outcome is rare (e.g., $P(D = 1 \mid A_0, A_1, L_1) < 0.03$ with probability 1), then $w(a_1, l_1, a_0)$ is, in general, approximately equal to $w^*(a_1, l_1, a_0) = 1/P(A_1 = a_1 \mid L_1 = l_1, A_0 = a_0, Y = 0)$, which can be estimated from case–control data with unknown control sampling fraction. Thus, to obtain approximately valid estimates and tests for a the direct effect of A_0 controlling for A_1, one fits the above logistic model using estimated weights $\widehat{W}^* = \widehat{w}^*(A_1, L_1, A_0)$.

APPENDIX: CAUSAL GRAPHS AND CONFOUNDING

Robins (1986, 1987) proposed a set of counterfactual models based on event trees, one of which is a causal graph as defined below. Let G be a directed acyclic graph (DAG) with nodes (vertices) $V = (V_1, \dots, V_M)$, where a DAG is a graph in which all edges are directed and there are no directed cycles. If on a DAG G there is a directed edge (equivalently, arrow or arc) from V_k to V_m, we say V_k is a parent of V_m. If there is a sequence of directed edges from V_k to V_m, we say that V_k is an ancestor of V_m and V_m is a descendant of V_k. We let G^* be the complete graph (i.e., graph with no missing arrows) in which $\bar{V}_{m-1} \equiv (V_1, \dots, V_{m-1})$ are V_m's parents, where we always choose the order of the V_m's such that G is a subgraph of G^* (i.e., G is obtained from G^* by removing arrows).

Statistical DAGs. A law F_V of V is represented by a DAG G, if the law has a density $f_V(v)$ that factorizes as

$$(A.1) \qquad f_V(v) = \prod_{j=1}^{M} f_{V_j \mid Pa_j}(v_j \mid pa_j),$$

where Pa_j are the parents of V_j on G and pa_j and v_j are realizations. Geiger, Verma and Pearl

(1990) showed that if (A.1) holds, then, for any disjoint sets of variables X, Y and Z contained in V, X and Y are conditionally independent given Z (i.e., $X \coprod Y \mid Z$) if X and Y are d-separated by Z on G (written $(X \coprod Y \mid Z)_G$). D-separation is a purely graphical criterion described by Pearl (1995) as follows.

Definition (d-Separation). Let X, Y and Z be three disjoint subsets of nodes in a directed acyclic graph G, and let p be any path between a node in X and a node in Y, where by "path" we mean any succession of arcs, regardless of their directions. Then Z is said to block p if there is a node w on p satisfying one of the following two conditions: (i) w has converging arrows along p, and neither w nor any of its descendants is in Z; (ii) w does not have converging arrows along p, and w is in Z. Further, Z is said to d-separate X from Y, in G, written $(X \coprod Y \mid Z)_G$, if and only if Z blocks every path from a node in X to a node in Y.

Definition of causal graphs. For any subset of variables $X \subset V$, let $V_m(x)$ be the random variable encoding the value of the variable V_m had, possibly contrary to fact, X been set to x. Note here we have assumed that the variables X are manipulable and that the counterfactuals $V_m(x)$ are well defined.

Definitions (Robins, 1986, pages 1419–1423). We say the following:

(a) The complete DAG G^* is a finest causal graph if (i) V_i and all one-step-ahead counterfactuals $V_m(\overline{v}_{m-1})$ exist for $m > 1$ and (ii) the observed variables V and, for any subset X contained in V, the counterfactual variables $V_m(x)$ are obtained by recursive substitution, for example, $V_3 \equiv V_3\{V_1, V_2(V_1)\}$, $V_3(v_1) = V_3\{v_1, V_2(v_1)\}$, $V_3(v_2) = V_3\{V_1, v_2\}$;

(b) DAG G is a finest causal graph if G^* is a finest causal graph and $V_m(\overline{v}_{m-1}) = V_m(pa_m)$ depends on $\overline{v}_{m-1}$ only through V_m's parents on G;

(c) a finest causal graph is a finest fully randomized causal graph if, for all m,

$$\{V_{m+1}(\overline{V}_{m-1}, v_m), \ldots, V_M(\overline{V}_{m-1}, v_m, \ldots, v_{M-1})\} \coprod V_m \mid \overline{V}_{m-1}.$$

Definition. We simply say G is a causal graph if it is a finest fully randomized causal graph.

Pearl (1995) originally gave an alternative, but equivalent definition of a causal graph as a nonparametric structural equations model.

It is easy to show that if G is a causal graph over variables V, then the density $f_V(v)$ factorizes as in (3.1). Furthermore if we let $V = (\overline{A}_K, \overline{L}_{K+1})$ with $\overline{A}_K, \overline{L}_{K+1}$ as defined in Section 3.1.2, so that the variables in $\overline{A}_k$ and $\overline{L}_{k+1}$ are nondescendants of A_{k+1}, then G being a causal graph implies that the assumption (3.7) of no unmeasured confounders holds.

A.1 Confounding

The results in this section provide the formal justification for our informal analysis of the thought experiment in Section 1.

Remark. Our formal results will only use the fact that the law of V is represented by a given DAG G and that the assumption (3.7) of no unmeasured confounders holds. In particular they do not require that G be a causal graph. Indeed they do not require that any other counterfactuals other than the counterfactuals $Y(\overline{a})$ be well defined, so we do not need to think of the nontreatment variables $\overline{L}_{K+1}$ as manipulable.

Suppose we believe (3.7) holds but that a subset $\overline{U}_{K+1}$ of the variables $\overline{L}_{K+1}$ is not observed. The observed subset $\overline{O}_{K+1}$ of $\overline{L}_{K+1}$ includes the outcome variable $Y = L_{K+1}$. Define $\overline{U}_k = \overline{L}_k \cap \overline{U}_{K+1}$ and $\overline{O}_k = \overline{L}_k \cap \overline{O}_{K+1}$ to be the unobserved and observed nontreatment variables through time t_k. The goal of this section is to define restrictions on the joint distribution of $V = (\overline{L}_{K+1}, \overline{A}_K) \equiv (\overline{L}, \overline{A})$ such that (3.7) will imply that

$$Y(\overline{a}) \coprod A_k \mid \overline{O}_k, \overline{A}_{k-1} = \overline{a}_{k-1} \tag{A.2}$$

because, then, by Theorem 3.1, $f_{Y(\overline{a})}(y)$ is identified from data $(\overline{A}_K, \overline{O}_{K+1})$ and can be computed by the g-computation algorithm formula (3.9) with o substituted for ℓ.

We now present a sufficient graphical condition for this result under the assumption that the law of V is represented by the statistical DAG G; that is, (3.1) holds. Let $G_k^{\overline{a}}$ be the DAG that has no arrows out of A_k, has no arrows into the A_m for $m > k$ and is elsewhere identical to G. We then have the following.

Theorem A.1 (Pearl and Robins, 1995). *If*

$$\left(Y \coprod A_k \mid \overline{O}_k, \overline{A}_{k-1}\right)_{G_k^{\overline{a}}}, \quad k \le K, \tag{A.3}$$

then (3.7) *implies* (A.2).

Here $(A \coprod B \mid C)_{G_k^{\overline{a}}}$ stands for d-separation of A and B given C in $G_k^{\overline{a}}$ (Pearl, 1995). Note that checking d-separation is a purely graphical (i.e., visual) procedure. Pearl (1995) had earlier proved Theorem A.1 for non-sequential treatments (i.e., in the case $K = 0$) and named it the no-back-door path criterion.

A.1.1 *Application to the analysis of Section 2.* We repeatedly used Theorem A.1 in our analysis of the three hypothetical studies in Section 2 with $E = A_0$ and $D = Y$. For example, when E^* was temporally prior to E, we used (A.3) to determine whether (A.2) was true with $K = 0$ and $\overline{O}_0 = E^*$, because (A.2) implies, by (3.8) with o substituted for ℓ, that $OR_{ED\,|\,E^*}$ equals the conditional causal odds ratio $OR_{\text{causal},\,ED\,|\,E^*}$ as defined in Section 2

Equation (A.3) can be expressed in terms of the associations of the observed and unobserved variables in a manner more familiar to epidemiologists. Let $G^{\overline{a}.k}$ be the DAG that differs from $G^{\overline{a}}_k$ only in that arrows on G that were out of A_k are restored. We then have

Theorem A.2 (Pearl and Robins, 1995; Robins, 1997). *Equation* (A.3) *is true if and only if, for* $k \leq K$, $U_k = (U_{ak}, U_{bk})$ *for possibly empty mutually exclusive sets* U_{ak}, U_{bk} *satisfying* (i) $(U_{bk} \coprod A_k \,|\, \overline{O}_k, \overline{A}_{k-1})_{G^{\overline{a}\cdot k}}$ *and* (ii) $(U_{ak} \coprod Y \,|\, \overline{O}_k, \overline{A}_k, U_{bk})_{G^{\overline{a}\cdot k}}$.

Remark. The set U_{ak} need not be contained in $U_{a(k+1)}$, and similarly for U_{bk}. Also U_{bk} can always be taken to be the largest subset of U_k satisfying the assumption in (i). The special cases of Theorem A.2 in which treatment is time independent ($K = 0$) and either U_{a0} or U_{b0} is the empty set are the standard conditions for nonconfounding taught in first-year epidemiology courses.

Robins (1994) also proved the following.

Theorem A.3 *Equation* (A.3) *is true if and only if, for* $k \leq K$, U_k *can be divided into three possibly empty mutually exclusive sets* $U_k = (U_{1k}, U_{2k}, U_{3k})$ *as follows*:

(i) *nonancestors* U_{1k} *of both* A_k *and* Y *in* $G^{\overline{a}\cdot k}$;

(ii) *variables* U_{2k} *that are d-separated from* A_k *given* $(\overline{O}_k, \overline{A}_{k-1})$ *in* $G^{\overline{a}\cdot k}$;

(iii) *variables* U_{3k} *that are d-separated from* Y *in* $G^{\overline{a}\cdot k}$ *given* $(\overline{A}_k, \overline{O}_k)$;

(iv) U_{2k} *and* U_{3k} *are d-separated from one another given* $(\overline{O}_k, \overline{A}_{k-1})$ *in* $G^{\overline{a}\cdot k}$.

The special cases of Theorem A.3 in which $K = 0$, U_{1k} is the empty set and either U_{2k} or U_{3k} is the empty set are again the standard conditions for nonconfounding taught in first-year epidemiology courses. We summarize our discussion of confounding in the following definition.

Definition of nonconfounders and potential confounders. Suppose the assumed prior information is that the law F_V of $V = (\overline{L}_{K+1}, \overline{A}_K)$ is represented by a given DAG G such that the variables in $(\overline{A}_k, \overline{L}_{k+1})$ are nondescendants of A_{k+1} and the assumption (3.7) of no unmeasured confounders holds with $Y \equiv \overline{L}_{K+1}$. Then, when (A.3) is also assumed to hold, we say the U_k are nonconfounders for the effect of treatment $\overline{A}_K$ on Y given data on the O_k. When (A.3) is not assumed, we say the U_k are potential confounders for the effect of treatment on Y given data on the O_k.

This definition reflects the fact that, when (A.3) is false, there are many distributions represented by the DAG G for which (A.2) is false, even though (3.7) is true, and yet there generally are still a few special distributions represented by DAG G for which (A.2) is true. Thus, given (3.7), when (A.3) is false we cannot be guaranteed that the g-computation algorithm formula given by the right-hand side of (3.9) with o substituted for ℓ will equal the causal effect $f_{Y(\overline{a})}(y)$, but we also cannot be certain that it will not equal $f^{(Y)}_{Y(\overline{a})}$. However, because data on the U_k are not available, we have no way to test from the data whether the actual distribution of V is one of the special distributions for which they are equal. Thus, we cannot use the g-computation algorithm formula based on the observed data $(\overline{O}_{K+1}, \overline{A}_K)$ to estimate causal effects when the U_k are potential confounders.

The question remains as to whether we can use any other functional of the law of the observables to compute $f^{(Y)}_{Y(\overline{a})}$ when (A.3) is false. The answer is no if on G there are no missing arrows out of the A_k, because then $f_{Y(\overline{a})}(y)$ is not identified from the data $(\overline{O}_{K+1}, \overline{A}_K)$. That is, there will be distributions F_V represented by G which have the same marginal distribution for $(\overline{O}_{K+1}, \overline{A}_K)$ but different values for $f_{Y(\overline{a})}(y)$. Hence, in practice, we must treat potential confounders as actual confounders, and interpret causal effects as uncomputable.

Comment—Design Rules: More Steps Toward a Complete Theory of Quasi-experimentation

William R. Shadish and Thomas D. Cook

> *"You can't fix by analysis what you bungled by design."*—[*Light, Singer and Willett*, 1990, *page* v]

One of the most gratifying trends in statistics has been the increased attention statisticians like Paul Rosenbaum (and Paul Holland and Donald Rubin) have paid to causal inference in quasi-experiments—a term that is not quite synonymous with Rosenbaum's "observational study," but the differences are not crucial for present purposes. Those of us who toil in the trenches of fields like psychology, education and economics know that random assignment is what we would like to do, but that quasi-experiments are what we are sometimes forced to do for practical or ethical reasons. We struggle with how to justify causal inferences from such second-class designs.

Rosenbaum brings a statistician's expertise to bear on a theme dear to our hearts—that researchers can improve causal inference in quasi-experiments by thoughtful choice of design features such as control groups, pretreatment observations on the same scale as the outcome and deliberately varying when a treatment is implemented. We enthusiastically support his advice, in no small part because we come from a school of thought in psychology that for over 40 years has made thoughtful design choices the centerpiece of good experimental work for gaining control over alternative interpretations. We prefer such design choices to trying to measure alternative explanations directly or indirectly and then "somehow" controlling for them in the statistical analysis (Campbell, 1957; Campbell and Stanley, 1963; Cook and Campbell, 1979; Corrin and Cook, 1998; Shadish, Cook and Campbell, 1999). So this commentary is less a rejoinder to Rosenbaum and more a sympathetic extension of his thinking, and a request to make even more systematic the theoretical foundations of quasi-experimentation. Our key theme is reflected in our title's double entendre: in the inevitable give-and-take between statistical analysis and experimental design, it is design that should rule if quality causal inferences are to result; so we need theories of quasi-experimentation that describe which rules for designing quasi-experiments provide better justification for such inferences.

Researchers in economics and statistics are rediscovering this message, but for different reasons. For economists, the rediscovery stems from the failure of elegant statistical selection bias models to deliver predictably accurate effect estimates, despite diverse efforts to do this over several decades. (To quasi-experimentalists, some of these efforts seemed doomed to failure from the start because they involved contrasting a local program group of substantive interest with a purportedly matched control group abstracted from some national record-keeping system.) In any event, many economists eventually realized that these models might work better when supported by stronger design; so Heckman and Todd (1996, page 60) noted their selection-adjustment methods may "perform best when (1) comparison group members come from the same local labor markets as participants, (2) they answer the same survey questionnaires, and (3) when data on key determinants of program participation is (sic) available."

For statisticians, the need to rediscover quasi-experimental design stems from the lack of strong statistical theory for nonrandomized experiments that parallels the elegant theoretical grounding of randomization in the work of R. A. Fisher and his successors. Lacking such a hammer, statisticians swung at the quasi-experimental nail less often than at more congenial targets that suited their tool box better. One of many merits to Rubin's causal model (RCM) is that it provides such a ham-

William R. Shadish is Professor, Department of Psychology, University of Memphis, Campus Box 526400, Memphis, Tennessee 38152-6400 (e-mail: shadish@mail.psyc.memphis.edu). Thomas D. Cook is Professor, Institute for Policy Research, Northwestern University, 2040 Sheridan Road, Evanston, Illinois 60208-4100 (e-mail: t-cook@nwu.edu).

mer, the products of which are increasingly seen in articles like Rosenbaum's. Our brief article seeks to expand Rosenbaum's ideas on design choices in quasi-experiments. We hope to speed up the theoretical integration still much needed to increase the justification for, and use of, certain types of quasi-experiments, simultaneously decreasing the support for others that are not equal to responsible causal inference. We offer the following seven discussion points that complement Rosenbaum's work in hopefully productive and challenging ways.

1. CAUSAL INFERENCE IS MORE A MATTER OF LOGIC THAN STATISTICS

Naturally, statisticians will bring the tools with which they are familiar to bear on any problem. Most professionals do. Yet Rosenbaum's focus on design reminds us that drawing causal inferences without random assignment requires more than statistics alone. After all, scientists had successfully inferred causation long before Fisher's justifications for randomization—discovering both the causes of effects (e.g., Snow's demonstration that cholera was caused by polluted water) and the effects of causes (e.g., Triplett's 1898 demonstration that that the presence of audience and competitors tends to improve performance of bicyclists). They succeeded in this because the logic of causal inference depends primarily on qualitative judgment—such as whether a putative cause precedes its possible effect in time, whether alternative causes can be ruled out and whether we have some plausible idea of what would have happened had the cause not occurred (the counterfactual). Statistics can aid these judgments, often enormously, but they are servants not masters. Design rules because the active manipulation of design choices creates experiments with a structure that recreates the logic of causal inference through knowledge of causal counterfactuals, alternative interpretations and temporal precedence. Statistics cannot actively create such a structure; they can only help describe and interpret certain questions within it, most clearly those concerning covariation between the cause and effect.

2. WE NEED A GREATLY ELABORATED THEORY OF DESIGN RULES FOR QUASIEXPERIMENTS

Rosenbaum describes several design choices; our Table 1 lists more such choices about assignment, measurement, comparison groups and treatment that the experimenter can implement to improve the validity of a causal inference. The two lists overlap (e.g., compare Rosenbaum's Section 3.1 to complex predictions in Table 1, Section 3.2 to comparison groups and Section 3.3 to other nonrandom assignment) but both lists could borrow from each other. However, a theory of design rules for quasi-experiments must do far more than list design choices. It must also (1) describe the function each choice can serve (see the brief examples in Table 1), (2) show its advantages (e.g., nonequivalent dependent variables are often low cost and effective) and disadvantages (e.g., matching can create regression artifacts, treatment removals depend on treatment effects being temporary and multiple pretests sometimes impose a measurement burden that some respondents might refuse to bear) and (3) elaborate how specific design choices reduce specific doubts about causal inference. Rosenbaum's Section 1.2 recognizes the importance of these questions with its call to evaluate the quality of evidence. However, he is less clear about how such an evaluation is to be done when quasi-experiments are at issue. We turn to such questions now.

3. HOW ARE WE TO JUDGE THE QUALITY OF EVIDENCE IN QUASIEXPERIMENTS?

First, what are the features of causal inference that can be doubted when someone questions a causal inference? Rosenbaum's article alludes to such features without identifying them explicitly. For example, the quote from Campbell that starts his article refers in substantial part to doubts about the generalization of causal inferences, while the quote from Meyer and Fienberg on group comparability in comparative studies refers more to the logical warrant asserting that a particular kind of comparison group yields a valid counterfactual. Distinguishing among such causal inference features is important because different design elements help resolve different doubts about different features.

The typology we prefer characterizes each causal inference as having four related features:

- To what extent do the presumed cause and effect covary (statistical conclusion validity)?
- How plausible is the evidence indicating that this covariation is due to a causal relationship from the independent to the dependent variable (internal validity)?
- Which general constructs best characterize the nature of this cause-and-effect relationship given the contextual features of the experiment—features related to particular samples of persons, treatments, settings, observations and times (construct validity)?

TABLE 1

Design elements used in constructing quasi-experiments

Assignment (control of assignment strategies to increase group comparability)

- Cutoff-based assignment: controlled assignment to conditions based solely on one or more fully measured covariates; this yields an unbiased effect estimate.
- Other nonrandom assignment: various forms of “haphazard” assignment that sometimes approximate randomization (e.g., alternating assignment in a two-condition quasi-experiment whereby every other unit is assigned to one condition, etc.)
- Matching and stratifying: efforts to create groups equivalent on observed covariates in ways that are stable, do not lead to regression artifacts and are correlated with the outcome; preference is for pretreatment measures of the outcome itself.

Measurement (use of measures to learn whether threats to causal inference actually operate)

- Posttest observations
  - Nonequivalent dependent variables: measures that are not sensitive to the causal forces of the treatment, but *are* sensitive to all or most of the confounding causal forces that might lead to false conclusions about treatment effects (if such measures show no effect, but the outcome measures do show an effect, the causal inference is bolstered because it is less likely due to the confounds)
  - Multiple substantive posttests: used to assess whether the treatment affects a complex pattern of theoretically predicted outcomes
- Pretest observations
  - Single pretest: a pretreatment measure on the outcome variable, useful to help diagnose selection bias
  - Retrospective pretest: reconstructed pretests when actual pretests are not feasible—by itself, a very weak design feature, but sometimes better than nothing
  - Proxy pretest: when a true pretest is not feasible, a pretest on a variable correlated with the outcome—also often weak by itself
  - Multiple pretest time points on the outcome: helps reveal pretreatment trends or regression artifacts that might complicate causal inference
- Pretests on independent samples: when a pretest is not feasible on the treated sample, one is obtained from a randomly equivalent sample
- Complex predictions such as predicted interaction: successfully predicted interactions lend support to causal inference because alternative explanations become less plausible.
- Measurement of threats to internal validity: helps diagnose the presence of specific threats to the inference that A caused B such as whether units actively sought out additional treatments outside the experiment

Comparison groups (selecting comparisons that are “less nonequivalent” or that bracket the treatment group at the pretest(s))

- Single nonequivalent groups: compared to studies without control groups, using a nonequivalent control group helps identify many plausible threats to validity.
- Multiple nonequivalent groups: serve several functions; for instance, groups are selected that are as similar as possible to the treated group but at least one outperforms it initially and at least one underperforms it, thus bracketing the treated group.
- Cohorts: comparison groups chosen from the same institution in a different cycle (e.g., sibling controls in families or last year’s students in schools)
- Internal (versus external) controls: plausibly chosen from within the same population (e.g., within the same school rather than from a different school)

Treatment (manipulations of the treatment to demonstrate that treatment variability affects outcome variability)

- Removed treatments: showing an effect diminishes if treatment is removed
- Repeated treatments: reintroducing treatments after they have been removed from some group—common in laboratory sciences or where treatments have short-term effects
- Switched replications: reversing treatment and control group roles so that one group is the control while the other receives treatment, but the controls receive treatment later while the original treatment group receives no further treatment or has treatment removed
- Reversed treatments: provides a conceptually similar treatment that reverses an effect—for example, reducing access for some students to a computer being studied but increasing access for others
- Dosage variation (treatment partitioning): demonstrates that outcome responds systematically to different levels of treatment

- In what ways does any cause-and-effect relationship so labeled generalize to units, treatments, observations, settings and times different from those in the study (external validity)?

These four distinctions are interrelated. Each is a different component of the more general concept of dependable causation, and particular pairs among them share overlapping purposes. For instance, both statistical conclusion and internal validity deal with the nature of the association between observables within an experiment, while construct and external validity are both about the generalization of such an association.

What specific doubts can we have about each of these four features of a dependable causal inference? Many cognate phrases are used to characterize the doubts arising from quasi-experiments: third variable confounds; alternative explanations; treatment biases; potential falsifiers; and (in Campbell's theory) threats to validity. Some of the specific doubts have names that are by now almost universally recognized. Selection bias, for example, is widely understood to refer to the possibility that the population in one condition is systematically different from the population in another in ways that might pertain to the outcome. It is a key threat to internal validity. Other doubts are more subject-specific, as with Sackett's (1979) list of biases in epidemiological research. Campbell's lists of specific threats to each of the four validity types seem to apply very broadly, despite being generated from experimental practice in psychology and education. His is one attempt to list individual threats to the validity of each type of inference. For instance, violation of unit independence is a threat to statistical conclusion validity; treatment-correlated attrition from a study threatens internal validity; and experimenter expectancy effects threaten construct validity. The individual threats are too extensive to recite here and are perhaps best summarized in Cook and Campbell (1979).

How do various design features diagnose or reduce the doubts (i.e., rule out threats to validity)? Table 1 provides some examples of how this question might be addressed, and we can add others. For example, in a controlled study, the decision to compare treatment to a placebo rather than to a no-treatment control group is primarily done to clarify the construct validity of the treatment—to what extent should the intended treatment be characterized as the causal agent as opposed to receiving a placebo? This differential choice between placebo or no-treatment control is not done to assess whether anything causal happened in a study, which is the domain of internal validity. Similarly, when reason exists to think that pretreatment trends in an outcome may mimic treatment effects in the treatment group (e.g., participants were getting better by themselves anyway), adding pretest observations on the same measurement instrument at multiple prior time points helps diagnose this possibility. The use of multiple design features in one study can help to create a complex pattern of evidence that functions in a manner similar to Rosenbaum's discussion of creating a complex pattern of evidence by making design choices that stem from a broad theory.

The tradition within which we work has focused attention on creating a theory of quasi-experimentation that would constructively deal with issues like those just broached. This theory is not yet fully developed, and it may not be possible to formulate any theory of quasi-experimentation in a fully axiomatized way with the precision that many statisticians prefer and that has been approached for experiments with random assignment. However, we suspect that increased interchange between the design tradition we represent and the emergent statistical tradition Rosenbaum shares will improve the judgments he calls for in Section 1.2 about assessing the quality of evidence.

4. CAUSAL INFERENCE IS A FALLIBLE HUMAN JUDGMENT

Rosenbaum is clear, particularly in Section 4.1, that he does not intend to reduce design choices to statistics. His Section 1.2 emphasizes that scientists must take responsibility for judgments that inevitably go beyond statistics; indeed, when making these judgments he emphasizes a scientist's responsibility to both conscience and community. We want to make the same point even more broadly, acknowledging the less noble psychological influences contributing to the judgments scientists make, particularly in interpreting quasi-experiments. Judgments about causal inferences are all too human. In integrating evidence, logic, statistics and tacit knowledge of a study, there is much room for simple human biases to exert their influence—for example, the confirmation bias whereby all humans, scientists included, tend to recall evidence in favor of the inference they prefer and to overlook evidence against it (Cordray, 1986). Such judgments operate in randomized experiments where, before we can conclude that a treatment is effective, we have to judge such issues as whether any violations of normality were sufficient to question the validity of statistical tests, whether loss of some participants after random assignment was enough to vitiate group equivalence, whether certain generalizability-reducing features of the experimental context should be called to readers's attention or whether interactions between treatment and possible moderators should be sought out and reported or left unsaid. The role of human judgment in quasi-experiments is even greater and much more consequential, especially concerning the plausibility of those threats to internal validity that randomization would otherwise rule out.

5. THE EMPIRICAL PROGRAM OF QUASIEXPERIMENTS

A reasonable response to the preceding point is to claim that both design and statistics seek to reduce the role human judgment plays in causal inference. So, the solution is to develop better design choices and statistical methods. Randomization, for example, reduces the role of judgment about the plausibility of threats to internal validity (e.g., it ensures that selection bias between conditions is probabilistically minimized). We trust that statisticians will continue to develop such solutions from their repertoire; propensity scores and sensitivity analyses are apt examples. But given how much of causal inference is beyond statistics, we think it more likely that a theory of design rules for quasi-experiments will come mostly from empirical research more than statistical theory. By that, we mean evidence obtained from the empirical study of central design choices so as to examine how they help or hinder causal inference from quasi-experiments. We have already presented an extended discussion of what such an empirical program of research on quasi-experimentation might look like (Shadish, 1999), including (1) studies of the extent to which hypothesized threats to validity actually occur in practice (e.g., pretest sensitization, experimenter expectancy effects) and that identify those variables that moderate the influence of a threat and (2) studies of various strategies for improving effect estimates, including the use of "better" quasicontrol groups and improved methods for using matching or stratification strategies so as to avoid the past situations where matching proved to be misleading. A great deal of this empirical research has already been conducted. It includes the well-known efforts by some economists to compare estimates from randomized experiments to those from quasi-experiments that have been corrected using various selection bias models; it also includes metaanalytic and laboratory work comparing effect estimates from randomized and quasi-experiments to identify those design choices that improve estimates from the latter. As these examples suggest, the methodologies used to investigate such questions are eclectic, including surveys, literature reviews (quantitative and qualitative) and comparisons with randomized experiments. To the extent that a theory of design rules consists more of facts that are contingent (i.e., subject to revision given data) than logical (i.e., true by rational analysis), an empirical program of research is required to identify the "better" design rules that manifestly facilitate causal inference in quasi-experiments.

6. DESIGN RULES FOR CAUSAL GENERALIZATION

Rosenbaum's opening quote from Campbell touches on a fundamental problem in experimentation, how to generalize causal inferences from the highly constrained context of an experiment to the broader array of persons, treatments, observations, settings and times that are usually of both theoretical and policy interest. Some statisticians and economists write as though confident causal generalization depends on formal sampling procedures. For instance, Lavori, Louis, Bailar and Polansky (1986) suggest that experimenters should follow a two-step process: first randomly sample a subset of patients from the total population of patients, and then randomly assign the sampled patients to experimental or control conditions. Kish (1987) also prefers this two-step ideal, but he frankly acknowledges that this ideal is rarely feasible. It takes very little thought to see the difficulties of including in any one experiment a truly representative sample of persons, treatments, observations, settings and times (even if random samples of persons and settings would let themselves be assigned at random). Sometimes, informal sampling fallbacks are feasible, as when we deliberately sample instances that are heterogeneous in features thought to be related to an outcome, or when we sample the most frequently occurring instance to maximize potential relevance to a clear target of generalization.

Fortunately, there is more we can do to facilitate causal generalization (Cook, 1993, 1999; Shadish, Cook and Campbell, 1999). We should remember that the kind of generalization sampling promotes is not the only kind of generalization scientists do. For instance, we routinely generalize from the specific operations used in a study to the general terms we want to attach to the operations (construct validity). Such construct generalizations are sometimes given a formal statistical rationale, as with the domain sampling theory in classical test theory. More typically we justify generalizing from operations to constructs by bringing together a combination of quantitative and qualitative evidence based on five considerations:

- assessing the match between the operations achieved in a study (i.e., the units, treatments, observations, settings or times in the study) and the prototypical features of the constructs to which generalization is desired;
- demonstrating which experimental features are irrelevant to generalization in the sense that the causal relationship is replicated across all of them;
- demonstrating which specific features limit generalization in the sense that the causal relationship varies significantly from one to another;
- estimating how well the causal relationship can be interpolated (or extrapolated) over the range of some quantitative variable;
- identifying which parts of the treatment affect which parts of the outcome through which causal mediating processes, an explanatory process that should help identify those essential components of the cause that are required for transfer to other populations and settings;

These justifications for generalizing from operations to constructs also provide some justification for generalizing from operations that were studied to operations not included in the experiment's sampling frame. That is, they suggest a logic of generalization at the same kind of conceptual level as the logic for making descriptive causal inferences that underlies internal validity (i.e., treatment precedes outcome, alternative explanations ruled out etc.).

Hedges (1997) suggests some connections between these ideas and some pertinent statistical concepts. Those who prefer statistical terms might think of the first four of these five points above as a conceptual explication that parallels response surface modeling. However, our belief that causal inferences, including causal generalizations, are more logical than statistical leads us to doubt whether such statistical models will be a feasible solution to the problem of causal generalization on a regular and widespread basis in the next decades. In single studies, these five points suggest the importance to causal generalization of (1) thoroughly describing the target of generalization, (2) including in a study the prototypical features of the target that are thought to be crucial to successful generalization, (3) also including features thought to be irrelevant to generalization and (4) creating explanatory models that identify the essential elements that must be present for generalization to occur. Doing so allows partial tests of generalizability within single studies, for example, by examining how effect sizes vary with different types of person characteristics. This is not to say that causal generalization *should* be a primary goal of every study because doing so takes resources from other design choices that are frequently more important (e.g., including more observations to increase power). It is only to say that we can make design choices that improve such generalizations despite the fact that formal sampling procedures are rarely feasible.

However, assuming that generalization only occurs from *single* experiments is counterproductive. Few single studies have the large and heterogeneous samples of persons, settings, times, treatments and outcome measures that would be useful for confident causal generalization. By contrast, many systematic literature reviews include numerous studies with a clearly wider range of samples. They typically have used many more versions of a construct both as manipulations and outcome measures, addressed more diverse populations and been conducted in many more settings and have different times, both in terms of follow-up intervals and historical time periods. In addition, these studies will likely have used different methodologies, some stronger than others, allowing the researcher to examine how outcome differs as a function of methodological variables such as type of control group, reactivity of outcome measures or method of soliciting subjects for study. All this variation in substantive and methodological features allows the researcher to examine how well a cause–effect relationship generalizes and to identify causal contingencies influencing the size and direction of the relationship. We are fans of quantitative methods for research synthesis (metaanalysis) developed in the last two decades exactly because of their usefulness in addressing the causal generalization problem.

7. DESIGN RULES, NOT STATISTICS

Given what we have written, it should not be surprising that our only strong disagreement with Rosenbaum is with his last sentence, that the principles of design choice are properly viewed as part of the subject of statistics. No doubt, of course, statistics must be a major part of a theory of design. But statistics has not yet displayed the breadth of theory capable of incorporating all the complexities of causal inference in quasi-experiments that we describe. No single discipline has, of course; so no slight is intended to statistics. But for the present, an adequate theory of design rules for causal inference in quasi-experiments must be interdisciplinary. To force it into any of the molds currently available in statistics is to impoverish it and prevent it from incorporating exactly those nonstatistical features that are the most important part of causal inference. When it comes to causal inference from quasi-experiments, design rules, not statistics.

Rejoinder

Paul R. Rosenbaum

I thank the discussants for their comments. I am delighted that there are distinguished discussants from three fields and that there are areas of agreement, disagreement and new topics in their discussion.

1. MANSKI

1.1 Worst–Case Bounds on Treatment Effects

Although economists have proposed interesting methods for testing certain testable assumptions, either by contrasting estimates under alternative assumptions (Hausman, 1978) or using overidentified estimation equations (Hansen, 1982; Newey, 1985), much less attention has been paid to the consequences of assumptions that cannot be adequately tested. In light of this, Manski's worst-case bounds on treatment effects and his more general discussion of the information in underidentified models (Manski, 1995) are important contributions to the literature.

In the simplest cases, analogous worst-case bounds are obtained from a sensitivity analysis by letting the magnitude of the departure from a randomized experiment increase without bound, that is, by letting $\Gamma \to \infty$ in Section 4.3 of the paper; see Rosenbaum (1995b, Section 2.4) for technical details. So, worst-case bounds and sensitivity analyses are not logically incompatible. However, in practice they can leave very different impressions. As an example, imagine a study in which n subjects receive the treatment and all n survive, whereas an additional n subjects receive the control and all n die. If this pattern continued to hold as the sample size increased, $n \to \infty$, the sensitivity analysis would say that no fixed departure from a randomized experiment—that is, no finite Γ—could explain away such a strong association between treatment and survival. In contrast, the worst-case bounds would say that for every n it is possible that the treatment has no effect, because it might happen that every treated subject would have survived if assigned to control, and every control would have died if assigned to treatment; see Manski (1995, page 41). Again, these are logically compatible statements, but they leave very different practical impressions. For instance, the worst-case bounds would say it is logically possible that heavy smoking does not cause lung cancer, whereas the sensitivity analysis would say that extremely large hidden biases would need to be present to explain away the strong observed association.

In his book, Manski is not primarily interested in the simplest cases of his worst-case bounds, but rather in strategies for using information that might narrow the bounds. The simple, worst-case bounds then become a yardstick against which narrower bounds are measured to appraise the value of the added information. See, for example, the potential narrowing of the bounds when an instrumental variable is available that satisfies an exclusion restriction (Manski 1995, pages 36–37). Of course, an instrumental variable may be incorporated into a sensitivity analysis as well, yielding exact nonparametric inferences (Rosenbaum 1999a).

In his discussion, Manski says: "Where Rosenbaum and I differ is that I do not view the assumption of ignorable treatment selection to have a special status in observational studies of treatment effects." I am unsure whether we do differ here, because I am unsure about the meaning of the term "special status." When treatment assignment is ignorable, an appropriate estimate of a treatment effect is obtained from an analysis that compares people who looked comparable prior to treatment in terms of observed covariates. For instance, such an analysis might adjust for observed covariates by matching or covariance adjustment. In virtually all observational studies, an analysis that compares people who looked comparable prior to treatment is one of the several analyses presented, and comparisons of this sort may be found in almost any issue of, say, the *American Economic Review*—a high-quality empirical journal. At the same time, in most observational studies, departures from ignorable assignment are quite plausible, at times even likely, and sensitivity analyses are an aid in clarifying the consequences of such departures. While departures from ignorable assignment (i.e., departures from $\Gamma = 1$) are plausible, I do not think the worst case, $\Gamma \to \infty$, is especially plausible. The interesting value of Γ, it seems to me, is the value where the qualitative conclusions of a particular study begin to change. It is this value of Γ, varying from one study to the next, that we need to think about, because our judgments about it determine what we conclude from a particular study.

1.2 Treatment Effects That Vary from Person to Person

Manski argues that we need representative samples from populations because a treatment may have heterogeneous effects in a heterogeneous population. As is clear from the paper, I agree with half of this. If a treatment has substantially different effects on different types of people, then we need to discover this; see, for instance, the discussion of effects that vary with $\mathbf{x}_\omega$ in Section 4.4 of the paper and the discussion of dilated effects in Rosenbaum (1999b). However, having a few people from Philadelphia and a few more from Chicago, as might happen in a national survey, does not ensure that the relevant kinds of heterogeneity are present with sufficient frequency. If what matters for the effectiveness of a treatment is a person's ejection fraction, human capital, HIV status, attributional style or addiction to cocaine, then the entire study could be conducted in Philadelphia—or in Chicago. For example, randomized clinical trials routinely look for heterogeneous treatment effects, and they do this without representative samples from national populations (Piantadosi, 1997, Section 13). On the other hand, if what matters for the effectiveness of a treatment is a person's language, religion, culture, attitude toward the curative powers of magic or whether sexual transmission of the AIDS virus is primarily homosexual or heterosexual, then Philadelphia and Chicago are far too similar—and a study confined to the U.S. population is far too provincial. The sort of heterogeneity we need depends on the specific competing theories (Section 3.4) that we intend to contrast.

A survey may be representative of a population at a particular moment and yet be unrepresentative of nature. For instance, a survey of HIV prevalence in the United States in 1999 is misleading as a description of the AIDS epidemic and is inadequate even as a basis for public policy in the United States. Stanton (1997) claims 66% of the global HIV burden is in Africa and writes: "In many African cities, seroprevalence rates exceed 30% among women aged 20 to 24 years." In China, also, the spread of AIDS follows a pattern unlike the United States; see Yu et al. (1996). Scientific theories make broad claims about the past, present and future of the natural world, claims that can be refuted but not sampled. "Pure observation lends only negative evidence, by refuting an observation categorical that a proposed theory implies" (Quine, 1992, page 13).

1.3 Generalizing versus Theorizing

Manski raises the issue of internal and external validity. My view of this distinction is contained in Sections 3.1 and 4.4 of the paper and it differs from Manski's description of my view. For clarity, I briefly restate my view here.

I do not think scientists in any field start with a theory for a small population and then generalize it to a large population. I think we start with a general theory and scrutinize it in particular circumstances. The scrutiny may alter, narrow, refute or corroborate the theory. Popper (1990, pages 6–7) speaks of "bold, adventurous theorizing, followed by exposure to severe testing" and asserts:

> Our theories, our hypotheses, are our adventurous trials. Admittedly, most of them turn out to be errors: under the impact of our tests their falsity may be revealed. Those theories we cannot refute by the severest tests, we *hope* to be true. And, indeed, they *may be true*; but new tests may still falsify them.

A broad theory that is constantly narrowed by successive empirical studies is described by Lakatos (1970) as "degenerating."

An economist in Cleveland notices an important but unstudied aspect of behavior. She formulates a theory of this sort of behavior in terms of a rational choice to optimize expected utility. Living in Cleveland herself, she begins to collect data in Cleveland. Later a grant permits study sites all over Ohio. Another grant supports examination of data from the Current Population Survey of the entire United States. Still later, parallel studies are conducted in Singapore. But the theory was never, not even for an instant, a theory about optimizing behavior in Cleveland. From the first, it was a theory about thought, action and behavior. The theory was subjected to severe scrutiny in one circumstance or another, and perhaps the theory in its full generality was corroborated or perhaps refuted; if refuted, perhaps the theory became more focused. The scrutiny was broadened and intensified, but the theory was not generalized from a small population of people to a larger one.

Consider again the two studies of the psychological consequences of bereavement by Lehman, Wortman and Williams (1987) and Lichtenstein et al. (1996). Both studies contrasted the same two theories, one theory claiming that bereavement has only short-term effects on mental functioning and the other that it has long-term effects. The first study looked at sudden deaths of close relatives in car crashes and the second at the loss of a spouse by one of a pair of twins. Both studies are deliberately unrepresentative: car crashes are an atypical cause of bereavement, and twins are also atypical. If one wanted to generalize from a study population to a natural population, one would look for the typical, not the atypical, situation. But, for reasons discussed in the paper, car crashes and twins provide a sharper contrast of the two theories than typical situations permit.

2. ROBINS

Robins reasons about observational studies with time-varying treatments with reference to an analogous experiment, depicted in the simplest case in his Figure 11, in his Section 3. Although in his discussion he emphasized the experiment, a recent paper (Keiding et al., 1999, Section 3) discusses the role of such an experiment as an aid to thinking about observational studies. In these experimental analogs, at each node in the tree, a treatment branch is selected with probabilities that depend only on quantities observed prior to that node. These quantities may be baseline covariates, previous treatment decisions and outcomes observed prior to the node. Within this framework, Robins develops models for treatment effects and devices for reweighting observed outcomes to estimate the consequences of treatment sequences other than those actually imposed. It is an attractive, sensible approach to shedding light on a challenging class of problems.

One might also consider other analogous experiments for time-varying treatments. For instance, in a discussion of early studies of heart transplantation, in which patients must wait for a heart to become available, Gail (1972) imagines a different type of experiment:

> If one were interested in assessing survival from the time of transplantation (rather than from the time of admission as a transplant candidate), one could randomly assign an available heart to one of a pair of recipients (perhaps the pair that had waited longest). Then the survival of the two groups, transplant and control subjects, could be measured from the time of operation.

Here, some subjects are permanently assigned to control and are not at constant risk of receiving the treatment. Perhaps that would be a useful analog of an observational study in which some subjects categorically reject the treatment. For instance, if one were studying the effects of cocaine use, most people categorically reject the treatment and might possibly be viewed as permanent controls. Yet another alternative experimental analog has a single experimental assignment at the root of the tree, with subsequent treatment decisions being viewed as part of an adaptive treatment regime to which a person is assigned. Robins mentions Maclure's (1991) case-crossover design, which is yet another very interesting class of observational study designs with experimental analogs. With all attempts to reason about observational studies using analogous experiments, my own view is that one needs to entertain the possibility of departures from the imagined random assignment using some form of sensitivity analysis. Mittleman et al. (1997) mention the possibility of sensitivity analyses in case-crossover studies.

3. SHADISH AND COOK

The literature on quasi-experimentation, to which Shadish and Cook have made extensive contributions, is a cornerstone of the interdisciplinary literature on observational studies. Campbell (1957) insisted on a certain logic, a certain standard for criticism, namely, that objections to an observational study be expressed as specific, credible threats to validity, or "grounds for doubt" in Wittgenstein's phrase (1972, page 18). Anticipating specific grounds for doubt, the quasi-experimentalist improves the design to address these specific issues. Instead of assuming that hidden biases are absent, that what is not visible is equal—the so-called ceretis paribus clause—the quasi-experimentalist assumes specific biases may be present and investigates them. This is similar to advice offered by Lakatos (1970, page 110):

> How can one test a *ceteris paribus* clause severely? By assuming that there *are* other influencing factors, by specifying such factors, and by testing these specific assumptions. If many of them are

> refuted, the *ceteris paribus* clause will be regarded as well-corroborated.

For instance, in the study of bereavement following car crashes, if one objected to the absence of measures of depression prior to the crash, then one might use a design that addresses this specific concern; see Toedter, Lasker and Campbell (1990) for discussion of such a design. Specific threats to validity lead to specific improvements in design.

Quasi-experimentalists emphasize structured designs, assembled from components, to address specific threats to validity. The designs somewhat resemble designs used in the statistical theory of experimental design. This aspect of quasi-experimentation has not yet diffused into other fields that conduct observational studies, but it deserves attention as a design option in all fields. I look forward to the continuation of this important tradition in the forthcoming book by Shadish, Cook and Campbell (1999).

Once in a while, one sees the issues that arise in quasi-experimentation described in terms of dichotomies determined with certainty. Either a specific threat to validity is present or it is not. If a corresponding design feature is in place, then one is completely protected from this threat to validity, but if it is not, then the study is invalid. If a design is described in this way, then two questions go unasked, hence unanswered. First, does the design feature have the ability, given the sample size and variability, to address the specific threat to validity if it is present in a magnitude sufficient to affect the study? When the design features are the use of multiple control groups or unaffected outcomes (i.e., nonequivalent dependent variables), this question can be answered in conventional statistical terms such as the unbiasedness and monotonicity of the power function of tests for hidden bias, and the impact of these design features on confidence intervals that allow for hidden bias (Rosenbaum 1987b, 1989 a, b, 1992). Second, if a hidden bias is present or is suspected, is it of a magnitude sufficient to alter the conclusions of the study? One may be unable to rule out hidden bias, but biases of plausible size may or may not be able to account for the ostensible effects of the treatment. For instance, smokers and nonsmokers are known and suspected to differ in many ways not controlled in observational studies, and yet biases of plausible size cannot account for the extremely strong association between heavy smoking and lung cancer; see Cornfield et al., (1959) for a sensitivity analysis. The answers to these two questions seem to me to be part of the answer to the question: "How are we to judge the quality of evidence in quasi-experimentation?"

ADDITIONAL REFERENCES

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *J. Amer. Statist. Assoc.* **92** 1171–1177.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin* **54** 297–312.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. In *Understanding Causes and Generalizing About Them* (L. Sechrest and A. G. Scott, eds.) 39–82. Jossey-Bass, San Francisco.

Cook, T. D. (1999). Towards a practical theory of external validity. In *Validity and Social Experiments: Donald Campbell's Legacy* **1** (L. Bickman, ed.) Sage, Thousand Oaks, CA. To appear.

Cook, T. D. and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Rand-McNally, Chicago.

Cordray, D. W. (1986). Quasi-experimental analysis: a mixture of methods and judgment. In *Advances in Quasi-Experimental Design and Analysis* (W. M. K. Trochim, ed.) 9–27. Jossey-Bass, San Francisco.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. and Wynder, E. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* **22** 173–203.

Corrin, W. J. and Cook, T. D. (1998). Design elements of quasi-experimentation. *Advances in Educational Productivity* **7** 35–57.

Gail, M. (1972). Does cardiac transplantation prolong life? A reassessment. *Annals of Internal Medicine* **76** 815–817.

Geiger, D., Verma, T. and Pearl, J. (1990). The logic of influence diagrams (with comments). *Influence Diagrams, Belief Nets and Decision Analysis* 67–87.

Greenland, S., Pearl, J. and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* 10 37–48.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica* **78** 1251–1272.

Heckman, J. J. and Todd, P. E. (1996). Assessing the performance of alternative estimators of program impacts: a study of adult men and women in JTPA. Unpublished manuscript. (Available from the author, Dept. Economics, Univ. Chicago).

Hedges, L. V. (1997). The role of construct validity in causal generalization: the concept of total causal inference error. In *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences* (V. R. McKim and S. P. Turner, eds.) 325–341. Univ. Notre Dame Press.

Keiding, N., Filiberti, M., Esbjerg, S., Robins, J. M. and Jacobsen, N. (1999). The graft versus leukemia effect after bone marrow transplantation: a case study using structural nested failure time models. *Biometrics* **55** 23–28.

Kish, L. (1987). *Statistical Design for Research.* Wiley, New York.

Lavori, P. W., Louis, T. A., Bailar, J. C. and Polansky, H. (1986). Designs for experiments: parallel comparisons of treatment. In *Medical Uses of Statistics* (J. C. Bailar and F. Mosteller, eds.). New England Journal of Medicine, Waltham, MA.

Light, R. J., Singer, J. D. and Willett, J. B. (1990). *By Design: Planning Research in Higher Education.* Harvard Univ. Press.

MACLURE, M. (1991). The case-crossover design: a method for studying transient effects on risk of acute events. *American Journal of Epidemiology* **133** 144–153.

MANSKI, C. (1994). The selection problem. In *Advances in Econometrics, Sixth World Congress* (C. Sims, ed.) 143–170. Cambridge Univ. Press

MANSKI, C. (1996). Learning about treatment effects from experiments with random assignment of treatments. *Journal of Human Resources* **31** 707–733.

MANSKI, C. (1997a). Monotone treatment response. *Econometrica* **65** 1311–1334.

MANSKI, C. (1997b). The mixing problem in programme evaluation. *Rev. Econom. Stud.* **64** 537–553.

MANSKI, C. (1998). Treatment choice in heterogeneous populations using experiments without covariate data," In *Uncertainty in Artificial Intelligence, Proceedings of the Fourteenth Conference* (G. Cooper and S. Moral, eds.) 379–385. Morgan Kaufmann, San Francisco.

MANSKI, C. (1999). Identification problems and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice. *J. Econometrics.* To appear.

MANSKI, C. and NAGIN, D. (1998). Bounding disagreements about treatment effects: a case study of sentencing and recidivism," *Sociological Methodology* **28** 99–137.

MANSKI, C. and PEPPER, J. (2000). Monotone instrumental variables: with an application to the returns to schooling. *Econometrica*. To appear.

MANSKI, C., SANDEFUR, G., MCLANAHAN, S. and POWERS, D. (1992). Alternative estimates of the effect of family structure during adolescence on high school graduation. *J. Amer. Statist. Assoc.* **87** 25–37.

MITTLEMAN, M. A., MALDONADO, G., GERBERICH, S. G., SMITH, G. S. and SOROCK, G. S. (1997). Aternative approaches to analytical designs in occupational injury epidemiology. *American Journal of Industrial Medicine* **32** 129–141.

NEWEY, W. K. (1985). Generalized method of moments specification testing. *Journal of Econometrics* **29** 229–256.

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco.

PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–688.

PEARL, J. and ROBINS, J. M. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty in Artificial Intelligence. Proceedings of the 11th Conference on Artificial Intelligence* 444–453. Morgan Kaufmann, San Francisco.

PEARL, J. and VERMA, T. (1991). A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference* (J. A. Allen, R. Fikes and E. Sandewall, eds.) 441–452. Morgan Kaufmann, San Francisco.

PIANTADOSI, S. (1997). *Clinical Trials: A Methodologic Perspective*. Wiley, New York.

POPPER, K. R. (1990). *A World of Propensities*. Thoemmes, Bristol, UK.

QUINE, W. (1992). *Pursuit of Truth*. Harvard Univ. Press.

ROBINS, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Mathematical Modelling* **7** 1393–1512.

ROBINS, J. M. (1987). Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect." *Computers and Mathematics with Applications* **14** 923–945.

ROBINS, J. (1989a). The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS* (L. Sechrest, H. Freeman, and A. Mulley, eds.). NCHSR, U.S. Public Health Service.

ROBINS, J. M. (1994). Confounding and DAGS. Technical report, Dept. Epidemiology, Harvard School of Public Health.

ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statist..* **120** 69–117. Springer, New York.

ROBINS J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology* (E. Halloran, ed.) Springer, New York. To appear.

ROBINS, J. and GREENLAND, S. (1996). Comment on Angrist, Imbens, and Rubin's "Identification of causal effects using instrumental variables." *J. Amer. Statist. Assoc.* **91** 456–458.

ROSENBAUM, P. R. (1987a). The role of a second control group in an observational study (with discussion). *Statist. Sci.* **2** 292–316.

ROSENBAUM, P. R. (1989a). On permutation tests for hidden biases in observational studies. *Ann. Statist.* **17** 643–653.

ROSENBAUM, P. R. (1989b). The role of known effects in observational studies. *Biometrics* **45** 557–569.

ROSENBAUM, P. R. (1992). Detecting bias with confidence in observational studies. *Biometrika* **79** 367–374.

ROSENBAUM, P. R. (1995b). Quantiles in nonrandom samples and observational studies. *J. Amer. Statist. Assoc.* **90** 1424–1431.

SACKETT, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases* **32** 51–63.

SALZBERG, A. (1999). Removable selection bias in quasiexperiments. *Amer. Statist.* **53** 103–107.

SHADISH, W. R. (1999). The empirical program of quasiexperimentation. In *Validity and Social Experimentation: Donald Campbell's Legacy* (L. Bickman, ed.). Sage, Thousand Oaks, CA.

SHADISH, W. R., COOK, T. D. and CAMPBELL, D. T. (1999). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton-Mifflin, Boston. To appear.

SHADISH, W. R., COOK, T. D. and LEVITON, L. C. (1995). *Foundations of Program Evaluation*. Sage, Thousand Oaks, CA.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation, Prediction, and Search*. Springer, New York.

STANTON, B. (1997). Editorial: good news for everyone? *American Journal of Public Health* **87** 1917–1919.

TRIPLETT, N. (1898). The dynamogenic factors in pacemaking and competition. *American Journal of Psychology* **9** 507–533.

TOEDTER, L. J., LASKER, J. N. and CAMPBELL, D. T. (1990). The comparison group problem in bereavement studies and the retrospective pretest. *Evaluation Review* **14** 75–90.

WITTGENSTEIN, L. (1972). *On Certainty*. Harper and Row, New York.

YU, E., XIE, Q., ZHANG, K., LU, P. and CHAN, L. (1996). HIV infection and AIDS in China. *American Journal of Public Health* **68** 1116–1122.