

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,

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
	$OR=9$				$OR=9$

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 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

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

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

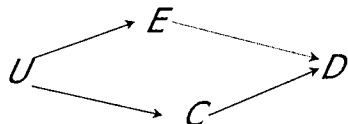


FIG. 2.

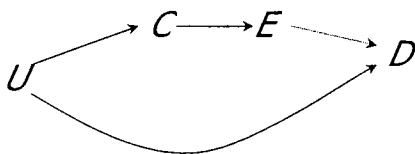


FIG. 3.

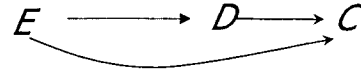


FIG. 4.

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.

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  $C$  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

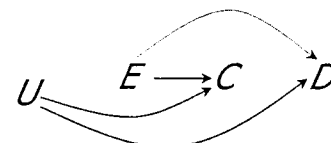


FIG. 5.

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 condi-

tional 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.

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  $OR_{E \cdot D|E} = 0.6$  represents not a protective

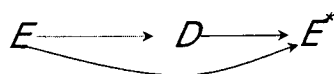


FIG. 6.

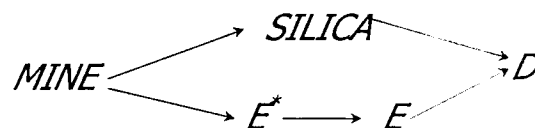


FIG. 7.

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 inde-

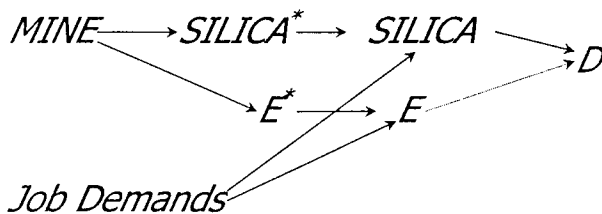


FIG. 8.

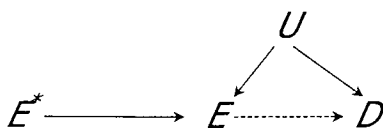


FIG. 9.

pendent. 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  $D$ .

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 epi-

demologists. 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:

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

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

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

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

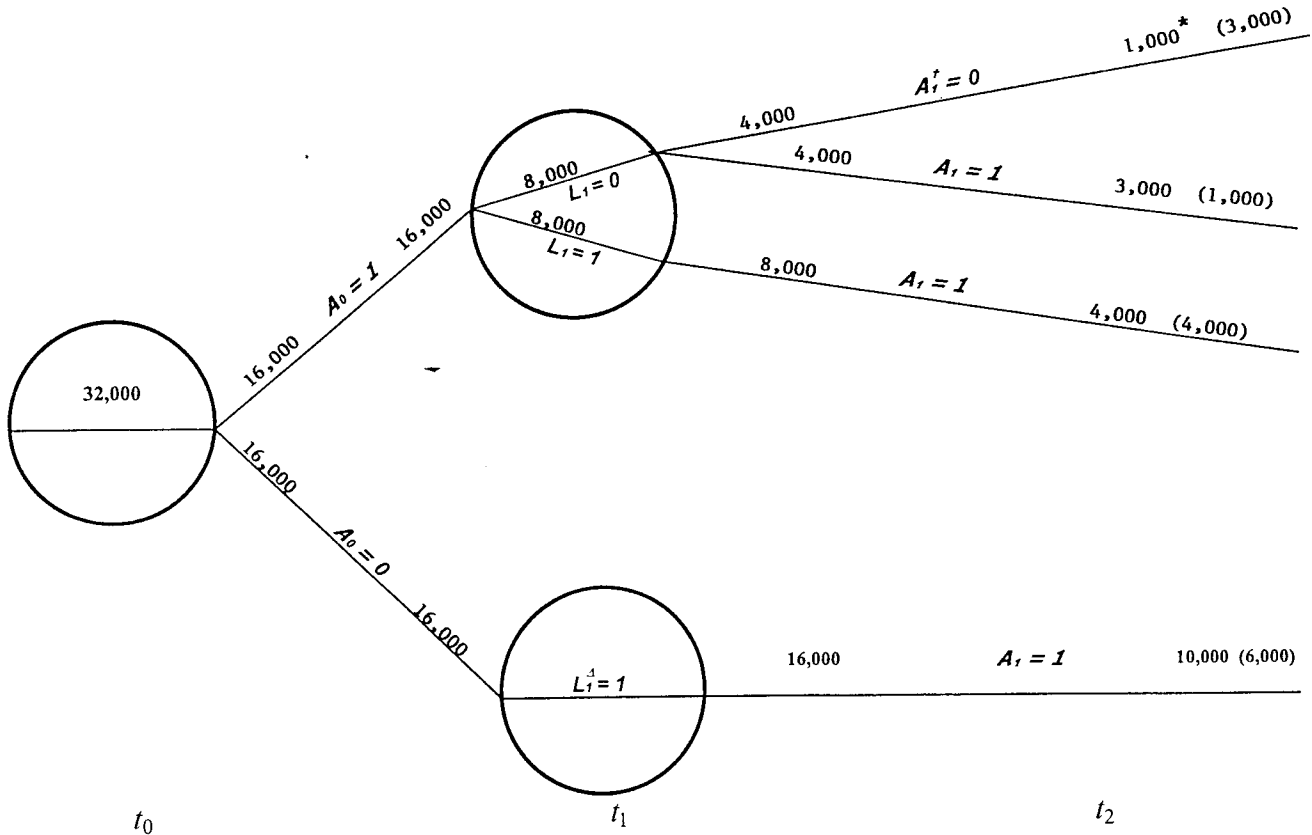


FIG. 10. Data from a hypothetical study: (\*) survivors ( $Y = 1$ ) at  $t_2$  [Deaths ( $Y = 0$ ) at  $t_2$  in parentheses]; (†)  $A_k$  measured just after time  $t_k$ ,  $k = 0, 1$ ; ( $\Delta$ )  $L_1$  measured at time  $t_1$ .

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:

$$\begin{aligned}
 (3.3) \quad & 0.5 = P[Y = 1 | L_1 = 1, A_0 = 1, A_1 = 1] \\
 & \neq P[Y = 1 | L_1 = 0, A_0 = 1, A_1 = 1] \\
 & = 0.75
 \end{aligned}$$

and

$$\begin{aligned}
 (3.4) \quad & 1 = P[A_1 = 1 | L_1 = 1, A_0 = 1] \\
 & \neq P[A_1 = 1 | L_1 = 0, A_0 = 1] = 0.5.
 \end{aligned}$$

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

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

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,

$$\begin{aligned}
 (3.6) \quad & 0.5 = P[L_1 = 1 | A_0 = 1] \\
 & \neq P[L_1 = 1 | A_0 = 0] = 1.
 \end{aligned}$$

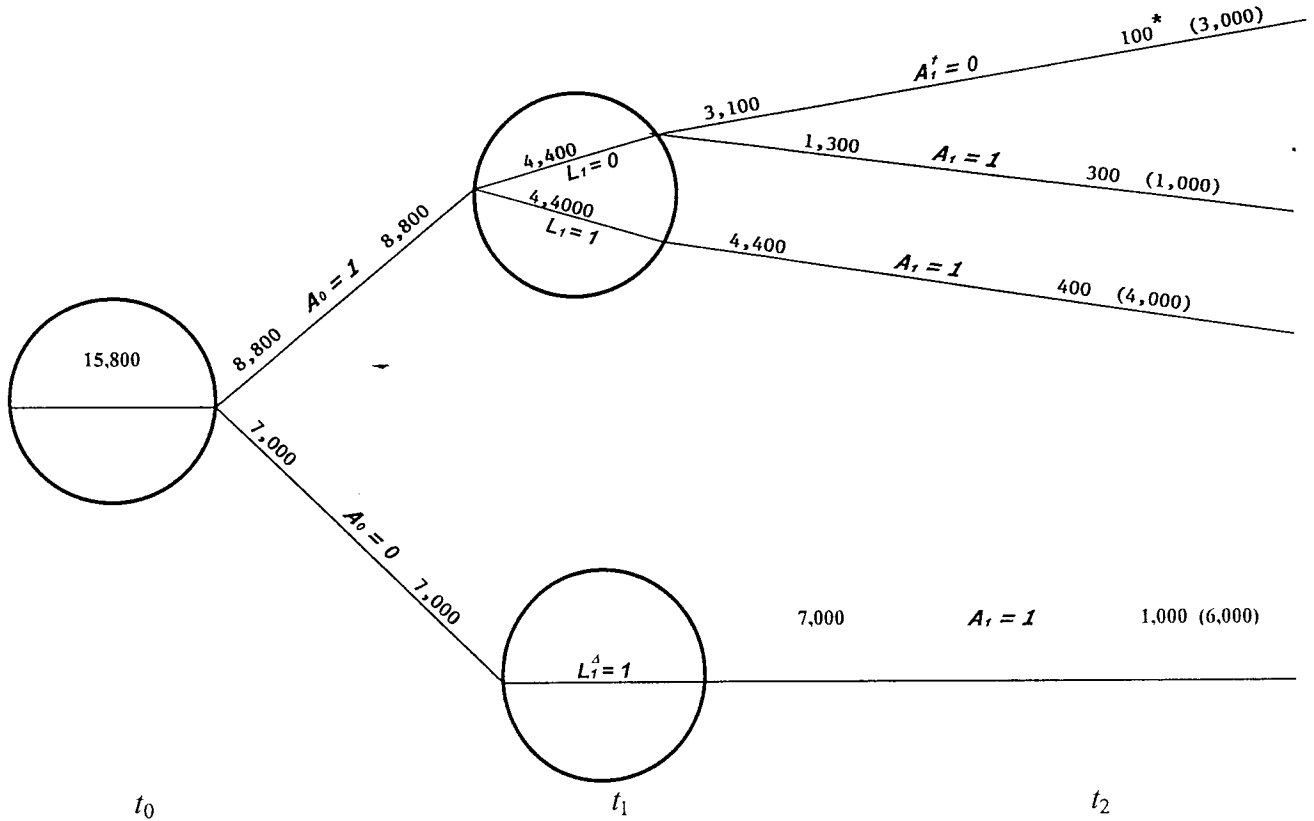


FIG. 11. Data from a hypothetical study: (\*) survivors ( $Y = 1$ ) at  $t_2$  [Deaths ( $Y = 0$ ) at  $t_2$  in parentheses]; (†)  $A_k$  measured just after time  $t_k$ ,  $k = 0, 1$ ; ( $\Delta$ )  $L_1$  measured at time  $t_1$ .

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 appropri-

ately 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  $\bar{A}_K = (A_0, A_1, \dots, A_K)$  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,  $\bar{L}_{K+1} = (L_0, \dots, L_{K+1})$  is the vector of all nontreatment variables. For notational convenience, define  $\bar{A}_{-1}, \bar{L}_{-1}, \bar{V}_{-1}$  to be identically 0 for all subjects. We view  $\bar{A}_K$  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  $\bar{a} = \bar{a}_K = (a_0, \dots, a_K)$  denote possible realizations of the random vector  $\bar{A}_K$ . Let  $Y(\bar{a})$  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  $\bar{L}_k, \bar{A}_{k-1} = \bar{a}_{k-1}$ :

$$(3.7) \quad Y(\bar{a}) \perp\!\!\!\perp A_k \mid \bar{L}_k, \bar{A}_{k-1} = \bar{a}_{k-1},$$

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*

$$(3.8) \quad \begin{aligned} & f_{Y(\bar{a})}(y \mid \bar{L}_k = \bar{\ell}_k, \bar{A}_{k-1} = \bar{a}_{k-1}) \\ &= \int \cdots \int f(y \mid \bar{\ell}_K, \bar{a}_K) \\ & \quad \cdot \prod_{j=k+1}^K f(\ell_j \mid \bar{\ell}_{j-1}, \bar{a}_{j-1}) d\mu(\ell_j), \end{aligned}$$

$$(3.9) \quad \begin{aligned} & f_{Y(\bar{a})}(y) = \int \cdots \int f(y \mid \bar{\ell}_K, \bar{a}_K) \\ & \quad \cdot \prod_{j=0}^K f(\ell_j \mid \bar{\ell}_{j-1}, \bar{a}_{j-1}) d\mu(\ell_j), \end{aligned}$$

where  $\mu$  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),

$$\begin{aligned} & \sum_{\bar{\ell}_1} f(y = 1 \mid \bar{\ell}_1, \bar{a}_1) f(\bar{\ell}_1 \mid \ell_0, \bar{a}_0) \\ &= f(y = 1 \mid \ell_1 = 1, a_1 = 1, a_0 = 1) \\ & \quad \cdot f(\ell_1 = 1 \mid a_0 = 1) \\ & \quad + f(y = 1 \mid \ell_1 = 0, a_0 = 1, a_1 = 1) \\ & \quad \cdot f(\ell_1 = 0 \mid a_0 = 1) \\ &= \left(\frac{4,000}{8,000}\right) \left(\frac{8,000}{16,000}\right) \\ & \quad + \left(\frac{3,000}{4,000}\right) \left(\frac{8,000}{16,000}\right) \\ &= \frac{10,000}{16,000}. \end{aligned}$$

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  $k$ ,*

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

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  $\hat{\beta}$  of the parameter  $\beta$  of the logistic regression model  $\log \text{it } P(D = 1 \mid A_0, A_1) = \beta_0 + \beta_1 A_0 + \beta_2 A_1 + \beta_3 A_0 A_1$  by fitting our case-control data by weighted logistic regression with subject-specific weights  $W$ , then  $\exp(\hat{\beta}_1 a_0 + \hat{\beta}_2 a_1 + \hat{\beta}_3 a_0 a_1)$  will, under (3.7) with  $L_0 \equiv 0$ , converge to  $OR_{\text{causal}}(a_0, a_1)$ . In particular  $\hat{\beta}_1$  and  $\hat{\beta}_3$  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  $\hat{W}^* = \hat{w}^*(A_1, L_1, A_0)$ .

### 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

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

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.,  $X \perp\!\!\!\perp Y \mid Z$ ) if  $X$  and  $Y$  are d-separated by  $Z$  on  $G$  (written  $(X \perp\!\!\!\perp Y \mid Z)_G$ ). 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  $(X \perp\!\!\!\perp Y \mid Z)_G$ , 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  $V_m(\bar{v}_{m-1})$  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  $V_m(\bar{v}_{m-1}) = V_m(pa_m)$  depends on  $\bar{v}_{m-1}$  only through  $V_m$ 's parents on  $G$ ;

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

$$\{V_{m+1}(\bar{V}_{m-1}, v_m), \dots, V_M(\bar{V}_{m-1}, v_m, \dots, v_{M-1})\} \\ \perp\!\!\!\perp V_m \mid \bar{V}_{m-1}.$$

**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 non-parametric 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  $V = (\bar{A}_K, \bar{L}_{K+1})$

with  $\bar{A}_K, \bar{L}_{K+1}$  as defined in Section 3.1.2, so that the variables in  $\bar{A}_k$  and  $\bar{L}_{k+1}$  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  $Y(\bar{a})$  be well defined, so we do not need to think of the nontreatment variables  $\bar{L}_{K+1}$  as manipulable.

Suppose we believe (3.7) holds but that a subset  $\bar{U}_{K+1}$  of the variables  $\bar{L}_{K+1}$  is not observed. The observed subset  $\bar{O}_{K+1}$  of  $\bar{L}_{K+1}$  includes the outcome variable  $Y = L_{K+1}$ . Define  $\bar{U}_k = \bar{L}_k \cap \bar{U}_{K+1}$  and  $\bar{O}_k = \bar{L}_k \cap \bar{O}_{K+1}$  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  $V = (\bar{L}_{K+1}, \bar{A}_K) \equiv (\bar{L}, \bar{A})$  such that (3.7) will imply that

$$(A.2) \quad Y(\bar{a}) \perp\!\!\!\perp A_k \mid \bar{O}_k, \bar{A}_{k-1} = \bar{a}_{k-1}$$

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

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  $G_k^{\bar{a}}$  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*

$$(A.3) \quad \left( Y \perp\!\!\!\perp A_k \mid \bar{O}_k, \bar{A}_{k-1} \right)_{G_k^{\bar{a}}}, \quad k \leq K,$$

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

Here  $(A \perp\!\!\!\perp B \mid C)_{G_k^{\bar{a}}}$  stands for d-separation of  $A$  and  $B$  given  $C$  in  $G_k^{\bar{a}}$  (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  $\bar{O}_0 = E^*$ , because (A.2) implies, by (3.8) with  $o$  substituted for  $\ell$ , that  $OR_{ED|E^*}$  equals the conditional causal odds ratio  $OR_{causal, ED|E^*}$  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  $G^{\bar{a},k}$  be the DAG that differs from  $G_k^{\bar{a}}$  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)  $(U_{bk} \amalg A_k | \bar{O}_k, \bar{A}_{k-1})_{G^{\bar{a},k}}$  and (ii)  $(U_{ak} \amalg Y | \bar{O}_k, \bar{A}_k, U_{bk})_{G^{\bar{a},k}}$ .*

**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  $G^{\bar{a},k}$ ;*
- (ii) *variables  $U_{2k}$  that are  $d$ -separated from  $A_k$  given  $(\bar{O}_k, \bar{A}_{k-1})$  in  $G^{\bar{a},k}$ ;*
- (iii) *variables  $U_{3k}$  that are  $d$ -separated from  $Y$  in  $G^{\bar{a},k}$  given  $(\bar{A}_k, \bar{O}_k)$ ;*
- (iv)  *$U_{2k}$  and  $U_{3k}$  are  $d$ -separated from one another given  $(\bar{O}_k, \bar{A}_{k-1})$  in  $G^{\bar{a},k}$ .*

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  $V = (\bar{L}_{K+1}, \bar{A}_K)$  is represented by a given DAG  $G$  such that the variables in  $(\bar{A}_k, \bar{L}_{k+1})$  are nondescendants of  $A_{k+1}$  and the assumption (3.7) of no unmeasured confounders holds with  $Y \equiv \bar{L}_{K+1}$ . Then, when (A.3) is also assumed to hold, we say the  $U_k$  are nonconfounders for the effect of treatment  $\bar{A}_K$  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  $\ell$  will equal the causal effect  $f_{Y(\bar{a})}(y)$ , but we also cannot be certain that it will not equal  $f_{Y(\bar{a})}^{(Y)}$ . 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  $(\bar{O}_{K+1}, \bar{A}_K)$  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  $f_{Y(\bar{a})}^{(Y)}$  when (A.3) is false. The answer is no if on  $G$  there are no missing arrows out of the  $A_k$ , because then  $f_{Y(\bar{a})}(y)$  is not identified from the data  $(\bar{O}_{K+1}, \bar{A}_K)$ . That is, there will be distributions  $F_V$  represented by  $G$  which have the same marginal distribution for  $(\bar{O}_{K+1}, \bar{A}_K)$  but different values for  $f_{Y(\bar{a})}(y)$ . Hence, in practice, we must treat potential confounders as actual confounders, and interpret causal effects as uncomputable.