CHAPTER SIXTEEN

USING CAUSAL DIAGRAMS TO UNDERSTAND COMMON PROBLEMS IN SOCIAL EPIDEMIOLOGY

M. Maria Glymour

Epidemiologists typically seek to answer causal questions using statistical data: we observe a statistical association between poverty and early mortality and seek to determine whether poverty causes early death. An essential component of epidemiologic training is therefore learning what statistical relations imply, or do not imply, about causal relations. This is why the cliché "correlation does not imply causation" is the mantra of introductory epidemiology classes. But correlations, and other forms of statistical association, do give us information about causal relations, and this is why—despite the oft-repeated warnings—quantitative statistical analyses are the mainstay of epidemiology.

Diagrams are routinely used informally to express beliefs and hypotheses about relations among variables. These informal uses can be greatly expanded by adopting formal rules for drawing the diagrams so that they meet the criteria for causal Directed Acyclic Graphs (DAGs). Causal DAGs are a simple, flexible device for demonstrating the statistical associations implied by a given set of assumptions about the causal structure relating variables. Knowing this, we can also move in the other direction: given a set of statistical associations observed in the data, we can identify all of the causal structures that could have given rise to these associations. Learning the rules for reading off statistical associations from the causal assumptions represented in a DAG can take a little time and practice. Once mastered, though, these rules turn out to be extremely practical for a number of tasks (for example, choosing regression covariates, understanding selection bias,

interpreting tests of "direct" effects, or assessing natural experiments). Using DAGs makes it easier to recognize and avoid mistakes in these and a number of other analytic decisions. The rules linking causal relations to statistical associations are grounded in mathematics, and one way to think of the usefulness of causal diagrams is that they allow non-mathematicians to draw rigorous, mathematically based conclusions about certain types of statistical relations.

In this chapter, I first introduce some language and background assumptions; I then describe the rules for drawing causal DAGs and the associated rules linking the causal assumptions encoded in a DAG to the statistical relations implied by these structural assumptions; and finally, I discuss a few applications of DAGs within social epidemiology. Some readers may prefer to begin with the examples and refer back to the definitions and rules for DAGs as needed; however, the material described in the section on the d-separation rules is essential for following the examples. A number of excellent and more comprehensive introductions to DAGs, many written by the researchers who developed the ideas, are available elsewhere (Greenland et al. 1999; Pearl 2000; Robins 2001; Spirtes et al. 2000). My goal in this chapter is to provide a basic introduction to demonstrate the utility of DAGs for applied social epidemiology researchers.

Some Background Definitions

Causal inference is an important problem in many applied disciplines, and much of the work written on the topic has been addressed to readers in fields other than epidemiology. The writing on causal inference can sometimes be dense or technical. I will begin by explaining how I use key terms. Note that some of my uses (for example, the definition of *cause*) are controversial and I encourage the reader to see others who disagree. Debating the definitions is beyond the scope of this chapter, and little of the discussion of DAGs would be affected by adopting such alternative definitions.

Define X and Y as random variables. We say X **causes** Y if, had X taken a different value than it actually did—*and* nothing else temporally prior to or simultaneous with X differed—then Y would have taken a different value. To accommodate the possibility that causation is not deterministic, we can say that had X taken a different value, this would have resulted in a different probability distribution for Y.

It is invaluable to frame our research question in terms of a hypothetical intervention on X. For example, instead of asking, "Does income affect diabetes risk among Cherokee tribal members?" we ask, "Would sending each tribal member an annual check for $4,000 from the Cherokee Nation government change their diabetes risk?" The effect of such a check might differ from other ways of changing income—for example, increasing wages or providing in-kind donations or changing

tax rates—even if these approaches had identical net monetary value. Most importantly, referring to a hypothetical intervention distinguishes the causal question from related statistical questions, such as "Do high-income individuals have lower diabetes risk compared with low-income individuals?" The hypothetical intervention must directly affect only the exposure X, although other things might also change if they are consequences of exposure (Pearl 2000; Spirtes et al. 2000). For example, an intervention such as sending a check may affect diet as well as income, but only because recipients use the extra income to buy different foods. It need not be possible for the researcher to conduct the intervention; there must merely be some conceivable way that X could take a different value, even if by random assortment. The definition of "cause" is the topic of heated and extensive debate; see for example (Dawid 2000; Glymour 1986; Hernán 2004; Holland 1986a, 1986b; Kaufman and Cooper 1999, 2001; Parascandola and Weed 2001; Pearl 2000; Woodward 2003).

We say X and Y are **statistically independent** if knowing the value of X does not provide any information about the value of Y (if X is independent of Y, Y is also independent of X). Conversely, we say X and Y are **statistically dependent** if knowing the value of X gives us some information about the likely value of Y, even if this information is very limited and amounts to a modest change in the probability distribution of Y. If there is some value X of X that is informative about the probability distribution of Y, we say that X and Y are statistically dependent. Note that statistical dependency may be assessed with various statistical parameters, some of which depend on additional assumptions (for example, regression coefficients, odds ratios, t tests, chi-square tests, or correlation coefficients).

It is very helpful to distinguish between words that denote causal relations and words that denote statistical relations (Pearl 2001). "Cause," "influence," "change," "increase," "decrease," and "promote" are all examples of **causal language.** Association, prediction, and any specific measures of statistical association such as regression coefficients and so forth are examples of **statistical language.** When a statistical association is reported in an epidemiology article, it is generally with the hope (sometimes unstated) of using this to give insight into a causal relation. Surveillance reports and predictive (as opposed to etiologic) models are exceptions; in these cases, causal inference is not of primary interest.

If we examine the distribution of one variable, Y, *within* levels of a second variable X, we say that we are examining the distribution of Y **conditional** on X. Conditional relations are often denoted in equations with the symbol "|". For example, if $p(Y)$ denotes the probability distribution of Y, a formal definition of statistical independence is:

$$p(Y \mid X) = p(Y) \tag{1}$$

which would be read "the probability distribution of Y conditional on X equals the marginal (or unconditional) probability distribution of Y." In other words, knowing

the value of X does not give us information about the distribution of Y (for any value x of X).

Similarly, if we examine the relations between two variables within levels of a third variable—for example, the relation between income and mortality within levels of education—we say we are examining the conditional relation. Stratification, restriction, matching, and covariate adjustment in regression models are all statistical techniques that are special types of conditioning. If two variables X and Y are statistically independent without conditioning on any other variables, we say X and Y are **marginally independent.** If X and Y are independent, conditional on Z, then:

$$p(Y \mid X, Z) = p(Y \mid Z) \tag{2}$$

Although causal dependence and statistical dependence are not the same, they are related phenomena. To understand how causal and statistical relations are linked, note that statistical dependency between two variables X and Y could reflect any of five situations (or combinations of these):

1. Random fluctuation.
2. X caused Y.
3. Y caused X.
4. X and Y share a common cause[1]
5. The statistical association was induced by conditioning on a common effect of X and Y (as in selection bias).

The task epidemiologists typically face is to decide which of these explanations is consistent with our data and background knowledge and rule out all others. Often we are especially interested in demonstrating that X likely caused Y (perhaps because this may offer the best prospects for publication). Confidence intervals and p-values are used to assess the plausibility of the first explanation for a statistical association. Temporal order can rule out explanation three, and this is why longitudinal studies are advantageous for demonstrating causation. Ruling out common prior causes, explanation four, is the goal of most covariate

[1]Note that a variation on this situation is the possibility that the sample is composed of two subsamples, each of which has a different marginal probability distribution of X and of Y. In the combined population, X and Y may be statistically dependent even if they were independent in each of the subsamples. This is sometimes considered a sixth possibility to explain a statistical dependency between X and Y. I treat this as a special case of situation four, however. To frame it this way, consider subsample membership to be a variable that is a common cause of X and Y.

adjustment in regression models. Covariate stratification is also frequently, though not universally, motivated by the desire to eliminate the possibility of common prior causes. Explanation five—the association was induced by conditioning on a common effect—is confusing for many people, and it is perhaps for this reason that this possibility is often ignored. This phenomenon is crucial in many settings, though, so I will try to give an intuitive explanation here (it will come up again in the examples section of the chapter).

Why does conditioning on a common effect of two variables induce a statistical association between those variables? The easiest way to hold onto this idea is to find a simple anecdote that describes the phenomenon. For example, suppose you believe that two factors determine basketball prowess: height and speed. Exceptional players must be either extremely tall or extremely fast. If you examined everyone in the world, height and speed might be statistically independent. Short people are not necessarily fast, nor are tall people; however, if you look only at professional basketball players, you would confidently guess that the short ones are *very* fast. People without the advantage of height must compensate with lightening speed in order to become great ball players. By restricting to pro basketball players, you have conditioned on a common effect of height and speed, and within this stratum of pro ball players, height and speed are (inversely) associated. This is not a perfect association, because some of the tall players may also be fast. And it is also possible that speed and height are correlated in the general population. The point is merely that, whatever the association between speed and height in the general population, it is quite different among professional basketball players.

This phenomenon—the change in association between two variables when conditioning on their common effect—is sometimes called **collider bias** because the two causes "collide" at the common effect. It can be induced by sample selection, stratification, or covariate adjustment if some of the covariates are effects of the other independent variables (Hernán et al. 2004).

We say the association between X and Y is **confounded** if the statistical association between X and Y does not equal the causal relation between the two variables. For example, if X and Y are both influenced by Z, the crude (marginal) relation between X and Y is likely confounded, although the relation between X and Y conditional on Z may be unconfounded. If conditioning upon a set of covariates $\boldsymbol{Z}$ will render the association between X and Y unconfounded, then we say $\boldsymbol{Z}$ is a **sufficient** set of covariates for estimating the relation between X and Y. A sufficient set may be empty (if the crude relation between X and Y is unconfounded), or it may contain one or many variables. Furthermore, there may be several alternative sufficient sets for any pair of variables X and Y (Greenland and Robins 1986; Greenland et al. 1999).

Graphical Models

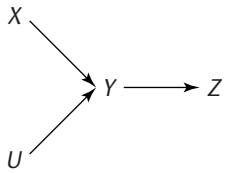
With this background and common language, we now turn to causal DAGs. First, I outline the rules for expressing causal assumptions in a DAG. Next, I explain the d-separation rules, which describe how to read from the DAG the set of statistical associations implied by the causal assumptions encoded in that DAG. Formal introductions to graphical models, explanations of how DAGs relate to conventional structural equation models, and proof of the mathematical equivalence between the rules we apply to DAGs and Robins' g-computation formula can be found elsewhere (Greenland et al. 1999; Pearl 2000; Robins 1987, 1995; Spirtes et al. 2000).

Drawing a Causal DAG[2]

Causal DAGs visually encode an investigator's a priori assumptions about causal relations among the exposure, outcomes, and covariates. In a causal DAG, we say that a variable X causes a variable Y *directly* (relative to the other variables in the DAG) if there is an arrow from X to Y or *indirectly* if there is a sequence of directed arrows that can be followed from X to Y via one or more intermediate variables. In Figure 16.1, X causes Y directly and Z indirectly. The **descendants** of a variable are the other variables in the DAG affected either directly or indirectly by that variable. If two variables shown in a DAG share a common cause, that common cause must also be included in the DAG or else the DAG is not considered "causal." It is not necessary to include all causes of individual variables in the DAG; only causes of two or more variables in the DAG must be included. If unknown or unmeasured common causes are assumed to exist, these should be represented in the diagram as unknown common causes with arrows to the variables that they are thought to affect. The absence of a sequence of directed arrows linking two variables in a DAG represents the assumption that there is no causal relation between the two variables. If a prior value of Y affects X, which affects a subsequent value of Y, these must each be shown as separate variables (for example, $Y_0 \rightarrow X_1 \rightarrow Y_2$). Directed acyclic graphs must not have any cycles between variables, consistent with the general intuition that if X causes Y, Y cannot also cause X at the same moment.

[2]This section, the following section on d-separation rules, and Figure 16.1 are taken substantially from Appendix Two in Glymour et al., "When is baseline adjustment useful in analyses of change? An example with education and cognitive change." *American Journal of Epidemiology,* 2005, 162(3) 267–278, by permission of Oxford University Press.

FIGURE 16.1. CAUSAL DIAGRAMS DEPICTING A VALID INSTRUMENT.



Causal assumptions represented in DAG 1:

- *X* and *U* are each direct causes of *Y* (direct with respect to other variables in the DAG).
- *Y* is a direct cause of *Z*.
- *X* is not a direct cause of *Z,* but *X* is an indirect cause of *Z* via *Y*.
- *X* is not a cause of *U* and *U* is not a cause of *X*.
- *U* is not a direct cause of *Z*, but *U* is an indirect cause of *Z* via *Y*.
- No two variables in the DAG (*X*, *U*, *Y*, or *Z*) share a prior cause not shown in the DAG, e.g., no variable causes both *X* and *Y,* or both *X* and *U*.

Statistical relations implied by the assumptions in the example causal DAG (note that this is not a comprehensive list of all the conditional relations and that the statistical dependencies listed here assume faithfulness):

- *X* and *Y* are statistically dependent.
- *U* and *Y* are statistically dependent.
- *Y* and *Z* are statistically dependent.
- *X* and *Z* are statistically dependent.
- *U* and *Z* are statistically dependent.
- *X* and *U* are statistically independent (the only path between them is blocked by the collider *Y*).
- *X* and *U* are statistically dependent, conditional on *Y* (conditioning on a collider unblocks the path).
- *X* and *U* are statistically dependent, conditional on *Z* (*Z* is a descendant of the collider *Y*).
- *X* and *Z* are statistically independent, conditional on *Y* (conditioning on Y blocks the path between *X* and *Z*).
- *U* and *Z* are statistically independent, conditional on *Y*.

The d-Separation Rules Linking Causal Assumptions to Statistical Independencies

After drawing a DAG to represent our causal assumptions, we can apply the d-separation rules to find the statistical relations implied by these assumptions. Before introducing the d-separation rules, I mention three assumptions adopted throughout the rest of the chapter. These assumptions are discussed in more detail at the conclusion of this section.

1. The Causal Markov Assumption (CMA): any variable X is independent of any other variable Y conditional on the direct causes of X, unless Y is an effect of X. The CMA is consistent with most accounts of causation and, although rarely stated in these words, is often implicitly invoked in applied research.
2. Faithfulness: positive and negative causal effects never *perfectly* offset one another; that is, if X affects Y through two pathways, one positive and one negative, the net statistical relation between X and Y will be either positive or negative. If the two paths perfectly offset one another, the net statistical association would be zero, in which case we say the statistical associations are **unfaithful** to the causal relations. Under faithfulness, we assume this situation never occurs.
3. Negligible randomness: statistical associations or lack of associations are not attributable to random variation or chance (that is, we assume a large sample size).

The DAG expresses a set of assumptions about the causal relations or absence of causal relations among the variables. If the assumptions of a causal DAG are correct, then two variables in the DAG will be statistically independent conditional on a set of covariates if every **path** between the two variables is **blocked.** What is a path and what does it mean to block it? A path is any sequence of lines (also called edges) connecting two variables *regardless of the direction of the arrowheads.* The direction of arrowheads is important to identify variables on a path that are **colliders.** If arrowheads from A and B both point to a variable C (as in: $A \rightarrow C \leftarrow B$), then C is referred to as a collider on that path between A and B: the causes collide at C. In other words, a collider is a common effect of two variables on the path (the collider itself must also be on the path). All other variables on a path are non-colliders. A path is blocked by conditioning on a proposed set of variables $\boldsymbol{Z}$ if either of two conditions holds:

1. One of the non-colliders on the path is in the set of variables $\boldsymbol{Z}$, or;
2. There is a collider on the path, and neither the collider nor any of the collider's descendants is in $\boldsymbol{Z}$.

These rules fit with the intuition that two variables will be correlated if one causes the other or there is an uncontrolled common prior cause of the two variables. The rules also reflect the fact that a statistical association between two variables can be induced by conditioning on a common effect of the two variables (Greenland et al. 1999; Hernán et al. 2002), as described in the pro basketball example. Note that if a collider on a path is in the proposed covariate set, this collider does not block the path. If a DAG contains no unblocked paths between A and B, the two variables will be marginally independent; that is, without conditioning on any other variables, A and B will be independent. If we assume faithfulness, two variables in a DAG will be statistically dependent

if there is an unblocked path between them. Rule (2) implies that conditioning on a variable may unblock a path between *A* and *B* and induce a correlation if that variable is a collider or a descendant of a collider on a path between *A* and *B*.

To make these ideas more concrete, consider the example DAG in Figure 16.1. This figure shows a causal DAG and lists the causal assumptions represented by that DAG and the statistical associations implied, under the d-separation rules, by those causal assumptions. For example, the assumptions encoded in the DAG imply that *X* and *U* are marginally independent but become statistically associated after conditioning on either *Y* or *Z*. In contrast, *X* and *Z* are marginally dependent but become statistically independent after conditioning on *Y*.

The Assumptions for Using Causal DAGs

Now we return to the assumptions we stated earlier: Causal Markov, Faithfulness, and negligible randomness. Why do we need these assumptions and should we accept them? The Causal Markov Assumption (CMA) is consistent with intuition: if we hold constant the factors that are direct causes of a variable *X*, then other factors will be independent of fluctuations in *X*, unless these other variables are themselves influenced by *X*. Imagine a string of dominos with letters from *A* to *Z* lined up in order. Flipping domino A will cause all of the downstream dominos to fall as well. You can interrupt the sequence of falling dominos by removing one in the middle (or holding it up so it doesn't fall). If you hold F up, then flipping E will not affect G or any subsequent domino; however, holding F will not interrupt the effect of flipping G on H or I (see Glymour 2001, pp. 21–27, for a more extensive but accessible discussion of the CMA). Standard epidemiologic reasoning often appeals to the CMA. For example, the injunction against conditioning on mediators if you wish to estimate the total effect of an exposure on the outcome implicitly relies on CMA.

The faithfulness assumption, that positive and negative effects never perfectly offset one another, is valuable because, formally, the d-separation rules define the statistical *independencies* implied by the assumptions in the DAG. Although the statistical independencies are interesting, we would often like to know about the statistical *dependencies*. These do not automatically follow from the d-separation rules, because two variables in a DAG might be statistically independent even though this independence is not implied by the causal structure. If two pathways with equal and opposite counterbalancing effects link two variables in a DAG, these two variables will be statistically independent despite their causal connection. To extend the d-separation rules to define the statistical dependences implied by a DAG, we must assume faithfulness. Some researchers contend that faithfulness is commonly violated in the real world. Nonetheless, the major implications from the examples in the rest of the chapter would stand if we did not assume faithfulness.

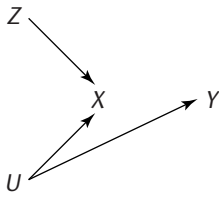
We assume negligible randomness, because DAGs give no information on whether statistical relations are likely to have arisen by chance due to random variation. To focus on DAGs we will assume that effects due to random variation can be ignored (for example, because you are looking at statistical associations in a very large sample). Without this assumption, the examples in the rest of the chapter would hold asymptotically.

These three assumptions should be clearly distinguished from the content-specific causal assumptions encoded in DAGs, which relate specifically to the substantive question at hand. By assuming CMA, faithfulness, and negligible randomness, we can link the causal assumptions in the DAG to probability statements about the variables. The CMA is fundamental for the d-separation rules. Faithfulness allows us to predict statistical associations instead of just statistical independencies. Negligible randomness lets us ignore random variations that would appear in small samples.

Applying DAGs to Answer Questions in Social Epidemiology

Why are DAGs useful? In general, we wish to test a hypothesis about how the world works within the context of our prior beliefs. This is linked, sometimes implicitly, to a desire to know what would happen if we intervened to change the value of some treatment or exposure. Directed acyclic graphs help us answer the question: under my prior assumptions, would the statistical analysis I am proposing provide a valid test of my causal hypothesis? Consider Figure 16.2 and imagine you are interested in testing whether X has a causal effect on Y (that is, you are unsure if there should be an arrow from X to Y). Other than this question, you believe the causal structure is as drawn in Figure 16.2. It is immediately evident from the DAG that the analysis must condition on U; U confounds the effect of X on Y. But suppose that you are interested in estimating the effect of Z on Y. In this case, you need not condition on U. The relation between Z and Y is unconfounded (as is the relation

FIGURE 16.2. CAUSAL DIAGRAMS DEPICTING INVALID INSTRUMENTS.



between Z and X). Directed acyclic graphs provide a way to state explicitly one's prior beliefs about causal relations or alternative sets of plausible prior assumptions. We base decisions such as selection of covariates on these priors, although the way in which priors shape these decisions is not always explicit.

We now turn to a number of examples in which DAGs can be used to clarify epidemiologic ideas. In some cases, the DAGs simply provide a convenient way to express well-understood concepts. In other examples, the DAGs illuminate a point of common confusion regarding the biases introduced by proposed analyses or study designs. In all these cases, the findings can be demonstrated mathematically or by using any number of informal arguments. The advantage of DAGs is that they provide a simple, common tool for understanding an array of different problems.

Why Conventional Rules for Confounding Are Not Reliable

Earlier, I defined confounding in terms of contrasting statistical and causal associations. A statistical association between two variables is confounded if it differs from the causal relation between the two variables. This definition implies graphical criteria for choosing a **sufficient set** of covariates, which is a set such that within strata of the covariates the statistical relation between exposure and outcome is unconfounded. That is, after specifying background causal assumptions using a DAG, we can identify from the DAG a sufficient set of covariates $\boldsymbol{Z}$ for estimating or testing for an effect of X on Y; $\mathbf{Z}$ is such a sufficient set if (1) no variable in $\mathbf{Z}$ is a descendant of X and (2) every path between X and Y that contains an arrow into X is blocked by $\mathbf{Z}$.

These rules are often called the "back-door" criteria, tapping the idea that paths with arrows into X are "back-doors" through which a spurious (non-causal) statistical association between X and Y might arise. When the back-door criteria are fulfilled by a set of measured covariates, it is possible to estimate the total average causal effect of X on Y. Under the graphical criteria, it is clear that there may be several alternative sufficient sets to control confounding. Thus, it is possible that a given variable is included in one sufficient set but not in another. A related point is that these rules do not define a "confounder" but instead describe when a conditional statistical association between two variables will be confounded (see Maldonado 2002 for a helpful discussion of this distinction). Detailed discussion of the graphical criteria can be found in Greenland et al. (1999) and Pearl (2000, p. 79).

How do the graphical criteria relate to conventional criteria for identifying confounders? In both intuition and application, the graphical and conventional

criteria overlap substantially. For example, Hennekens and Buring explain that confounding occurs when "an observed association . . . is in fact due to a mixing of effects between the exposure, the disease, and a third factor . . ." (Hennekens and Buring 1987, p. 35). Rothman and Greenland describe confounding as "a distortion in the estimated exposure effect that results from differences in risk between the exposed and unexposed that are not due to exposure." (Rothman and Greenland 1998, p. 255) The intuitions are similar.

Variations on the following specific criteria for identifying confounders are frequently suggested, although it is often noted that these criteria do not "define" a confounder:

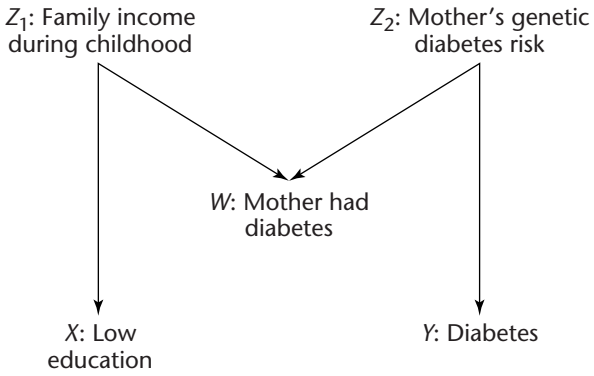
1. A confounder must be associated with the exposure under study in the source population.[3]
2. A confounder must be a risk factor for the outcome, though it need not actually cause the outcome.
3. The confounding factor must not be affected by the exposure or the outcome.

These rules are based on statistical associations, and we will refer to them as the conventional statistical criteria for confounding (a slight misnomer because criterion [3] refers to a causal relation). As it turns out, these statistical criteria often agree perfectly with the back-door criteria—that is, you would choose the same set of covariates using either criteria. For example, in Figure 16.2, both the graphical and statistical criteria indicate that one should condition on U to derive an unbiased estimate of the effect of X on Y. It fulfills the graphical criteria because U is not an effect of X, and the only path between X and Y that contains an arrow into X is blocked by U. It fulfills the statistical criteria because U and X will be statistically associated, U will also predict Y, and U is not affected by X or Y. There are cases when the statistical and graphical criteria disagree, however, and when they diverge, it is the statistical criteria that fail.

The DAG in Figure 16.3 gives one example. We are interested in whether having low education increases risk of type II diabetes; the DAG in Figure 16.3 depicts the causal null that education has no effect on diabetes. We have measured mother's diabetes status, but we do not have measures of the family's income when

[3]Sometimes this criterion states, instead, that the confounder must affect the outcome under study. Under this alternative statement of the statistical criteria, the basic argument still follows, in that there are situations in which the statistical and graphical criteria differ, and when this occurs the graphical criteria are correct. The DAGs under which such a discrepancy emerges are slightly more complicated than that in Figure 16.3, but an example is discussed in detail in Greenland et al. (1999).

FIGURE 16.3. A DAG UNDER WHICH CONVENTIONAL CONFOUNDING RULES FAIL.



the individual was growing up or if the individual's mother had any genes that would increase risk of diabetes. Under the assumptions in the DAG in Figure 16.3, should we adjust our analysis for mother's diabetes status? First we consider how we would answer this question with the statistical criteria for a confounder, and then we address it with the graphical criteria. The DAG in Figure 16.3 reflects the assumption that family income during childhood affects both educational attainment and mother's diabetes status. The reasoning is that if an individual was poor as a child, his or her mother was poor as an adult, and this poverty increased the mother's risk of developing diabetes (Robbins et al. 2001, 2005). Mother's diabetes status will be statistically related to the respondent's education, because under these assumptions they share a common prior cause. It will also be related to the risk that the respondent has diabetes, because the mother's genetic risk profile affects both her own and her offspring's diabetes risk. Mother's diabetes is not affected by the respondent's own education level or the respondent's own diabetes status. Thus, mother's diabetes meets all three statistical criteria for a confounder. With the statistical criteria, you would choose to adjust the analysis for mother's diabetic status.

What about the graphical criteria? Would conditioning on mother's diabetes block the back-door path between low education and diabetes? First, note that there is one path between low education and diabetes, and mother's diabetes is a collider on that path. If we do *not* adjust for mother's diabetes, it blocks the path between our exposure and outcome. Adjusting for mother's diabetes *unblocks* this

path and induces a spurious statistical association between low education and diabetes. Under the graphical criteria, one should not include mother's diabetic status as a covariate.[4]

The intuition here is very similar to the reasoning that pro-basketball players who are short will tend to be very fast. Assume that mothers developed diabetes owing either to a genetic predisposition or to experiencing poverty as adults (while raising their children). There may be other reasons as well, but assume these are two non-trivial determinants of a mother's diabetic status. Consider respondents whose mothers had diabetes but no genetic risk factors. These people's mothers likely developed diabetes owing to poverty, implying that the respondents themselves grew up in poverty. Conversely, among respondents with diabetic mothers who did not grow up in poverty, there is probably a genetic risk factor. Conditional on mother's diabetic status (for example, examining only those whose mothers were diabetic), childhood poverty and genetic risk factors will tend to be inversely related; individuals whose mothers did not carry a genetic risk factor will tend to have grown up in poverty. Because of this association, among people with diabetic mothers, low education will be inversely associated with diabetes risk. If low education increases diabetes risk, adjusting for mother's diabetic status (under the assumptions in Figure 16.3) will underestimate this effect. Appendix 16.1 provides some example Stata code to generate data consistent with the causal assumptions in DAG 3 in order to demonstrate this phenomenon.

[4]A variation on the statistical criteria can be used to determine whether, given a sufficient set of covariates $\boldsymbol{Z}$, it is possible to drop any variables from $\boldsymbol{Z}$ and still have a sufficient set for identifying the effect of X on Y. Assume that the sufficient set $\boldsymbol{Z}$ consists of two subsets **A** and **B**, and no variable in set **A** or set **B** is affected by either X or Y. It is unnecessary to adjust for the variables in **B**, given the variables in **A**, if **B** can be broken into two disjoint subsets $\mathbf{B_1}$ and $\mathbf{B_2}$ (no variable in $\mathbf{B_1}$ can be in $\mathbf{B_2}$ and all variables in **B** must be in either $\mathbf{B_1}$ or $\mathbf{B_2}$) such that 1) $\mathbf{B_1}$ is independent of X within strata defined by **A** *and* 2) $\mathbf{B_2}$ is independent of Y within strata defined by X, **A**, and $\mathbf{B_1}$. The implications of these criteria are consistent with the graphical criteria (Greenland et al. 1999). To apply this to the situation in Figure 16.3, imagine that we know conditioning on W, Z_1, and Z_2 is sufficient to identify the effect of X on Y. We would like to know whether conditioning on the empty set (call this set **A**; note that a set of variables can be broken down into two sets—one empty and the other the same as the original set) is sufficient. Now break set **B** (W, Z_1, and Z_2) into $\mathbf{B_1}$ (Z_2) and $\mathbf{B_2}$ (W and Z_1). Z_2 is marginally independent of X, meeting the first criteria above. Z_1 and W are both independent of Y within strata defined by X and Z_2, meeting the second criteria. Thus, if we know that conditioning on all three variables is sufficient, we can use these statistical criteria to establish that conditioning on none of the three variables would also be sufficient. The result might be more easily established using the graphical criteria, however.

Why Sample Selection Threatens Internal Validity as well as Generalizability

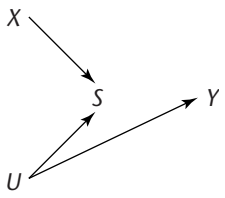
Samples for observational epidemiologic studies are drawn using a variety of criteria. For example, the sample may be drawn from members of a certain occupation (for example, nurses, doctors, or nuns) or residents of a certain community (for example, Framingham or Leisure World Laguna Woods). The possibility that such selection criteria might compromise generalizability is widely recognized. What is sometimes overlooked, however, are the circumstances under which selection criteria can affect internal validity. The sample selection process may sometimes result in spurious statistical associations (that is, associations that do not reflect causal relations between variables measured on the sample population). This potential for bias is of special interest to social epidemiologists, because some of the sample population selection rules use socially relevant characteristics.

On a DAG, we represent selection into the sample as a variable and say that all analyses of a sample are conditioned on selection into that sample. That is, we conceptualize selection as a variable with two values, zero = not selected and one = selected; analyses are restricted to observations where selection = one. The value of this selection variable may be influenced by any number of other variables, including the exposure, the outcome, or other factors that influence the exposure or the outcome (or both). Selection bias may occur if the likelihood of being admitted to the sample depends on both the exposure and the outcome or their respective causes.

To take an extreme example, imagine a study of education's effect on Alzheimer's dementia (AD). Suppose the eligibility criteria for the study are (1) college education or higher, *or* (2) memory impairment. Within the sample, you find a strong inverse correlation between education and AD. In fact, everyone with less than a college education has memory impairment (strongly associated with AD), because otherwise they would not have been eligible for the study. All the sample members with good memory turn out to have high education. Thus, in this sample, higher education is associated with lower risk of AD. Obviously, this is a completely spurious statistical relationship, induced by conditioning sample membership on education and memory impairment. All analyses of the sample are conditional on sample membership, and sample membership is a common effect of the exposure and outcome of interest. No matter what the causal relation between education and Alzheimer's, the statistical associations in the selected sample will differ substantially.

Note that the bias in this example was not a result of drawing a nonrepresentative sample from the "target population" and was not simply a problem

FIGURE 16.4. A DAG FOR SELECTION BIAS.



of generalizability. Instead, this bias arises from how the target population is defined, regardless of whether a representative sample is drawn from that target population. One may well define the target population to be college graduates or those with memory impairment and ask whether, for these people, education protected against AD. Within this population, however, the statistical associations between education and AD will not equal the causal relations.

This example is obvious because the selection criteria were direct measures of the exposure and outcome. Selection may be more subtly related to factors that influence exposure and outcome, however. Imagine that you choose to test the hypothesis that education affects AD risk in a sample with selection based on membership in a high-prestige occupation. Achievement of a high-prestige occupation is likely to be influenced by education, but many people with limited education obtain prestigious jobs by virtue of native talent or intellect (or any number of other explanations, but we will focus only on the intelligence factor). Some evidence indicates that intelligence protects against diagnosis of AD (Schmand et al. 1997). Consider the DAG in Figure 16.4. In this DAG, S represents selection into the sample (based on occupation), and it is influenced by X (representing education) and U (intellect), which is itself a cause of Y (AD). Among the high-prestige job holders, people with limited education are likely to have high intellect, whereas those with low intellect are likely to have quite a lot of education. This is not to say that everyone in the sample with extensive schooling will be dim or that all the smart people will be high-school dropouts. The selection process will merely bias the education–intellect association away from the association in the population as a whole. The strength of the spurious association will depend on the details of the selection process, that is, how strongly education and intellect each affect occupation and whether they interact in any way to determine occupation. Note, however, that if high-education sample members are slightly less likely to have high intellect than low-education sample members, this will increase the AD risk of high-education sample members relative to the low-education sample

members commensurately. Whatever the true causal relation between education and AD, in a study of high-prestige job holders, that relation will tend to be underestimated, *unless it is possible to also condition on intellect.* Alternatively, if the effect of intellect on AD is mediated entirely by some measured covariate, adjusting for that covariate will eliminate the selection bias. This problem is not resolved by using a longitudinal study design unless the effect of intellect on AD is mediated entirely by some measured baseline variable.

Telling the story as in the preceding paragraphs is complicated and prone to generating confusion, but analyzing the DAG is quite straightforward. Given the DAG in Figure 16.4, we can see that S is a collider between X and U; X and U are statistically associated conditional on S. Thus, conditional on S, X and Y are also statistically associated, even under the assumption shown in this DAG that X has no causal effect on Y (the null hypothesis). Note that whether selection exacerbates or reduces bias in estimating a specific causal effect depends crucially on the causal relations among variables determining selection. If we added an arrow from U to X to the DAG in Figure 16.4, selection on S might reduce bias in estimating the effect of X on Y. The relation between collider bias and selection bias is described by Spirtes et al. (1993) and Pearl (1995) and explicated within the framework of epidemiologic study designs by Hernán et al. (2004).

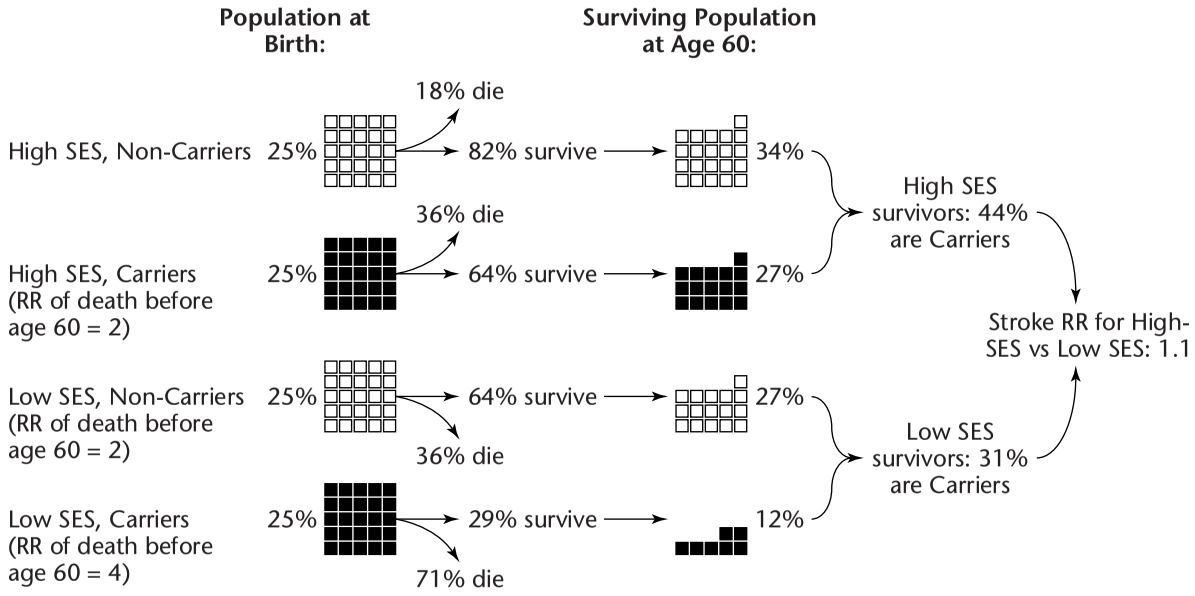
Survivor bias can be thought of as a special case of selection bias. In life-course research on early life exposures and health in old age, a large fraction of the exposed are likely to die before reaching old age, so survivor bias could be influential. Effect estimates for many exposures–outcome combinations are larger among the young and middle-aged than among the old (Elo and Preston 1996; Tate et al. 1998). An especially striking example of this phenomenon is the black–white mortality crossover: mortality is greater for blacks and other disadvantaged groups relative to whites at younger ages, but the pattern reverses at the oldest ages (Corti et al. 1999; Thornton 2004). Does the diminishing magnitude of effect estimates among the elderly indicate that the early life exposures become less important causes of the outcome among the old? Not necessarily. Selective survival models show that attenuated estimates among aged cohorts need not imply diminished effects (Howard and Goff 1998; Mohtashemi and Levins 2002). In a selected group of survivors to old age, observed coefficients for early life exposures may differ from the causal coefficients in the following situations: (1) probability of survival is influenced by early life exposure and some other unmeasured factor, (2) the combined effect of the unmeasured factor and early life exposure on *survival* is not perfectly multiplicative, and (3) the unmeasured factor influences the outcome of interest. This can occur even if the unmeasured factor is statistically independent of exposure at birth (as in the numerical example) and thus would not be considered a confounder.

Consider a simple numerical example of this phenomenon (illustrated in Figure 16.5). If interest is in how mother's socioeconomic status (SES) affects one's stroke risk, and we enroll surviving members of the 1920 birth cohort when they are age sixty, roughly 40 percent of the birth cohort will have died prior to enrollment (Arias 2004). Suppose that those whose mothers had low SES were twice as likely to die as those whose mothers had high SES. Furthermore, suppose there is a "bad" gene, carriers of which have twice the chance of dying before age sixty as non-carriers and also have twice the chance of incident stroke after age sixty. Suppose that at birth, these two risk factors are independent and exactly one-half the population are carriers of each (thus 25 percent of the population are high-SES non-carriers, 25 percent are high-SES carriers, 25 percent are low-SES non-carriers, and 25 percent are low-SES carriers). These factors are perfectly multiplicative for death; that is, risk of death before age sixty for high-SES non-carriers is 18 percent, risk of death for low-SES non-carriers is 36 percent, risk of death for high-SES carriers is 36 percent, and risk of death for low-SES carriers is 72 percent. Given this pattern of death, what are the associations among the survivors? The population, which was 25 percent of each risk combination at birth, at age sixty is 34 percent high-SES non-carrier, 27 percent low-SES non-carrier, 27 percent high-SES carrier, and 12 percent low-SES carrier. Thus, 44 percent of the high-SES group are carriers, whereas only 31 percent of the low-SES group are carriers. Suppose high SES actually had no effect on stroke risk after age sixty (that is, if, for everybody in the sample, had we intervened to flip their mother's SES, they would nonetheless have had the same stroke outcome). Even under this assumption of no causal effect, we would observe that high-SES survivors had an elevated risk of stroke compared with low-SES survivors. Although the spurious statistical association between SES and stroke would vanish within strata of the gene, if the gene is unmeasured, the crude association is biased. Whatever protection (or risk) having a high-SES mother might have conferred against having a stroke after age sixty, it will be biased toward looking harmful among the survivors (in this case, the bias is not very large).

This reasoning follows immediately from a causal DAG such as that in Figure 16.4, showing survival (S) affected by mother's SES (X) and an unmeasured risk factor (U) that also affects stroke (Y). Although the numerical example here makes high SES seem spuriously harmful, survivor bias can operate in either direction, depending on how mother's SES and the unmeasured risk factor combine to affect survival (that is, whether there is interaction). The direction and magnitude of the bias can be estimated under various assumptions about the causal structure, although the assumptions needed are more detailed than those shown in DAGs. In some cases, the plausible range of the bias may be too small to be of concern, but this is not always the case.

FIGURE 16.5. A SIMPLE NUMERICAL EXAMPLE OF SURVIVOR BIAS.

This example assumes that high-SES carriers of the "bad" gene have a relative risk of 2 for death before age 60 compared to high-SES non-carriers; low-SES non-carriers have a relative risk of death of 2 compared to high-SES non-carriers; and low-SES carriers have a relative risk of death of 4 compared to high-SES non-carriers. 40% of the birth cohort is assumed to die before age 60. The gene is assumed to double risk of stroke after age 60, while SES has no effect on stroke after age 60. Very few low-SES carriers will survive to age 60. If the RR for stroke of high vs low-SES individuals is calculated among survivors at age 60 without conditioning on the carrier status, high-SES will be associated with increased risk of stroke. The numbers do not total exactly due to rounding.



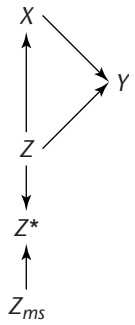
Why Handling Missing Data with Indicator Variables Is Biased Even If the Data Are Missing Completely at Random

Even the best of studies are usually compromised by missing data. Often, the missingness comes here and there, scattered across a large percentage of the observations in the data set. Earl didn't want to reveal his income, Esther was happy to report her income but refused questions on sexual behaviors, Viola broke into tears when asked about participation in community activities such as bridge, and the medical record forms for twelve other sample members were lost. Several methods for handling missing data are available, many of which are unbiased under some assumptions but biased under alternative scenarios (Greenland and Finkle 1995; Little and Rubin 1987). To many researchers, two goals are of preeminent importance: (1) retain everybody in the study so there is still a good chance of getting a statistically significant result, and (2) avoid a lot of extra work. A popular approach to handling missing data that fulfills both goals is to create indicator variables for missingness on each variable (0 = observed, 1 = missing). The variable in question is centered at its mean and all missing values are set to zero. In this way, we can retain everybody in a regression analysis, even if they skipped one or more items. As many a tired researcher has discovered, this approach is also pretty easy to implement. But, we might well ask, does it produce the right answer? Suppose we optimistically assume that the data are missing completely at random. In other words, Viola's shyness regarding social relations had nothing to do with her actual social isolation or any other observed or unobserved characteristic of Viola. The missing data are completely random with respect to exposure and outcome. In this case, would using the missing indicator method to adjust for a putative confounder provide an unbiased effect estimate?

Examine the DAG in Figure 16.6. We are interested in estimating the effect of X on Y, and we recognize that it is important to adjust for Z, a common prior cause of X and Y. Unfortunately, we do not have measures of Z for everyone in our sample. When Z is missing, the variable Z_{ms} takes the value of 1; otherwise it is 0. Because the data are missing completely at random, there are no arrows pointing into Z_{ms} in the DAG. We define a new variable, Z^*, that equals Z whenever Z is observed and equals c (the mean value of observed Z) everywhere else; Z^* is thus determined by both Z and Z_{ms}, and Z^* is thus influenced by both Z_{ms} and Z. Using the missing indicator method, we examine the statistical association between X and Y conditional on Z^* and Z_{ms}.

We can see from this DAG that conditioning on Z^* does not block the backdoor path from X to Y via Z; Z^* is correlated with Z, and that correlation is proportional to the fraction of the sample with observed values of Z. If Z does in fact

FIGURE 16.6. CONDITIONING ON A MISSING VARIABLE INDICATOR.



confound the association between X and Y, there will be residual confounding when adjusting for Z^*, and this residual confounding will be proportional to the fraction of missing. A similar issue will arise in general when confounders are mismeasured. The limitations of this approach to handling missing data are well-demonstrated in the literature (Greenland and Finkle 1995; Little and Rubin 1987); the DAG here is merely a device for clarifying the concepts. It is also clear from the DAG that a complete case analysis, in which we condition on Z and consider only observations where $Z_{ms} = 0$, is unbiased under these assumptions (that is, missing completely at random). The DAG can be extended to consider alternative assumptions about the determinants of missingness.

Why Adjusting for a Mediator Does Not Necessarily Estimate the Indirect Effect

Heated arguments in social epidemiology often focus on questions of mediation. Is the effect of sex on depression mediated by hormonal differences between men and women or differences in social conditions? Are education effects on health in old age mediated by credentials, cognitive differences, or behaviors? Is the association between occupational status and heart disease attributable to psychological consequences of low occupational status or material consequences of low-paying low-status jobs? Mediation tests are crucial for identifying the paths between social factors and health differences. We are often at somewhat of a loss as to how to change the "fundamental" cause of the outcome, but have more optimism that we could change a putative mediator, and the preferred policy

response would obviously depend on the primary mediators. Implicitly, the question of what mediates observed social effects informs our view of which types of inequalities are socially acceptable and which types require remediation by social policies. For example, a conclusion that women are "biologically programmed" to be depressed more than men may ameliorate the social obligation to try to reduce gender inequalities in depression. Yet if people get depressed whenever they are, say, sexually harassed—and women are more frequently sexually harassed than men—this suggests a very strong social obligation to reduce the depression disparity by reducing the sexual harassment disparity.

One definition of the direct effect of exposure X on outcome Y *not mediated by* Z is the effect of X on Y when everyone in the population is forced to receive the same level of Z. A slightly different definition of direct effects, which I adopt here, is the effect of X on Y when everyone in the population is forced to receive the level of Z they would have received for a specific, constant level of X (for example, if X were 0). The distinction between these definitions is important when discussing the decomposition of a total effect into direct and indirect effects. For a discussion of alternative definitions and issues that arise when the exposure interacts with the mediator, see (Kaufman et al. 2004; Robins and Greenland 1992). Although it is possible that the direct effect of X on Y differs depending on the value of Z, I assume for the remainder of this discussion that it does not.

When Z is believed to partially mediate the effect of X on Y, a common approach to quantifying the direct effect is to compare the regression coefficients for X predicting Y in a model simultaneously adjusted for Z to the regression coefficients for X in a model not adjusted for Z (Baron and Kenny 1986; Judd and Kenny 1981). That is:

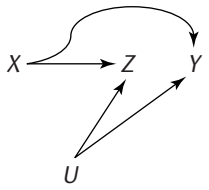
$$E(Y) = \beta_0 + \beta_1 X + \beta_2 Z \tag{3}$$

Assuming that it is known that X affects Z, rather than that Z affects X, the coefficient B_1 is interpreted as the direct effect of X on Y. To calculate the mediated effect, a second regression, unadjusted for Z, is estimated:

$$E(Y) = \gamma_0 + \gamma_1 X \tag{4}$$

The contrast between γ_1 and β_1 is interpreted as the portion of the effect of X on Y that is mediated by Z. Clearly, this interpretation is not correct if Z is a common prior cause of X and Y or to the extent that Z is measured with error. A more subtle problem occurs if X affects Z but there are unmeasured common causes of Z and Y. In this case, the approach described above does not generally give correct estimates of either the direct or indirect effects of X on Y. These

FIGURE 16.7. TESTS FOR DIRECT VERSUS MEDIATED EFFECTS.



unmeasured common causes may be completely unassociated with X; if they affect both Z and Y they will nonetheless bias the estimate of the direct effect of X on Y. This may be surprising because we are not used to considering carefully whether our mediator covariates might have unidentified confounders with the outcome.

The reason the standard approach to testing for mediation fails whenever the putative mediator is confounded is immediately evident from the DAG in Figure 16.7. The variable Z is a common effect of X and U. Within levels of Z, X and U become statistically associated, even if they were marginally statistically independent, and this introduces a spurious statistical association between X and Y within levels of Z. Whatever the causal relation between X and Y, when Z is held constant the statistical association will reflect this causal relation plus the spurious association via U.

We can describe this same phenomenon with an example. Suppose we are interested in knowing whether the relation between education and systolic blood pressure (SBP) is mediated by adult wealth (say, at age sixty). Unfortunately, we do not have any measure of occupational characteristics, and it turns out that having a high autonomy job promotes the accumulation of wealth and also lowers SBP (perhaps owing to diminished stress). Returning to Figure 16.7, now X represents education, Y represents SBP, Z represents wealth at age sixty, and U represents job autonomy. To estimate the effect of education on SBP not mediated by wealth, we need to compare the SBP in people with high and low education if the value of wealth were not allowed to change in response to education. For example, if we gave someone high education but intervened to hold their wealth to the wealth they would have accumulated had they had low education (but changed no other characteristics of the situation), how would SBP change compared with giving the person less education? Unfortunately, we cannot conduct such an intervention. The mediation analysis described previously instead compares the SBP of people with high versus low education but who

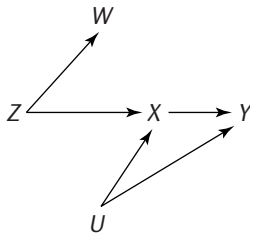
happened to have the same level of adult wealth. Overall, someone with high education will also tend to be wealthier than someone with low education. A high-education person with the same wealth as a low-education person is likely to have accumulated less wealth than expected for some other reason, such as a low autonomy job. Thus, the mediation analysis will be comparing people with high education but low job autonomy to people with low education and average job autonomy. If job autonomy affects SBP, the high-education people will seem to be worse off than they would have been if they had average job autonomy. This will in effect underestimate the direct effect of education on SBP. Under the traditional analysis plan, if we underestimate the direct effect, we will automatically overestimate the mediated effect. This same phenomenon can be explained more formally using counterfactual language. My point here is to note that with a causal DAG, one can see quickly that adjusting for a confounded mediator will induce a spurious association (which may be in either direction) between the primary exposure and outcome.

This observation can be frustrating, because estimating mediation is so important in social epidemiology. In fact, it is so frustrating that researchers sometimes prefer to ignore the problem because, if honestly confronted, it seems to render progress impossible. This is a mistake. First, the injunction that hypothesized mediators be unconfounded in order to draw causal inferences is not any more severe than the demand that primary exposures be unconfounded in order to draw causal inferences. We accept the latter injunction without irritation. Second, if the hypothesized mediators are confounded, we can conduct sensitivity analyses to understand our true uncertainty about the magnitude of the direct or mediated effects. Cole and Hernán (2002) wrote an accessible discussion of this problem walking through a numerical example. Blakely (2002), in a response to Cole and Hernán, called for careful sensitivity analyses to determine whether substantial bias is introduced under realistic assumptions about the strengths of the causal relations.

When Is an Alleged Natural Experiment Valid?

Observational epidemiologists are (or at least should be) constantly concerned that they have not adequately measured and controlled for all common prior causes of their exposure and outcome. For this reason, randomized experiments are strongly preferred to observational studies for demonstrating causality (despite the many other limitations of randomized trials). A randomized trial is represented in the DAG in Figure 16.8. Here Z represents random assignment to treatment group. We will ignore the variable W on this DAG for the moment. Random assignment

FIGURE 16.8. IDENTIFYING A VALID NATURAL EXPERIMENT.



affects treatment received (X), although it does not perfectly determine X because some participants do not adhere to their assigned treatment. There are no causal connections between Z and Y except via X. In the DAG, we show an unmeasured variable U that confounds the association between X and Y, thus forcing us to use the experimental design to test whether X affects Y. The crucial assumption here is that, if we find that average Y differs by treatment assignment Z, this implies that Z affects Y. If Z covaries with Y, this implies that X affects Y, because there is no other possible pathway which would lead to an association between Z and Y except the one via X.

Note that the causal assumptions for a valid trial may be met even if the researcher did not assign the values of Z: the crucial assumption is simply that Z was assigned in a manner otherwise unrelated to the outcome, and its association with X is the only plausible reason it might predict Y. Various natural experiments may fulfill this assumption. We may think that the day of the week one falls ill determines the quality of hospital care received, but there is no other reason for day of illness to influence ultimate health outcomes. In this case, day of symptom onset provides a natural experiment for the effect of quality of hospital care on outcome. A similar idea using hour of birth as an instrument for postpartum length of stay is developed in the study by Malkin et al. (2000). We may think that the weather in the summer before a subsistence farmer's child is born determines the calories that child receives in his first year of life, but weather during that period should have no other effect on the child's health at age ten. Weather then provides a natural experiment for the effect of early caloric intake on later health. We may believe that infants born in hospitals that provide lactation counseling to postpartum mothers are more likely to be breastfed but that being born in such a hospital has no other effect on child health. In this case, being born in a hospital with lactation counseling provides a natural experiment for the effect of breastfeeding on child health. We may believe that women whose mothers or sisters had

breast cancer are unlikely to take hormone therapy at menopause but that having relatives with breast cancer has no other causal link to cardiovascular disease risk. If so, having female relatives with breast cancer is a natural experiment for the effect of hormone therapy on cardiovascular disease.

These examples highlight the core criteria upon which putative natural experiments must be assessed: is there any other reason for the treatment assignment (that is, day of symptom onset, weather summer before birth, birth in a hospital with lactation counseling) to influence the outcome besides via the exposure of interest? For example, if we believe that hospitals with lactation counselors also tend to provide better care in other respects, then we cannot attribute a difference in health between children born at lactation-counseling or non-counseling hospitals strictly to breastfeeding. The natural experiment is not valid. Data from natural experiments are often analyzed with an Instrumental Variables (IV) analysis, in which treatment assignment is referred to as an **instrument** for the effect of X on Y. Specifically, given a causal DAG, we say Z is a valid instrument for the effect of X on Y if Z and X are statistically dependent and if every unblocked path connecting Z and Y contains an arrow pointing into X. An IV effect estimate can be calculated as the ratio of the relation between the instrument and the outcome (the intent to treat effect estimate) and the relation between the instrument and the treatment. To interpret this parameter, we assume that some people would have been treated regardless of the value of the instrument, other people would *not* have been treated no matter what value the instrument took, whereas still a third group, sometimes called the cooperators, would receive the treatment if and only if assigned to receive it by the instrument. We assume nobody in the population is a contrarian (that is, receives treatment only if assigned *not* to receive treatment and avoids treatment only if assigned to receive it). Under these assumptions, the IV estimate provides a consistent estimate of the average effect of receiving treatment on those who received the treatment owing to the value of the instrument.

One interesting and somewhat surprising observation from the DAGs is that an instrument need not directly affect exposure. In Figure 16.8, the relation between W and Y may provide a valid test of the hypothesis that X affects Y even though W does not itself directly affect X but rather shares a common prior cause with X. Here Z affects both W and X, and they are thus statistically associated. Neither W nor Z has any other pathways linking them to Y. If W and Y are statistically associated, under these assumptions it implies that X affects Y. Natural experiments and IV analyses are discussed in more detail in chapter 17 of this book. For accessible discussions of the use of IV analyses to estimate causal effects see (Angrist and Krueger 2001; Currie 1995; Greenland 2000).

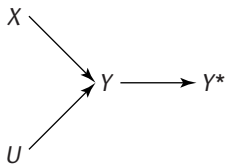
Why It Is a Mistake to Condition on the Dependent Variable

For various reasons, it may be appealing to examine relations between X and Y within a certain range of values of Y. For example, one might want to know whether the effect of education on mental status among individuals with below average mental status is the same as the effect of education among individuals with above average mental status. Alternatively, one might suspect that the outcome measurement available becomes increasingly unreliable at high levels and therefore wish to exclude any high-scoring respondents from the analysis.

These decisions can introduce important bias into an analysis, and this can be seen with a DAG such as that in Figure 16.9. In this DAG, we are interested in the effect of X on Y; Y is also influenced by U, but U is statistically independent of X. Under these assumptions, a simple analysis of the statistical relation between X and Y (without statistical adjustment for any other covariates) gives an unbiased estimate of the causal effect. Suppose however, that we condition on some values of Y. Let us define a variable Y^* that is one if Y is above a threshold value and zero if it is below. Now we examine the relation between X and Y only among those with $Y^* = 1$. This turns out to have an undesirable consequence: X and U are likely to be statistically associated among respondents with $Y^* = 1$. As a result, the statistical relation between X and Y will now be confounded by the effect of U on Y (although the direction of confounding will not necessarily be the same as the direction of the effect of U on Y).

Let us consider the question of education's effect on mental status, using the mini-mental status exam (MMSE) as a measure of mental status. The MMSE ranges from zero to thirty, and an MMSE score below twenty-four is considered a clinically important threshold for impairment (Folstein et al. 1975). Suppose we ask whether the effect of education on MMSE is the same for respondents with MMSE equal to or above twenty-four as for respondents with MMSE below twenty-four. We assume that MMSE score is influenced by education and also influenced by intelligence (IQ), although IQ is unrelated to

FIGURE 16.9. CONDITIONING ON THE DEPENDENT VARIABLE.



education (if IQ itself affects education, the analysis is obviously confounded, but we make the optimistic assumption here that IQ does not affect education). Thus, in the DAG in Figure 16.9, U represents IQ, X represents education, Y represents MMSE, and Y^* is an indicator for whether MMSE is above twenty-four. In general, under this setup, we will underestimate the association between education and MMSE in both strata of Y^* unless we are able to simultaneously adjust for IQ. Among the high-functioning individuals (those with high MMSE scores), those with low education are more likely to have unusually high IQ. Among the low-functioning individuals, those with high education are more likely to have unusually low IQ. Even though IQ and schooling are statistically independent in the population, they are inversely correlated within strata of their common effect, MMSE. Note that the rules for drawing causal DAGs described earlier in the chapter would not require that U in Figure 16.9 be shown, because U is not a direct cause of more than one other variable in the DAG. The rules for drawing causal DAGs specify what is required for the d-separation rules to be applicable, but this phenomenon is not addressed by the d-separation rules.

This phenomenon is also relevant when considering how to respond to an artificial ceiling on the measurement of Y. One tempting but erroneous approach is to drop all of the observations with ceiling values of Y. This is effectively conditioning the analysis on the value of Y and will bias the statistical association between X and Y. An important caveat is that the preceding discussion only applies if X actually does affect Y. If X has no effect on Y, then Y is not a common effect of X and U. In this situation, conditioning on Y should not influence the estimated relation between X and Y—it should be zero in every strata. This finding is discussed in introductory econometrics texts, including Kennedy (1998) and Wooldridge (2002), although it is not generally demonstrated with DAGs.

Why Adjusting for Baseline Values Can Bias Analyses of Change

Our final example of how DAGs can clarify otherwise confusing analysis decisions relates to analyses of change. When the substantive question is whether an exposure X, measured at baseline, affects changes in the value of Y over a follow-up time period, an important analytic decision is whether to condition on the value of Y as measured at baseline. This conditioning may take the form of restriction or stratification, but most frequently the decision is whether to include Y at baseline as an independent variable in a regression model. Let us take as a substantive example the effect of exposure to violence (ETV) in early childhood on changes

in depressive symptoms in adulthood. Suppose that adults at average age thirty are enrolled, and depressive symptoms are assessed with the Centers for Epidemiologic Studies Depression scale at baseline ($CESD_1$) and again after five years of follow-up ($CESD_2$). The CESD is a continuous scale ranging from zero to sixty, in which higher scores indicate worse depressive symptoms (Radloff 1977). Our (hypothetical) ETV measure is dichotomous and based on exposures before age fifteen. At baseline, when respondents are average age thirty, ETV is associated with higher average CESD scores. We would like to know if ETV also causes increases in depressive symptoms over the five-year follow-up of adults. That is, for any given person who was not exposed to violence in childhood, would her change over the five-year follow-up period have differed had she in fact been exposed to violence?

One possible analysis would be to estimate a baseline-adjusted change score model using regression, where the CESD change score is the difference between CESD at follow-up and CESD at baseline:

$$CESD_2 - CESD_1 = \gamma_0 + \gamma_1 ETV + \gamma_2 CESD_1 + \varepsilon_i \quad (5)$$

It has been shown elsewhere (Laird 1983) that the previous model provides the same coefficient for ETV(γ_1) as does a lagged-effects model such as:

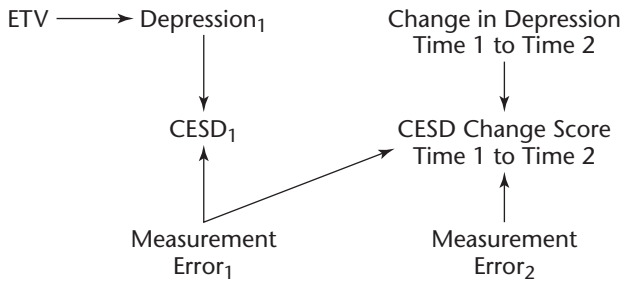
$$CESD_2 = \gamma_0 + \gamma_1 ETV + \gamma_2^* CESD_1 + \varepsilon_i \quad (6)$$

We will focus on whether the statistical analysis in equation (5) answers our causal question, but keep in mind that if the analysis in equation (5) fails to answer this question, estimation of equation (6) will also fail. Alternatively, we could estimate a change score model without baseline adjustment:

$$CESD_2 - CESD_1 = \beta_0 + \beta_1 ETV + \varepsilon_i \quad (7)$$

In either regression model a number of other covariates believed to directly affect ETV and change in depressive symptoms might also be included. It turns out that β_1 and γ_1 are frequently quite different numbers, so they both cannot represent the "right" answer to a specific causal question. Figure 16.10 is a causal DAG under which a baseline adjusted analysis (as in equation [5]) would give a positively biased estimate of the effect of ETV on change in depression, but an unadjusted analysis, as in equation (7), would give an unbiased estimate under the null. The major conceptual point in this DAG is that CESD is an imperfect measure of a latent construct: depressive symptoms. The CESD score is influenced both by true underlying depression and by some error in measuring that depression. This reflects the well-documented finding that the CESD scale has imperfect reliability (McDowell and Newell 1996). The phenomenon in the following description could also occur because of instability in the construct of depression, but that is outside

FIGURE 16.10. AN EXAMPLE WHEN BASELINE ADJUSTMENT BIASES ANALYSES OF CHANGE.



the scope of this discussion. Because ETV and $CESD_1$ are correlated and ETV is temporally prior to $CESD_1$, we assume that ETV affects baseline depressive symptoms. Over the five-year follow-up, some true change in depressive symptoms will occur. We are not privy to the true change, but we will observe the change in CESD scores, which is strongly influenced by true change. Unfortunately, the change in CESD scores is *also* influenced by the error in measuring CESD at baseline. If the baseline error was positive, the CESD change score will tend to be negative, purely due to regression to the mean. If the baseline error was negative, regression to the mean will tend to push the change score in a positive direction. The error in measuring $CESD_2$ will also influence the change score, and if the two errors are perfectly correlated, there will be no regression to the mean; however, psychometric assessments of the CESD scale indicate substantial measurement error that is uncorrelated across time periods. This is the reasoning for drawing the DAG as we did in Figure 16.10. Under these assumptions, if ETV has no effect on change in depression during the follow-up period, ETV and change in CESD score will be statistically independent: the β_1 estimated in equation (7) is unbiased. The only path in the diagram connecting ETV and change score (ETV—$Depression_1$—$CESD_1$—$error_1$—CESD change score) is blocked by $CESD_1$ (a collider). Thus, analyses not adjusted for $CESD_1$ provide unbiased estimates of the overall (that is, total) effect of ETV on change.

Conditional on $CESD_1$, however, ETV and CESD change score are spuriously correlated, because conditioning on $CESD_1$ "unblocks" the previously described path. The intuition is just as with the previous examples of conditioning on common effects. Anyone with a high $CESD_1$ has either severe baseline depression symptoms (high $depression_1$) or large positive measurement $error_1$ (or both). A person without depression who has a high $CESD_1$ must have a positive

$error_1$. If a person with severe depressive symptoms scores a low $CESD_1$, error is negative. Thus, *within levels of* $CESD_1$, $depression_1$ and $error_1$ are inversely correlated, and ETV and $error_1$ are inversely correlated. Because $error_1$ contributes negatively to change score, change score and $error_1$ are negatively correlated, an example of the regression to the mean phenomenon. Hence, conditional on $CESD_1$, ETV and CESD change score are positively correlated. Therefore baseline-adjusted ETV coefficients are positive, even when ETV does not affect change in depressive symptoms. The spurious correlation is proportional to the error in the CESD measure and the strength of the ETV-$CESD_1$ relationship. This finding has been demonstrated mathematically (Yanez et al. 1998) and with an applied example (Yanez et al. 2002). The issue is discussed in more detail using DAGs in (Glymour et al. 2005).

Caveats and Conclusion

Directed acyclic graphs do not convey information about important aspects of the causal relations, such as the magnitude or functional form of the relations (for example, linearity, interactions, or effect modification). This can be frustrating because not all biases are created equal, and it would be nice to establish which ones can safely be ignored. For example, Greenland (2003) compares the likely bias introduced by adjusting for a collider with the bias that would result from failing to adjust for a common prior cause. His findings suggest that, if the collider is not a direct effect of the exposure and the outcome, one might prefer to adjust on the grounds that the bias potentially introduced by *failing* to adjust for the variable is likely to be larger than the bias potentially introduced by mistakenly adjusting for it. Exploring the magnitude of potential biases under a range of assumptions is invaluable, and there are many approaches to doing this. One option is to generate simulated data sets based on DAGs, for example as in the simple code in Appendix 16.1. More sophisticated simulations can be conducted in many statistical packages, including freeware available online (TETRAD, 2005)[5].

Drawing a DAG that adequately describes our prior beliefs or assumptions is sometimes difficult. To the extent that using DAGs forces greater clarity about

[5]TETRAD is a project on statistical data and causal inference. The causal discovery algorithms used by the accompanying software package have been controversial (see for example, Robins and Wasserman, 1999 and rejoinders). Apart from this debate regarding the reliability of the causal inference algorithms, the software includes a convenient routine to simulate data sets based on assumptions in structural equation models without specifying programming code.

assumptions, this seems advantageous. Though it may seem an impossible task to draw a "true" DAG, to the extent that we are uncertain about how to specify the DAG, we should also be uncertain about the causal interpretation of our statistical tests.

Directed acyclic graphs are a convenient device for expressing ideas explicitly and understanding how causal relations translate into statistical relations. Causal DAGs provide a simple, flexible tool for thinking about many epidemiological problems. My goal in this chapter was to demonstrate how an array of apparently disparate problems in epidemiologic reasoning can be expressed and resolved with causal DAGs. This is part of the remarkable convenience of learning the d-separation rules. Rather than considering each case of a potential bias as a separate problem and struggling for the "right" answer, DAGs help provide a unified way of evaluating a potential analysis plan for any specific question of interest and set of causal assumptions. Although in some cases the issues raised are especially pertinent in research on social determinants of health, these problems are by no means limited to social epidemiology. The last two decades of progress on causal inference, of which the use of causal DAGs is only a part, has the potential to substantially enhance applied epidemiologic work, and these improvements may be especially beneficial in social and life-course epidemiology.

APPENDIX 16.1

The following Stata commands create a data set with five normally distributed variables: W, X, Y, $Z1$, and $Z2$. Variable $Z1$ affects X and W; $Z2$ affects W and Y. There are no other causal relations between variables (for example, we assume the null hypothesis that X has no effect on Y). This is the same causal structure as shown in Figure 16.3, although all variables are assumed to be continuous. Under these assumptions, W meets conventional statistical criteria for a confounder but not the graphical criteria. As shown in the two regressions, conditioning on W induces a negative statistical association between X and Y.

Manipulating the path coefficients can illustrate how the size of the bias induced by adjustment for W depends on the strength of these relations. Please note that several assumptions about the causal structure are implicit in the following code but *not* encoded in the corresponding DAG. For example, the code specifies linear and additive causal effects. The DAG encodes no such assumptions and would thus be consistent with other specifications. The magnitude of the bias

induced by conditioning on *W* is sensitive to these assumptions about functional form.

```
set obs 10000
* Generate constants that determine the magnitude of
the causal effects (that is, path coefficients).
gen Z1toX = 1
gen Z1toW = 1
gen Z2toW = 1
gen Z2toY = 1
* Generate Z1 and Z2 as normally distributed random
variables.
gen Z1 = invnorm(uniform())
gen Z2 = invnorm(uniform())
* Generate random components for all other variables:
W, X, and Y
gen W_random = invnorm(uniform())
gen X_random = invnorm(uniform())
gen Y_random = invnorm(uniform())
* Generate W as a function of Z1, Z2, and a random
component
gen W = Z1toW*Z1 + Z2toW*Z2 + W_random
* Generate X as a function of Z1 and a random component
gen X = Z1toX*Z1 + X_random
* Generate Y as a function of Z2 and a random component
gen Y = Z2toY*Z2 + Y_random
* Describe the data generated means
corr W X Y Z1 Z2
* Run regressions with and without adjustment for W to
estimate the effect of X on Y.
reg Y X
reg Y X W
```

Acknowledgments

I wish to thank the following people for helpful comments and suggestions on earlier drafts of this chapter: Jennifer Weuve, Katherine Hoggatt, Clark Glymour, James Robins, David Rehkopf, and Sander Greenland. All remaining errors and confusions are my own.

References

The asterisks indicate especially accessible or influential references.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives, 15,* 69–85.

Arias, E. (2004). United States life tables, 2001. *Natl Vital Stat Rep, 52.*

Baron, R. M., & Kenny, D. A. (1986). The moderator mediator variable distinction in social psychological-research—conceptual, strategic, and statistical considerations. *J Pers Soc Psychol, 51,* 1173–1182.

Blakely, T. A. (2002). Commentary: Estimating direct and indirect effects—fallible in theory, but in the real world? *International Journal of Epidemiology, 31,* 166–167.

*Cole, S. R., & Hernán, M. A. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology, 31,* 163–165.

Corti, M. C., Guralnik, J. M., Ferrucci, L., et al. (1999). Evidence for a Black-White crossover in all-cause and coronary heart disease mortality in an older population: The North Carolina EPESE. *American Journal of Public Health, 89,* 308–314.

Currie, J. (1995). Welfare and the well-being of children. In F. Welch (Ed.), *Fundamentals of pure and applied economics* (p. 165.). Chur, Switzerland: Harwood Academic Publishers.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association, 95,* 407–424.

Elo, I. T., & Preston, S. H. (1996). Educational differentials in mortality: United States, 1979–85. *Soc Sci Med, 42,* 47–57.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res, 12,* 189–198.

Glymour, C. (1986). Statistics and causal inference—statistics and metaphysics. *Journal of the American Statistical Association, 81,* 964–966.

Glymour, C., Spirtes, P., & Richardson, T. (1999). On the possibility of inferring causation from association without background knowledge. In C. Glymour, & G. Cooper (Eds.), *Computation, causation, and discovery* (p. 323–331), Menlo Park, CA, and Cambridge, MA: AAAI Press/The MIT Press.

Glymour, C., (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology.* Cambridge, MA: MIT Press.

Glymour, M. M., Weuve, J., Berkman, L. F., Kawachi, I., & Robins, J. M. (2005). When is baseline adjustment useful in analyses of change? An example with education and cognitive change. *Am J Epidemiol,* 162, 267–278.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology,* 14, 300–306.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology, 29,* 722–729.

Greenland, S., & Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology, 142,* 1255–1264.

Greenland, S., & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology, 15,* 413–419.

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

Hennekens, C. H., & Buring, J. E. (1987). *Epidemiology in medicine* (1st ed., p. 383). Boston/Toronto: Little, Brown and Company.

*Hernán, M. A. (2004). A definition of causal effect for epidemiological research. *J Epidemiol Community Health, 58*, 265–271.

*Hernán, M. A., Hernandez-Diaz, S. & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology, 15*, 615–625.

Hernán, M. A., Hernandez-Diaz, S. Werler, M. M. & Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology, 155*, 176–184.

Holland, P. W. (1986a). Statistics and Causal Inference. *Journal of the American Statistical Association, 81*, 945–960.

Holland, P. W. (1986b). Statistics and causal inference—rejoinder. *Journal of the American Statistical Association, 81*, 968–970.

Howard, G., & Goff, D. C. (1998). A call for caution in the interpretation of the observed smaller relative importance of risk factors in the elderly. *Annals of Epidemiology, 8*, 411–414.

Judd, C. M., & Kenny, D. A. (1981). Process analysis—estimating mediation in treatment evaluations. *Eval Rev, 5*, 602–619.

Kaufman, J. S., & Cooper, R. S. (1999). Seeking causal explanations in social epidemiology. *Am J Epidemiol, 150*, 113–120.

Kaufman, J., Maclehose, R., & Kaufman, S. (2004). A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiol Perspect Innov, 1*, 4.

Kaufman, J. S., & Cooper, R. S. (2001). Commentary: Considerations for use of racial/ethnic classification in etiologic research [see comment]. *American Journal of Epidemiology, 154*, 291–298.

Kennedy, P., (1998). *A guide to econometrics* (4th ed., p. 468). Cambridge, MA: The MIT Press.

Laird, N. (1983). Further comparative analyses of pretest-posttest research designs. *The American Statistician*, 37, 329–330.

Little, R., & Rubin, D. B. (1987). *Statistical analysis with missing data. Wiley series in probability and mathematical statistics applied probability and statistics.* New York: Wiley.

Maldonado, G., & Greenland, S. (2002). *Estimating causal effects. International Journal of Epidemiology; 31*, 422–429.

Malkin, J. D., Broder, M. S., & Keeler, E. (2000). Do longer postpartum stays reduce newborn readmissions? Analysis using instrumental variables. *Health Services Research, 35*, 1071–1091.

McDowell, I., & Newell, C. (1996). *Measuring health: A guide to rating scales and questionnaires* (2nd ed.). New York: Oxford University Press.

Mohtashemi, M., & Levins, R. (2002). Qualitative analysis of the all-cause Black-White mortality crossover. *Bull Math Biol, 64*, 147–173.

Parascandola, M., & Weed, D. L. (2001). Causation in epidemiology. *Journal of Epidemiology and Community Health, 55*, 905–912.

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

*Pearl, J. (2000). *Causality* (p. 384). Cambridge, UK: Cambridge University Press.

*Pearl, J. (2001). Causal inference in the health sciences: A conceptual introduction. *Health Services & Outcomes Research Methodology, 2*, 189–220.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 384–401.

Robbins, J. M., Vaccarino, V., Zhang, H. P., & Kasl, S. V. (2001). Socioeconomic status and type 2 diabetes in African American and non-Hispanic white women and men: Evidence from the Third National Health and Nutrition Examination Survey. *American Journal of Public Health, 91,* 76–83.

Robbins, J. M., Vaccarino, V., Zhang, H. P., & Kasl, S. V. (2005). Socioeconomic status and diagnosed diabetes incidence. *Diabetes Res Clin Pract, 68,* 230–236.

Robins, J. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases, 40,* 139S–161S.

Robins, J. M. (1995). Comment on Judea Pearl's paper, "Causal diagrams for empirical research." *Biometrika, 82,* 695–698.

Robins, J., & Wasserman, L. (1999). On the impossibility of inferring causation from association without background knowledge. In C. Glymour, & G. Cooper (Eds.), *Computation, causation, and discovery* (p. 305–321). Menlo Park, CA, and Cambridge, MA: AAAI Press/The MIT Press.

*Robins, J. M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology, 12,* 313–320.

*Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology, 3,* 143–155.

Rothman, K. J., & Greenland, S. (1998). *Modern epidemiology* (2nd ed.). Philadelphia, PA: Lippincott-Raven.

Schmand, B., Smit, J. H., Geerlings, M. I., & Lindeboom, J. (1997). The effects of intelligence and education on the development of dementia. A test of the brain reserve hypothesis. *Psychological Medicine, 27,* 1337–1344.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search. Lecture Notes in Statistics* (Vol. 81). New York: Springer-Verlag.

*Spirtes, P., Glymour, C. & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.

Tate, R. B., Manfreda, J., & Cuddy, T. E. (1998). The effect of age on risk factors for ischemic heart disease: The Manitoba Follow-Up Study, 1948–1993. *Annals of Epidemiology, 8,* 415–421.

TETRAD (2005). *The Tetrad project.* Retrieved on January 16, 2006, from www.phil.cmu.edu/projects/tetrad.

Thornton, R. (2004). The Navajo-US population mortality crossover since the mid-20th century. *Population Research and Policy Review, 23,* 291–308.

Woodward, J. (2003). *Making things happen: A theory of causal explanation* (p. 410). New York: Oxford University Press.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data (*p. 752). Cambridge, MA: Massachusetts Institute of Technology.

Yanez, N. D., Kronmal, R. A., & Shemanski, L. R. (1998). The effects of measurement error in response variables and tests of association of explanatory variables in change models. *Statistics in Medicine,* 17, 2597–2606.

Yanez, N. D., Kronmal, R. A., Shemanski, L. R., & Psaty, B. M. (2002). A regression model for longitudinal change in the presence of measurement error. *Annals of Epidemiology,* 12, 34–38.