American Journal of Epidemiology

Vol. 155, No. 2
Printed in U.S.A.

Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology

Miguel A. Hernán,[1] Sonia Hernández-Díaz,[2] Martha M. Werler,[2] and Allen A. Mitchell[2]

Common strategies to decide whether a variable is a confounder that should be adjusted for in the analysis rely mostly on statistical criteria. The authors present findings from the Slone Epidemiology Unit Birth Defects Study, 1992–1997, a case-control study on folic acid supplementation and risk of neural tube defects. When statistical strategies for confounding evaluation are used, the adjusted odds ratio is 0.80 (95% confidence interval: 0.62, 1.21). However, the consideration of a priori causal knowledge suggests that the crude odds ratio of 0.65 (95% confidence interval: 0.46, 0.94) should be used because the adjusted odds ratio is invalid. Causal diagrams are used to encode qualitative a priori subject matter knowledge. *Am J Epidemiol* 2002;155:176–84.

abnormalities; causality; confounding factors (epidemiology); inference; selection bias

In epidemiologic studies, statistical analyses are typically organized around three different sets of variables: the exposure, the outcome, and the confounder(s). The exposure and outcome are usually determined by the causal question under investigation. The confounders, on the other hand, are not so clearly defined; they must first be identified and then appropriately adjusted for in the analysis.

A number of authors have emphasized that confounder identification must be grounded on an understanding of the causal network linking the variables under study (i.e., a priori subject-matter or expert knowledge) (1–8). Yet, some widely used approaches to confounder identification are centered on statistical associations. One common approach, which we will call strategy 1, has been the application of automatic variable selection procedures, such as stepwise selection (9). The implicit assumption underlying this approach is that, although not all variables selected will be confounders, all important confounders will be selected. A second common approach, strategy 2, compares adjusted and unadjusted effect estimates. If the relative change after adjustment for certain variable(s) is greater than 10 percent, for example, then the variable(s) is selected (10). Implicit in this approach is that any variable substantially associated with an estimate change is worth adjusting for. Most epidemiology textbooks recommend a third approach, strategy 3, that consists of checking whether some necessary criteria for confounding are met. Generally, it is stated that a confounder is a variable associated with the exposure in the population, associated with the outcome conditional on the exposure (e.g., among the unexposed), and not in the causal pathway between the exposure and the outcome. A further refinement is to replace the second condition by the condition that the potential confounder is a causal risk factor or a marker for a causal risk factor (11). Strategies 1 and 2 rest only on statistical associations that can easily be identified from the data. Strategy 3 combines statistical associations from the data with some background knowledge about the causal network that links exposure, outcome, and potential confounders.

All three strategies may lead to bias from the omission of important confounders or inappropriate adjustment for nonconfounders (3–5, 7, 8, 12). Here, we will describe a real example from research on birth defects in which all three strategies prefer the adjusted effect estimate over the crude effect estimate. However, we will use our a priori subject-matter knowledge to argue that the crude estimate should probably be preferred. We will utilize causal diagrams (4, 5, 13, 14) to represent our qualitative a priori assumptions about the underlying biologic mechanisms. First, we briefly review confounding and causal diagrams.

CONFOUNDING, CONFOUNDERS, AND CAUSAL DIAGRAMS

Intuitively, two variables E and D will be statistically associated if one is a cause of the other (e.g., smoking and lung cancer), if they share a common cause (e.g., yellow fingers and lung cancer share smoking as a common cause), or

Received for publication August 17, 2000, and accepted for publication July 9, 2001.

Abbreviations: CI, confidence interval; DAG, directed acyclic graph; OR, odds ratio; RR, risk ratio.

[1] Department of Epidemiology, Harvard School of Public Health, Boston, MA.

[2] Slone Epidemiology Unit, Boston University School of Public Health, Brookline, MA.

Reprint requests to Dr. Miguel Hernán, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115 (e-mail: miguel_hernan@post.harvard.edu).

both. If *E* precedes *D*, the overall association between *E* and *D* will have two components: a spurious one that is due to the sharing of common causes and another due to the causal effect of *E* on *D*. The goal of etiologic research from observational data is to estimate the latter. The former component produces confounding (4, 5).

One way to eliminate a spurious association is to adjust, stratify, or condition on the common cause; for example, we would find no association between yellow fingers and lung cancer among nonsmokers. Confounders are variables that when stratified on or adjusted for will eliminate (or diminish) the spurious component of the association between exposure and disease.

The presence of common causes, and therefore of confounding, can be represented by causal diagrams known as directed acyclic graphs (DAGs) (12–14). Briefly, these diagrams link variables by arrows that represent direct causal effects (protective or causative) of one variable on another. Figures 1–8 are selected examples of causal diagrams that link the variables *E*, *D*, and *C*. We use *U* to depict unmeasured variables. Because causes precede their effects, these graphs are acyclic: One can never start from one variable and, following the direction of the arrows, end up at the same variable. In figure 1, *E* causes *D*, and both *D* and *E* are causes of *C*; in figure 2, *E* does not cause *C* but both share an unmeasured common cause U_1.

In figures 1–4, exposure and disease do not share common causes; that is, no variable connects both *E* and *D* by following only forward-pointing arrows. Their crude association lacks a spurious component and thus is wholly due to the causal effect of *E* on *D*. There is no confounding, and no adjustment for confounding is necessary. The odds ratio $(OR)_{ED}$ measures the causal effect of *E* on *D* (on the odds ratio scale). On the other hand, in figures 5–8, exposure and

FIGURE 1. Low folate intake (*E*) may increase the risk of preterm delivery and infant low birth weight (*C*) (Am J Clin Nutr 2000;71(suppl):1295s–303s), and many birth defects (*D*) result in preterm deliveries and low birth weight infants (Am J Dis Child 1991;145:1313–18).

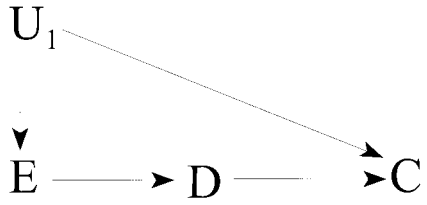


FIGURE 2. The association between multivitamin use (*E*) and pregnancy outcomes may be due to shared sociodemographic characteristics. For example, teen pregnancy age (U_1) can result in lower micronutrient intake (J Epidemiol Community Health 2000;54:17–23) and in poor prenatal weight gain and low birth weight infants (*C*) (J Sch Health 1998;68:271–5). Some malformations (*D*) involve incomplete or small fetuses, which may have an impact on birth weight and maternal weight gain.

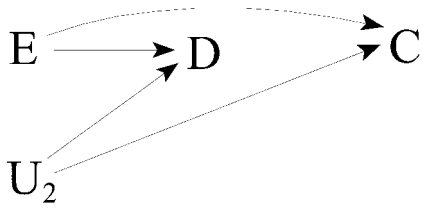


FIGURE 3. The association between the birth defect of interest (*D*) and maternal weight gain (*C*) may be due to shared genetic characteristics, such as an enzymatic polymorphism U_2, which may affect the risk of this birth defect and, independently, maternal weight gain. *E*, multivitamin use.

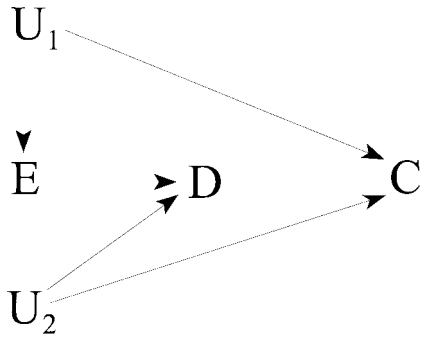


FIGURE 4. This could be a combination of figures 2 and 3.

outcome share a common cause. Hence, the association has a spurious component, there is confounding, and the crude OR_{ED} is a biased estimator of the causal effect of *E* on *D*. Does adjustment for *C* in each of figures 1–8 reduce confounding or does it introduce bias? A major strength of using DAGs is that a set of simple graphical rules can be applied to answer this question (Appendix). However, the judgment as to which variables on a DAG cause which others must in general be based on subject matter considerations.

Let us first concentrate on figures 5–8. Figure 5 depicts *C* as a common cause of *E* and *D*, whereas in figures 6–8 *U* is the common cause. We say that *C* is a confounder in figure 5 and that *U* is a confounder in figures 6–8. To eliminate the spurious component of the association between exposure and outcome, we can condition on the confounder and calculate the $OR_{ED|C}$; that is, we adjust for the common cause. Thus, in figure 5 the $OR_{ED|C}$ adjusted for *C* is a valid estimator of the causal effect on the odds ratio scale within levels of *C*. Furthermore, if (as we shall assume for simplicity) the stratified odds ratio $OR_{ED|C}$ is constant over levels of *C* and the disease is rare (at each joint level of *E* and *C*), then

FIGURE 5. Multivitamin use (*E*) may reduce the risk of certain birth defects (*D*), and maternal age (*C*) may affect multivitamin use (J Epidemiol Community Health 2000;54:17–23).

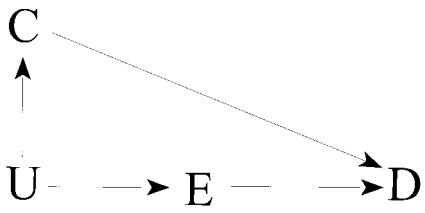


FIGURE 6. Maternal obesity (*C*) may cause certain defects (*D*) (JAMA 1996;275:1089–92). Advanced maternal age (*U*) may increase the risk of obesity and increases the chances of periconceptional multivitamin use (*E*) (J Epidemiol Community Health 2000;54:17–23).

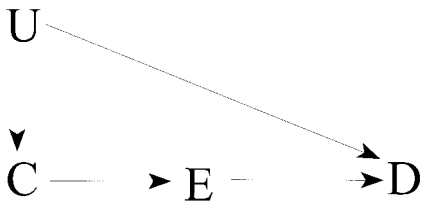


FIGURE 7. History of birth defects in the family or in previous pregnancies (*C*) may lead to more careful pregnancy planning and counseling, which will increase the chances of periconceptional multivitamin use (*E*) (Can Fam Physician 1999;45:2053–7). If a genetic factor (*U*) was a cause of previous malformations in the family, it may also affect the risk of that malformation (*D*) in the current pregnancy (N Engl J Med 1994;331:1–4).

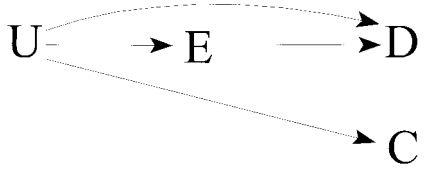


FIGURE 8. A certain enzymatic polymorphism (*U*) decreases plasma folate levels early in pregnancy (*E*) (Am J Hum Genet 1998;62:1044–51) and increases the risk of neural tube defects (*D*) (Am J Epidemiol 2000;151:862–77), perhaps through mechanisms other than folate levels. *C* represents enzymatic function measured after delivery. *C* may occur after *E* or *D* but still be a marker for a factor's occurrence before *E* and *D*.

the stratified odds ratio closely approximates the stratified risk ratio. Then $OR_{ED|C}$, unlike the crude OR_{ED}, also quantifies the causal effect of *E* on *D* in the whole population. But what about figures 6–8? Here the common cause is unmeasured, and therefore we cannot adjust for it. In figure 6, the causal pathway from *U* to *D* is mediated through *C*. Intuitively, if we condition on a specific value of *C*, then *U* cannot affect *D* because *U* only affects *D* by changing the value of *C*. In other words, within levels of *C*, *U* is no longer a cause of *D*, and therefore the spurious association (confounding) disappears. Therefore, $OR_{ED|C}$ adjusted for *C* is a valid estimator of the causal effect of the exposure *E* on the outcome *D*. (This would still be true even if *C* occurred temporally after *E*.) Similar reasoning can be used to deduce that *C* should be adjusted for in figure 7. In both cases, we say that *C* is a confounder, although it is not a causal confounder in the sense that *C* itself is not a common cause of exposure and disease. Once we adjust for the confounder *C*, *U* ceases to be a confounder because it no longer induces a spurious association between exposure and disease.

The situation is different in figure 8, where *C* is not in the causal pathway between the unmeasured confounder *U* and either *E* or *D*. As a result, adjusting for *C* will not remove the spurious association between *E* and *D* due to its common cause *U*. However, if *C* is strongly associated with *U*, adjusting for *C* will remove a large part of the confounding. In the limit, if *C* were perfectly correlated with *U*, then all confounding would be removed when adjusting for either *C* or *U*. We say that *C* is a surrogate confounder in figure 8. Often, when a confounder cannot be adequately measured, it is better to adjust for a surrogate confounder than to use the crude odds ratio (1). For example, if *C* were a misclassified version of *U*, the stronger the association between them (i.e., the smaller the measurement error), the better is confounding taken into account.

Let us now turn our attention to figures 1–4. Even though *C* is not a confounder, is the adjusted $OR_{ED|C}$ a valid estimator of the causal effect? No. Adjustment for *C* is not only unnecessary but harmful. To explain why, let us focus on figure 1. Suppose *E* represents being on a diet and *D* represents a recent diagnosis of a non-diet-related cancer. Let $C = 1$ if the person had recent weight loss greater than 5 kg, and $C = 0$ otherwise. Assume that dieting does not cause cancer and therefore erase the arrow from *E* to *D*. We have seen that two variables may be associated when one is the cause of the other or when they have common causes. Neither case is true in this example, so dieting and cancer are statistically independent (OR = risk ratio (RR) = 1). In other words, knowing that someone was dieting does not change the probability that she develops cancer. Now let us condition on the common effect *C* (common effects are known as colliders in causal graph theory) and check if *E* and *D* remain independent within levels of *C*. Among those who lost weight ($C = 1$), does the probability of someone's having cancer change if we know that she was not dieting? Yes, it does. Given that a person lost weight, it is more likely that she had cancer if she was not dieting. Thus, within those who lost weight, dieting and cancer are inversely associated. See tables 1 and 2 for a numerical example.

In general, conditioning on a common effect or collider *C* creates a spurious association between *E* and *D* (15). The

TABLE 1. Hypothetical study on dieting (*E* = 1) and non-diet-related cancer (*D* = 1), unconditional association

	D = 1	*D* = 0
E = 1	100	100
E = 0	200	200

RR_{ED}* = 1

* RR, risk ratio.

TABLE 2. Hypothetical study on dieting ($E = 1$) and non-diet-related cancer ($D = 1$) stratified by their common effect, recent weight loss ($C = 1$ if yes)

	$C = 1$		$C = 0$	
	$D = 1$	$D = 0$	$D = 1$	$D = 0$
$E = 1$	55	25	45	75
$E = 0$	70	10	130	190

$RR_{ED|C=1}$* = 0.79; $RR_{ED|C=0}$ = 0.92

* RR, risk ratio.

practical implication is that the adjusted odds ratio, unlike the crude odds ratio, will indicate a spurious noncausal association between E and D. A similar argument can be used in figure 2, where C is not a common effect of E and D but of U_1 and D. Conditioning on C will induce a spurious association between U_1 and D; since U_1 is associated with E (because U_1 is a cause of E), conditioning on C will thus induce a spurious association between D and E. Similar arguments apply to figures 3 and 4 (in the latter, there would be bias even if C occurred before E).

Thus, C is a confounder and one needs to adjust for it in figures 5–8, but it is a nonconfounder and one should not adjust for it in figures 1–4. This definition of confounding and confounders is not based on the statistical associations found in our data but rather on qualitative background knowledge about the causal structure of the problem under study, which we encoded in causal diagrams. This approach contrasts with the causally blind strategies 1 and 2, which use only statistical associations to decide whether C should be adjusted for, and with strategy 3, which uses statistical conditions supplemented with partial but insufficient a priori causal information.

In fact, the conditions implied by strategies 1–3 hold true for all figures 1–8. However, in figure 9, C would be excluded as a confounder by the causal restriction of strategy 3 (that C is not in the causal pathway). The additional causal restriction that C must be a causal risk factor or a marker for a causal risk factor (8) further restricts the set of possible causal structures in which C may be a confounder to those in figures 3–8. Another widely recognized restriction is that the potential confounder cannot be affected by either exposure or outcome (1, 6, 16, 17), which excludes figures 1–3. As more causal restrictions are applied, fewer causal diagrams are consistent with C's being a confounder. Whether C is or is not truly a confounder depends on the causal structure of the problem under study. No generally applicable statistical approaches will substitute for using a priori causal knowledge to characterize such structure.

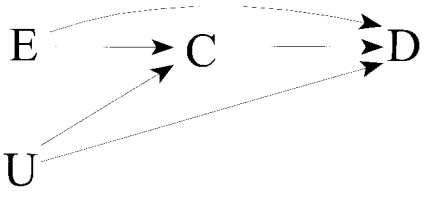


FIGURE 9. Antiepileptic drugs (E) may lower plasma folate levels (C) and may cause birth defects (D) through mechanisms not related to folate. Alcohol intake (U) may also decrease folate levels but may be associated with birth defects through other paths.

AN EXAMPLE FROM BIRTH DEFECTS EPIDEMIOLOGY

Supplementation with 0.4 mg of folic acid per day around the time of conception has been shown to decrease the risk of neural tube defects in randomized experiments (18) and observational studies (19, 20). We examined this relation using data from the Slone Epidemiology Unit Birth Defects Study. Since 1976, mothers of malformed children born in the greater metropolitan areas of Boston, Massachusetts, Philadelphia, Pennsylvania, and Toronto, Canada, have been interviewed about pregnancy events and exposures (19). For this analysis we included mothers of infants with neural tube defects as cases ($D = 1$) and mothers of infants with birth defects thought to be unrelated to folic acid as controls ($D = 0$). The exposure of interest was categorized as the presence ($E = 1$) or the absence ($E = 0$) of daily supplementation with folic acid during the first and second months after the last menstrual period. This period encompasses neural tube development. Because information on the folic acid-containing multivitamin was not collected before 1992, and because folate fortification of cereal grains began in 1998 (21), we restricted the analysis to infants born between 1992 and 1997. A third dichotomous variable C was also measured; for pedagogic reasons, its identity will be withheld until later. However, based on subject matter knowledge, C is known not to be in the causal pathway from exposure to disease. We use prevalence odds ratios and their 95 percent confidence intervals as estimators of causal effect. For simplicity, we assume that all variables are perfectly measured.

We found that 18 percent of cases and 25 percent of controls used folic acid daily during the exposure period (table 3). The crude (unadjusted) OR_{ED} was 0.65 (95 percent confidence interval (CI): 0.45, 0.94), which approximates the crude risk ratio. OR_{ED} can be obtained directly from table 3 or as $\exp(\beta_1)$ from the logistic model $\text{logit Pr}(D = 1|E) = \beta_0 + \beta_1 E$. Table 4 displays the data by levels of C. To estimate the adjusted odds ratio, we stratify (i.e., condition) on all levels of the third variable C, compute the stratum-specific odds ratio, and then calculate a pooled summary

TABLE 3. Periconceptional folic acid supplementation ($E = 1$) and neural tube defects ($D = 1$), Slone Epidemiology Unit Birth Defects Study, 1992–1997

	$D = 1$	$D = 0$
$E = 1$	43	239
$E = 0$	194	704

TABLE 4. Periconceptional folic acid supplementation ($E = 1$) and neural tube defects ($D = 1$), stratified by the covariate C, Slone Epidemiology Unit Birth Defects Study, 1992–1997

	$C = 1$		$C = 0$	
	$D = 1$	$D = 0$	$D = 1$	$D = 0$
$E = 1$	19	8	24	231
$E = 0$	100	46	94	658

measure across strata (e.g., using the Mantel-Haenszel method). We did not detect heterogeneity of the odds ratio between the two strata defined by C ($p = 0.43$ from the Breslow-Day test for homogeneity), so for our purposes we assume that the no interaction logistic model logit $\Pr(D = 1|E, C) = \beta_0 + \beta_1 E + \beta_2 C$ is correct. The $OR_{ED|C}$ adjusted by C ($\exp(\beta_1)$) was 0.80 (95 percent CI: 0.53, 1.20).

Which analysis is more appropriate, the crude or the adjusted? We first consider the three common strategies described above:

1. *Automatic variable selection*. We force exposure E as a covariate in the logistic model with D as the outcome. We consider an automatic forward selection procedure, available in most standard statistical software packages in which the variable C is added if the p value associated with its parameter estimate is less than 0.10. As the p value in our data set is less than 0.001, variable C is selected.
2. *Relative change in estimate greater than 10 percent*. The adjusted $OR_{ED|C}$ was 0.80, a 23 percent relative change with respect to the crude $OR_{ED} = 0.65$, so the adjusted estimate will be selected.
3. *Standard rules for confounding*. First, we check that C is associated with E in the population; in our data $OR_{CE|D=0}$ is 0.50 (95 percent CI: 0.23, 1.07). Second, we check that C is associated with D within the unexposed; in our data $OR_{CD|E=0}$ is 15.22 (95 percent CI: 10.09, 22.95). Third, we need to exclude the possibility that C may be in the causal pathway between E and D. The data by themselves are never sufficient to rule out the possibility; however, in our case it was known that C was not plausibly on the causal pathway. Because all three conditions are met, the adjusted estimate will be selected.

All three strategies require adjustment for C.

We have not as yet unveiled the variable encoded by C in table 4 in order to emphasize that no additional information about C beyond that contained in the data is required by stategies 1 and 2, and only limited external background information is required by strategy 3. However, we have seen in the previous section that knowledge of the causal structure is crucial if we are to decide whether C is a confounder and needs to be adjusted for. In fact, the adjusted $OR_{ED|C}$ is biased in four of our diagrams.

In our example, the variable C stands for the event that pregnancy ends either in stillbirth or therapeutic abortion. Should we regard C as a confounder? To answer this question, we would need the true, but possibly unknown, underlying causal structure. Most investigators would agree that figures 1–4 are more likely to represent the true causal structure than figures 5–8. In fact, figures 5–7 are rapidly eliminated because they assume that C occurs before the exposure E or the outcome D.

Yet it is not uncommon to find epidemiologic analyses that adjust for stillbirth/induced abortion, either by stratification or by restricting the analysis to livebirths. This practice is often the unintended consequence of the difficulty of identifying stillbirths and/or ascertaining their maternal exposures. Similarly, analyses of the effects of prenatal exposures frequently adjust for variables, such as maternal weight gain during pregnancy, gestational age, or birth weight, that are likely to be affected by either the exposure or the outcome. The decision to adjust is usually based on statistical criteria only. (Here we assume that the goal is to estimate the total effect. The section "Adjusting for Variables Affected by Exposure and Causal Diagrams" below discusses direct effects.) How much bias is introduced by this decision depends on the strength of the statistical associations between the potential confounder and exposure and outcome (22, 23). In our study, the apparent bias was moderate despite the fact that the association between the potential confounder and the outcome was very strong.

For expositional purposes, we have assumed throughout that there were no other confounders of the causal effect of E on D other than possibly the covariate C. This is not a realistic assumption, but it was useful to simplify the problem. In a more realistic analysis in which we adjusted for region, maternal age, whether the pregnancy was planned, and maternal education, the odds ratio was 0.72 (95 percent CI: 0.49, 1.05).

In general, crude and adjusted odds ratios can differ not because of confounding but because of the noncollapsibility property of the odds ratio; that is, the crude odds ratio does not necessarily equal a common stratified odds ratio even if the exposure and stratifying factor are unassociated in the population (24, 25). This is an additional reason to avoid the change-in-estimate method for the odds ratio. We did not consider this issue because we assumed that, in our study, the disease was rare so the odds ratio was approximately equal to the risk ratio, which is a collapsible measure.

SELECTION BIAS, RECALL BIAS, AND CAUSAL DIAGRAMS

Causal diagrams are useful to represent biases other than confounding, such as selection and recall bias (5). In research into birth defects, investigators sometimes restrict the analysis to liveborn infants, which has long been recognized as a potential source of bias (26, 27). We now use causal diagrams to show how this bias is introduced. First, note that restricting to livebirths is equivalent to conditioning on a particular value of the variable C's encoding stillbirth/abortion. If the true causal structure is represented by

any of figures 1–4, then C is an effect of either exposure E or outcome D, or else it shares a common cause with them. Conditioning on the collider C would produce a biased odds ratio, as described above, for example, with selection of liveborn infants. In fact, studies of birth defects are potentially subject to some degree of selection bias, because spontaneous abortions in which the presence or absence of neural tube defects is indeterminate are not included. (This may not be a problem if the goal is to estimate the public health burden due to the exposure, rather than its causal effect among all conceptuses, which we have taken to be the causal contrast of interest.) A similar selection bias may occur when C stands for participation in the study, and the analysis is restricted to those who agreed to participate. This latter is a particularly difficult bias to control.

Selection bias induces noncomparability or, equivalently, lack of exchangeability of the exposed and the unexposed, even if they were comparable before the selection. Many authors use noncomparability as a synonym for confounding (7). We are being careful to separate confounding due to unmeasured common causes from noncomparability induced by selection.

We have so far made the simplifying assumption that exposure is perfectly measured before the outcome occurs. However, case-control studies often ascertain the exposure E' after the outcome is known, as represented in figures 10 and 11. In figure 10, E' is determined by the actual exposure E but not by the outcome D, so misclassification is nondifferential. Any association between E' and D ($OR_{E'D}$) is therefore due solely to the causal effect of E on D. In figure 11, E' is determined by the actual exposure E and the outcome D, representing, for example, a setting in which there is recall bias because mothers of infants with birth defects have more complete recall of the exposure than mothers of healthy infants. In figure 11, $OR_{E'D}$ can differ from one even under the causal null hypothesis, because part of the association between E' and D is spurious as a result of the causal effect of D on E'. Our study used mothers of infants with other birth defects in an attempt to eliminate the arrow between D and E'. Another example of the situation depicted in figure 11 occurs when studying the effect of an exposure

FIGURE 10. Case-control studies of birth defects (D) often rely on maternal recall (E') of antenatal exposure (E).

FIGURE 11. When mothers of malformed infants (D) recall the information more completely (or less completely) than mothers of controls, D will influence E' (maternal recall). Study designs that use malformed controls try to avoid the arrow between D and E'. E, antenatal exposure.

through a biomarker. For example, women with disease-associated weight loss could have altered blood levels of a pesticide residue due to mobilization of residues stored in fat.

ADJUSTING FOR VARIABLES AFFECTED BY EXPOSURE AND CAUSAL DIAGRAMS

An inspection of the causal diagram in figure 9 reveals the two main reasons why adjustment for a variable on the causal pathway is discouraged in the epidemiologic literature (1–3). First, when one is interested in the overall effect of E on D, one does not want to adjust for C if part of the effect of E on D is mediated through C, because the adjusted $OR_{ED|C}$ will not reflect an overall effect. Second, adjusting for (conditioning on) the common effect (collider) C would create a spurious association between its causes, E and the unmeasured factor U, and therefore between E and D. This will produce a non-null noncausal association between E and D even if the effect of E on D were entirely through C (i.e., no direct effect) or C had no causal effect on D whatsoever (28).

The second argument makes clear that, even to estimate the direct effect of E (not mediated through C) on D, 1) it is not valid to adjust for C when there is an unmeasured common cause of C and D, and 2) C's being on a causal pathway from E to D is not a necessary condition for this spurious association to appear. The source of the problem is that C is a marker for an unmeasured causal risk factor U for the outcome, and C is either causally affected by exposure (figure 9) or shares common causes with the exposure (similar to figure 4).

CONCLUSION

We have argued that knowledge of the causal structure is a prerequisite to accurately label a variable as a confounder. Taken literally, this statement may impose such an unrealistically high standard on the epidemiologist that many studies simply could not be done at all. Instead, we wish to emphasize that causal inference from observational data requires prior causal assumptions or beliefs, which must be derived from subject-matter knowledge, not from statistical associations detected in the data.

Our goal was to highlight potential inconsistencies between beliefs and actions in data analysis. In general, investigators should not adjust for a variable C unless they believe it may be a confounder. At the very least, researchers should generally avoid stratifying on variables affected by either the exposure or the outcome. Of course, thoughtful and knowledgeable epidemiologists could believe that two or more causal structures, possibly leading to different conclusions regarding confounding, are equally plausible. In that case they should perform multiple analyses and explicitly state the assumptions about causal structure required for the validity of each. One can never be certain that the set of causal structures under consideration includes the true one; this uncertainty and the attendant model uncertainty are unavoidable with observational data.

Causal diagrams are a useful way to summarize, clarify, and communicate one's qualitative beliefs about the causal

structure. The use of causal diagrams in epidemiology has been proposed by Greenland et al. (4). The main advantage of this graphical method is that, while being a natural and simple way to approach causal inference from observational data, it is also rigorous, being mathematically identical to Robins' "g-computation theory" (29–31).

We have used causal diagrams to describe three possible sources of statistical association between two variables: cause and effect, sharing of common causes, and calculation of the association within levels of a common effect. There is confounding when the association between exposure and disease includes a noncausal component attributable to their having an uncontrolled common cause. There is selection bias when the association between exposure and disease includes a noncausal component attributable to restricting the analysis to certain level(s) of a common effect of exposure and disease or, more generally, to conditioning on a common effect of variables correlated with exposure and disease. In either case, the exposed and the unexposed in the study are not comparable, or exchangeable, which is the ultimate source of the bias. Statistical criteria are insufficient to characterize either confounding or selection bias.

ACKNOWLEDGMENTS

This work was supported in part by National Institutes of Health grant R01-AI32475. The Slone Epidemiology Unit Birth Defects Study was supported in part by National Institute of Child Health and Human Development grant HD27697 and National Heart, Lung, and Blood Institute grant HL50763. Additional support for the Slone Epidemiology Unit Birth Defects Study was provided by Hoechst Marion Roussel, Inc. (Kansas City, Missouri), Pfizer, Inc. (New York, New York), the Glaxo-Wellcome Company (Research Triangle Park, North Carolina), and Rhone Pouleuc Rorer (College Park, Pennsylvania).

The authors thank James Robins for inspiring them to write this article and for his many comments, which led to a substantial improvement in the manuscript. The authors also thank Sander Greenland for his detailed suggestions.

REFERENCES

1. Greenland S, Neutra R. Control of confounding in the assessment of medical technology. Int J Epidemiol 1980;9:361–7.
2. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. Am J Epidemiol 1986; 123:392–402.
3. Weinberg CR. Towards a clearer definition of confounding. Am J Epidemiol 1993;137:1–8.
4. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999;10:37–48.
5. Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiology 2001;12:313–20.
6. Rothman KJ, Greenland S. Modern epidemiology. 2nd ed. Philadelphia, PA: Lippincott-Raven, 1998.
7. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 1986;15:413–19.
8. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health 1989;79:340–9.
9. Kleinbaum DG, Kupper LL, Muller KE, et al. Applied regression analysis and other multivariable methods. 3rd ed. Pacific Grove, CA: Duxbury Press, 1998.
10. Grayson DA. Confounding confounding. Am J Epidemiol 1987;126:546–53.
11. Szklo M, Nieto FJ. Epidemiology: beyond the basics. Gaithersburg, MD: Aspen Publishers, Inc, 1999.
12. Pearl J. Causality. Cambridge, United Kingdom: Cambridge University Press, 2000.
13. Pearl J. Causal diagrams for empirical research. Biometrika 1995;82:669–710.
14. Spirtes P, Glymour C, Scheines R. Causation, prediction, and search. Lecture notes in statistics 81. New York, NY: Springer-Verlag, 1993.
15. Hernán MA, Robins JM. Method for conducting sensitivity analysis. (Letter). Biometrics 1999;55:1316–18.
16. Cox DR. Planning of experiments. New York, NY: John Wiley & Sons, 1958.
17. Miettinen O. Confounding and effect-modification. Am J Epidemiol 1974;100:350–3.
18. MRC Vitamin Study Research Group. Prevention of neural tube defects: results of the Medical Research Council Vitamin Study. Lancet 1991;338:131–7.
19. Werler AM, Shapiro S, Mitchell AA. Periconceptional folic acid exposure and risk of occurrent neural tube defects. JAMA 1993;269:1257–61.
20. Berry RJ, Li Z, Erickson JD, et al. Prevention of neural-tube defects with folic acid in China. N Engl J Med 1999;341: 1485–90.
21. Use of folic acid-containing supplements among women of childbearing age—United States, 1997. MMWR Morb Mortal Wkly Rep 1998;47:131–4.
22. Bross IDJ. Spurious effects from an extraneous variable. J Chronic Dis 1966;19:637–47.
23. Walker AM. Observation and inference: an introduction to the methods of epidemiology. Newton Lower Falls, MA: Epidemiology Resources, Inc, 1991.
24. Miettinen OS, Cook EF. Confounding: essence and detection. Am J Epidemiol 1981;114:593–603.
25. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Stat Sci 1999;14:29–46.
26. Khoury MJ, Flanders WD, James LM, et al. Human teratogens, prenatal mortality, and selection bias. Am J Epidemiol 1989;130:361–70.
27. Hook EB, Regal RR. Conceptus viability, malformation, and suspect mutagens or teratogens in humans. The Yule-Simpson paradox and implications for inferences of causality in studies of mutagenicity or teratogenicity limited to human livebirths. Teratology 1991;43:53–9.
28. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology 1992;3:143–55.
29. Robins JM. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. J Chronic Dis 1987;40(suppl 2):139s–61s.
30. Robins JM. Comment. Biometrika 1995;82:695–8.
31. Robins JM. Causal inference from complex longitudinal data. In: Berkane M, ed. Latent variable modeling with applications to causality. New York, NY: Springer-Verlag, 1997:69–117.
32. Scholl TO, Johnson WG. Folic acid: influence on the outcome of pregnancy. Am J Clin Nutr 2000;71(suppl):1295s–303s.
33. Mili F, Edmonds LD, Khoury MJ, et al. Prevalence of birth defects among low-birth-weight infants. A population study. Am J Dis Child 1991;145:1313–18.
34. Mathews F, Yudkin P, Smith R, et al. Nutrient intakes during pregnancy: the influence of smoking status and age. J Epidemiol Community Health 2000;54:17–23.
35. Roth J, Hendrickson J, Schilling M, et al. The risk of teen mothers having low birth weight babies: implications of recent medical research for school health personnel. J Sch Health 1998;68:271–5.

36. Werler MM, Louik C, Shapiro S, et al. Prepregnant weight in relation to risk of neural tube defects. JAMA 1996;275:1089–92.
37. Pastuszak A, Bhatia D, Okotore B, et al. Preconception counseling and women's compliance with folic acid supplementation. Can Fam Physician 1999;45:2053–7.
38. Lie RT, Wilcox AJ, Skjaerven R. A population-based study of the risk of recurrence of birth defects. N Engl J Med 1994; 331:1–4.
39. Van der Put NMJ, Gabreels F, Stevens EMB, et al. A second common mutation in the methylenetetrahydrofolate reductase gene: an additional risk factor for neural-tube defects. Am J Hum Genet 1998;62:1044–51.
40. Botto LD, Yang Q. 5,10-Methylenetetrahydrofolate reductase gene variants and congenital anomalies: a HuGE review. Am J Epidemiol 2000;151:862–77.

APPENDIX

We have used causal DAGs to encode our qualitative a priori assumptions about the underlying biologic mechanisms. These diagrams consist of nodes (variables) and directed edges (arrows). The absence of an arrow between two variables indicates that the investigator believes there is no direct effect of one variable on the other (i.e., a causal effect not mediated through other variables in the DAG). The presence of an arrow indicates that the investigator is unable to assume the absence of a direct effect of one variable on the other. Further, conditioning on its direct causes, each variable is statistically independent of all the variables it does not cause. DAGs are acyclic because the arrows never point from a given variable to any other variable in its past. If, for example, one is interested in representing the causal interplay between nutritional status and infection in children, the DAG could contain an arrow from a variable that represents nutritional status at time t to another one that represents infection at time $t + 1$ and a second arrow from infection at time $t + 1$ to nutritional status at time $t + 2$. We say that a DAG is causal if the common causes of any pair of variables in the graph are also in the DAG.

Our neural tube defects example may be partially represented by the DAG in figure 1. Subject-matter knowledge suggests that neural tube defects (D) are a direct cause of stillbirths/therapeutic abortions (C), and that folic acid supplementation (E) may prevent stillbirths/abortions (C) through its effects on birth defects other than neural tube defects. (In our study, stillborn infants and fetuses therapeutically aborted because of a malformation are identified through review of admissions and discharges at major referral hospitals and clinics and through regular contact with newborn nurseries in community hospitals. Medical records and autopsies, if available, are then reviewed to ascertain birth defects.)

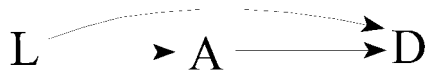
To draw a more realistic DAG according to our causal assumptions, we added the following variables: pregnancy planning (because women who are trying to conceive a pregnancy often take prenatal vitamins in preparation, and they may take better care of themselves in general, which in turn may affect the outcome), maternal education (for similar reasons to pregnancy planning), region (because both exposure and outcome may display geographic variations mediated through socioeconomic, behavioral, dietary, ethnic, cultural, and other factors), and maternal age (a risk factor for some birth defects, maternal age may also affect behavior regarding folic acid supplementation). In our hypothesized DAG, these preexposure variables would be the origin of arrows pointing to E and D. Note that the inclusion of these variables in the DAG does not imply that we are certain about the existence of their causal connections with E and D (e.g., maternal age may not affect the risk of neural tube defects relative to that of other birth defects), but that we are not willing to assume a priori that those connections are absent.

The arguments we used to support the statistical statements derived from causal DAGs were heuristic and relied on our causal intuitions. These arguments, however, have been formalized and mathematically proven (12–14). Here we present a brief overview of a graphical method called "d-separation" ("d-" stands for directional) (12, 13) that allows us to determine whether two given variables are (marginally or conditionally on other variables) independent.

The method of d-separation consists of a set of graphical rules to decide whether two variables are d-separated, which implies that they are independent, or are d-connected, which generally implies that they are not independent. If two variables are d-separated without conditioning on any other variables in the DAG, then they are marginally independent. If two variables are d-separated after conditioning on a set of third variables, then they are conditionally independent (i.e., independent within every joint stratum of the third variables). To explain the method we first need to define the terms "path" and "blocked path." A path is any arrow-based route between two variables in the graph. We define each path to be either blocked or open according to the following graphical rules.

Rule 1. If there are no variables being conditioned on, a path is blocked if and only if two arrowheads on the path collide at some variable on the path. For example, in appendix figure 1, the path $L \rightarrow A \rightarrow D$ is open, whereas the path $A \rightarrow D \leftarrow L$ is blocked because two arrowheads on the path collide in D. We call D a collider on the path $A \rightarrow D \leftarrow L$.

Rule 2. Any path that contains a noncollider that has been conditioned on is blocked. For example, in appendix figure 2, the path between L and D is blocked after conditioning on



APPENDIX FIGURE 1. A partial representation of our study on neural tube defects.

APPENDIX FIGURE 2. The path between L and D is blocked after conditioning on A.

A. We use a square box around a variable to indicate that we are conditioning on it.

Rule 3. A collider that has been conditioned on does not block a path. For example, in appendix figure 3, the path between *L* and *A* is open after conditioning on *D*.

Rule 4. A collider that has a descendant that has been conditioned on does not block a path. For example, in appendix figure 4, the path between *L* and *A* is open after conditioning on *S*, a descendant of the collider *D*.

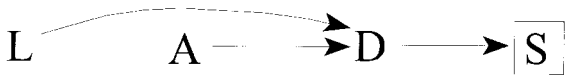
Rules 1–4 can be summarized as follows. A path is blocked if and only if it contains a noncollider that has been conditioned, or it contains a collider that has not been conditioned on and has no descendants that have been conditioned on.

Two variables are d-separated if all paths between them are blocked (otherwise they are d-connected). Thus, *A* and *L* are not marginally independent (d-connected) in appendix figure 1 because there is one open path between them ($L \rightarrow A$), despite the other path ($A \rightarrow D \leftarrow L$)'s being blocked by the collider *D*. In appendix figure 3, however, *A* and *L* are marginally independent (d-separated) because the only path between them is blocked by the collider *D*. In appendix figure 2, we conclude that *D* is conditionally independent of *L*, given *A*. From appendix figure 3 we infer that *L* is not conditionally independent of *A*, given *D*. Appendix figure 4 includes the variable *S*, representing the symptoms caused by the disease. If conditioning on *D* opens the path and therefore creates an association between *L* and *A*, then conditioning on an effect of the disease (*S*) also creates an association between *L* and *A*. In general, the farther the descendant of the collider is in the chain of causation, the weaker this association will be.

Some conclusions that follow from the method of d-separation are that causes (ancestors) are not independent of their effects (descendants) and vice versa, and that generally two variables are associated if they share a common cause. Another important conclusion is that sharing a common effect does not imply that two causes are associated. Intuitively, whether two variables (the common causes) are correlated cannot be influenced by an event in the future (their effect) (4), but two causes of a given effect generally become associated once we stratify on the common effect.

Finally, we explain why two variables that are not d-separated may actually be statistically independent. The reason is that it is logically possible that causal effects in opposite directions may exactly cancel out. For example, in appendix figure 5, if the arrow $L \rightarrow D$ is causative for half the population and preventive for the other half, and if the magnitude of the causative and protective effects is exactly the same, then *L* and *D* will be marginally independent despite the fact that they are not d-separated. Because exact cancellation of causal effects is probably a very rare event in epidenidologic applications, d-separation and independence may be treated in practice as equivalent concepts with little risk. In the probably rare occasions in which two variables are simultaneously d-connected and statistically independent, we say that the joint distribution of the variables in the DAG is not faithful to the DAG (14).

APPENDIX FIGURE 3. The path between *L* and *A* is open after conditioning on *D*.



APPENDIX FIGURE 4. The path between *L* and *A* is open after conditioning on *S*.

APPENDIX FIGURE 5. *A* and *D* may be unassociated under nonfaithfulness.