CHAPTER 9

# Enhancing Causal Inference in Observational Studies

Thomas B. Newman, Warren S. Browner, and Stephen B. Hulley

Most observational studies are designed to suggest that a predictor may be a cause of an outcome, for example, that eating broccoli may reduce the risk of colon cancer. (Exceptions are studies of diagnostic and prognostic tests, discussed in Chapter 12.) Causal associations between a predictor and an outcome are important because they can provide insights into the underlying biology of a disease, identify ways to reduce or prevent its occurrence, and even suggest potential treatments.

However, not every association that is found in an observational study represents cause–effect. Indeed, there are four other general explanations for an association between a predictor and an outcome in an observational study (Table 9.1). Two of these, **chance** and **bias**, create spurious associations between the predictor and the outcome in the study sample that do not exist in the population. Two others, **effect–cause** and **confounding**, create real associations in the population, but these associations are not causal in the direction of interest. Establishing that cause–effect is the most likely explanation for an association requires demonstrating that these other explanations are unlikely.

We typically quantify the causal effect of a predictor variable on an outcome using a measure of association, such as a risk ratio or odds ratio. For example, suppose that a study reveals that coffee drinking has a risk ratio of 2.0 for myocardial infarction (MI). One possibility—presumably the one that the investigator found most interesting—is that drinking coffee doubles the risk of MI. Before reaching this conclusion, however, the four rival explanations must be considered and dismissed.

With chance and bias, coffee drinking was associated with a doubled risk of MI in the study, but that association is not actually present in the population. Thus, chance and bias are explanations for spurious (i.e., not real) associations in a study.

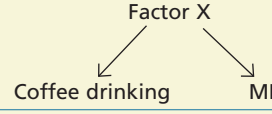
The other two alternatives—effect–cause and confounding—are true biological phenomena, which means that coffee drinkers in the population really do have twice the risk of MI. However, that increased risk is not due to a cause–effect relationship. In one situation, the association is due to effect–cause: having an MI causes people to drink more coffee. (This is just cause and effect in reverse.) The final possibility, confounding, occurs when a third factor, such as personality type, causes both coffee drinking and MI.

In the remainder of the chapter, we will discuss strategies for estimating and minimizing the likelihood of these four alternative explanations for finding an association in an observational study. These strategies can be used while designing a study or when analyzing its results. While this book emphasizes research design, understanding the analytic options can influence the choice of design, so both topics will be considered in this chapter.

## SPURIOUS ASSOCIATIONS DUE TO CHANCE

Suppose that in reality there is no association between coffee drinking and MI among members of a population, 45% of whom drink coffee. If we were to select 20 cases with MI and 20 controls, we would expect that about 9 people in each group (45% of 20) would drink coffee.

**TABLE 9.1 THE FIVE EXPLANATIONS FOR AN OBSERVED DOUBLING OF THE RISK OF MI ASSOCIATED WITH COFFEE DRINKING**

| EXPLANATION | TYPE OF ASSOCIATION | WHAT'S REALLY GOING ON IN THE POPULATION? | CAUSAL MODEL |
|---|---|---|---|
| 1. Chance (random error) | Spurious | Coffee drinking and MI are not related. | — |
| 2. Bias (systematic error) | Spurious | Coffee drinking and MI are not related. | — |
| 3. Effect–cause | Real | MI is a cause of coffee drinking. | MI → Coffee drinking |
| 4. Confounding | Real | A third factor causes both coffee drinking and MI. |   |
| 5. Cause–effect | Real | Coffee drinking is a cause of MI. | Coffee drinking → MI |

However, *by chance alone,* we might enroll 12 coffee drinkers among the 20 MI cases, but only 6 in the 20 controls. If that happened, we would observe a spurious association between coffee consumption and MI in our study.

Chance is sometimes called **random error**, because it has no underlying explanation. When an association due to random error is statistically significant, it's known as a **type I error** (Chapter 5).

Strategies for reducing random error are available in both the design and analysis phases of research (Table 9.2). *Design strategies*, such as increasing the **precision of measurements** and increasing the **sample size**, are discussed in Chapters 4 and 6, respectively. The *analysis strategy* of calculating ***P* values** and **confidence intervals** helps the investigator quantify the magnitude of the observed association in comparison with what might have occurred by chance

**TABLE 9.2 STRENGTHENING THE INFERENCE THAT AN ASSOCIATION IS DUE TO CAUSE–EFFECT BY REDUCING AND EVALUATING THE LIKELIHOOD OF SPURIOUS ASSOCIATIONS**

| TYPE OF SPURIOUS ASSOCIATION | DESIGN PHASE (HOW TO PREVENT THE RIVAL EXPLANATION) | ANALYSIS PHASE (HOW TO EVALUATE THE RIVAL EXPLANATION) |
|---|---|---|
| Chance (due to random error) | Increase sample size and other strategies to increase precision (Chapters 4 and 6) | Calculate *P* values and confidence intervals and interpret them in the context of prior evidence (Chapter 5) |
| Bias (due to systematic error) | Carefully consider the potential consequences of each difference between the research question and the study plan (Figure 9.1); alter the study plan if necessary | Check consistency with other studies (especially those using different designs) |
| | Collect additional data that will allow assessment of the extent of possible biases | Analyze additional data to see if potential biases have actually occurred |
| | Do not use variables affected by the predictor of interest as inclusion criteria or matching variables | Do not control for variables affected by your predictor variable |

alone. For example, a *P* value of 0.10 indicates that chance alone could cause a difference at least as large as the investigators observed about 10% of the time. Even more useful than *P* values, confidence intervals show the possible values for statistics describing an association that fall within the range of random error estimated in the study.

## SPURIOUS ASSOCIATIONS DUE TO BIAS

Many kinds of bias—systematic error—have been identified, and dealing with some of them is a major topic of this book. Along with the specific strategies described in Chapters 3, 4, 7 and 8, we now add a general approach to reducing the likelihood of bias.

### Minimizing Bias

As was discussed in Chapter 1, there are almost always differences between the original research question and the one that is actually answered by the study. Those differences reflect the compromises that were made for the study to be feasible, as well as mistakes in the design or execution of the study. Bias occurs when those differences cause the answer provided by the study to differ from the right answer to the research question. Strategies for minimizing bias are available in both the design and analysis phases of research (Table 9.2).

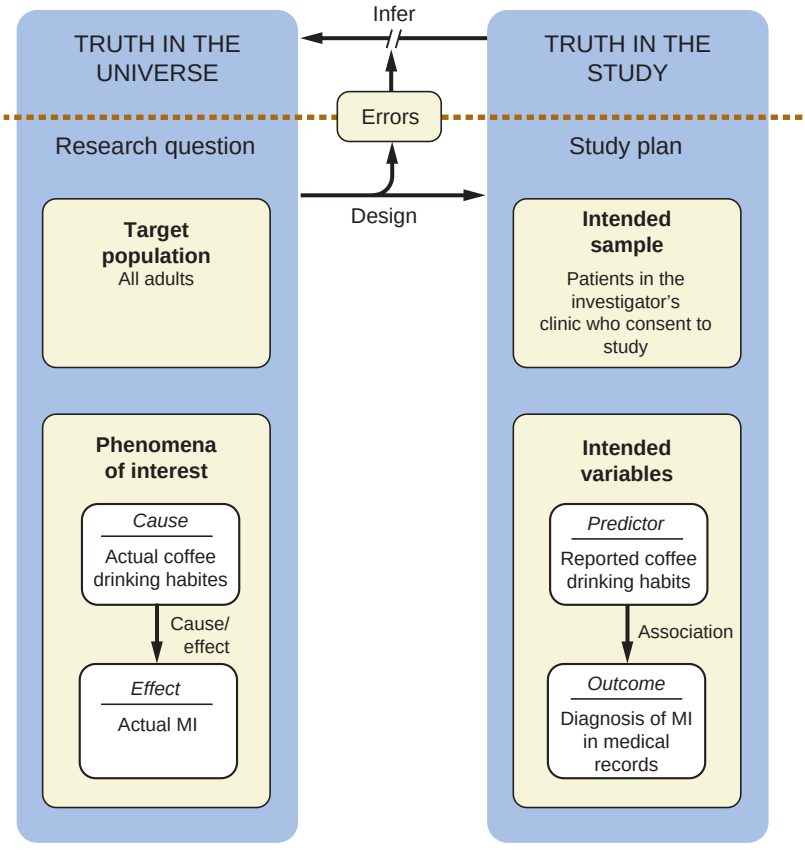
- **Design phase.** Begin by writing the research question next to the study plan, as in Figure 9.1. Then think through the following three concerns as they pertain to the research question:
  1. Do the **samples** of study subjects (e.g., cases and controls, or exposed and unexposed subjects) represent the population(s) of interest?
  2. Do the **measurements** of the **predictor variables** represent the predictors of interest?
  3. Do the **measurements** of the **outcome variables** represent the outcomes of interest?

  For each question answered "No" or "Maybe not," consider whether the bias applies similarly to one or both groups studied (e.g., cases and controls, or exposed and unexposed subjects) and whether the bias is likely to be large enough to affect the answer to the research question.

  To illustrate this with our coffee and MI example, consider a case–control study in which the control subjects are sampled from patients hospitalized for diseases other than MI. If many of these patients have chronic illnesses that led them to reduce their coffee intake, the sample of controls will not represent the target population from which the MI cases arose: There will be a shortage of coffee drinkers. And if esophageal spasm, which can be exacerbated by coffee, is misdiagnosed as MI, a spurious association between coffee and MI could be found because the measured outcome (diagnosis of MI) did not accurately represent the outcome of interest (actual MI).

  The next step is to think about possible strategies for preventing each potential bias, such as selecting more than one control group in a case–control study (Chapter 8) or the strategies for reducing measurement bias described in Chapter 4. In each case, judgments are required about the likelihood of bias and how easily it could be prevented with changes in the study plan. If the bias is easily preventable, revise the study plan and ask the three questions again. If the bias is not easily preventable, decide whether the study is still worth doing by judging the likelihood of the potential bias and the degree to which it will distort the association you are trying to estimate.

  Potential biases may either be unavoidable or costly to prevent, or it may be uncertain to what extent they will be a problem. In either case, the investigator should consider designing the study to collect additional data that will allow an assessment of the seriousness of the biases. For example, if the investigator is concerned that the cases in a study of pancreatic cancer may over-report recent exposures to toxic chemicals (perhaps because these individuals are searching desperately for an explanation for why they have pancreatic cancer),




■ **FIGURE 9.1** Minimizing bias by carefully considering differences between the research question and the study plan.

they could also be asked about exposures (such as coffee drinking!) that previous studies have shown have no effect on the risk of pancreatic cancer. If the investigator is concerned that a questionnaire does not accurately capture coffee drinking (perhaps because of poorly worded questions), she could assign a blinded interviewer to question a subset of the cases and controls to determine the agreement with their questionnaire responses. Similarly, if she is concerned that rather than causing MI, coffee increases survival among MI patients (which could lead to coffee drinkers being over-represented in a sample of MI survivors), the investigator could identify MI patients who died and interview their surviving spouses about their previous coffee-drinking habits.

- **Analysis phase.** Once the data have been collected, the goal shifts from minimizing bias to assessing its likely severity. The first step is to analyze data that have been collected for that purpose. For example, an investigator anticipating imperfect memory of coffee-drinking habits may have included questions about how sure the cases and controls are of their answers. The association between coffee drinking and MI could then be examined after stratifying on certainty about coffee intake, to see whether the association is stronger among those more certain of their exposure history.

The investigator can also look at the results of other studies. If the conclusions are consistent, the association is less likely to be due to bias. This is especially true if the other studies have used different designs and are therefore unlikely to share the same biases. However, in many situations the potential biases turn out not to be a major problem. The decision on how

vigorously to pursue additional information and how best to discuss these issues in reporting the study are matters of judgment for which it is helpful to seek advice from colleagues.

## REAL ASSOCIATIONS OTHER THAN CAUSE–EFFECT

In addition to chance and bias, the two types of associations that are real but do not represent cause–effect must be considered (Table 9.3).

### Effect–Cause

One possibility is that the cart has come before the horse—the outcome has caused the predictor. Effect–cause is often a problem in cross-sectional and case–control studies: Does a sedentary lifestyle cause obesity, or vice versa? Effect–cause can also be a problem in case-crossover studies. For example, in the study of mobile phone use and motor vehicle accidents described in Chapter 8 (1), a car crash could cause the driver to make a mobile phone call reporting the crash, rather than the crash having been caused by an inattentive driver. To address this possibility, the investigators asked drivers about phone use before and after the crash, and verified the responses using phone records.

Effect–cause is less commonly a problem in cohort studies of disease causation because risk factor measurements can be made among subjects who do not yet have the disease. Even in cohort studies, however, effect–cause is possible if the disease has a long latent period and those with subclinical disease cannot be identified at baseline. For example, Type II diabetes is associated with subsequent risk of pancreatic cancer. Some of this association may well be effect–cause, because pancreatic cancer could affect the pancreatic islet cells that secrete insulin, thus causing diabetes. Consistent with effect–cause, the risk of pancreatic cancer is highest just after diabetes is diagnosed (2). The association diminishes with the duration of diabetes, but some association persists even 4 years or more after the onset of diabetes (2–4) suggesting that at least some of the relationship may be cause–effect.

This example illustrates a general approach to ruling out effect–cause: looking for a diminution in the association with increasing time between the presumed cause and its effect. A second approach is to assess the biologic plausibility of effect–cause versus cause–effect. In this example effect-cause was plausible because pancreatic cancer could damage the pancreas, but the observation that having diabetes for more than 10 years is associated with an increased risk of a variety of other cancers as well as pancreatic cancer (4) increases the biologic plausibility of diabetes causing pancreatic cancer, rather than being only one of its effects.

**TABLE 9.3 STRENGTHENING THE INFERENCE THAT AN ASSOCIATION HAS A CAUSE–EFFECT BASIS: RULING OUT OTHER REAL ASSOCIATIONS**

| TYPE OF REAL ASSOCIATION | DESIGN PHASE (HOW TO PREVENT THE RIVAL EXPLANATION) | ANALYSIS PHASE (HOW TO EVALUATE THE RIVAL EXPLANATION) |
|---|---|---|
| Effect–cause (the outcome is actually the cause of the predictor) | Do a longitudinal study to discover which came first<br>Obtain data on the historic sequence of the variables<br>(Ultimate solution: do a randomized trial) | Consider biologic plausibility<br>Compare the strength of the association immediately after the exposure to the predictor with the strength later on<br>Consider findings of other studies with different designs |
| Confounding (another variable causes both the predictor and outcome) | See Table 9.4 | See Table 9.5 |

## Confounding

The other rival explanation in Table 9.3 is confounding, which occurs when a third factor is a real cause of the outcome and the predictor of interest is associated with, but not a cause of, this third factor. For example, if certain personality traits cause people to drink more coffee and also to be at higher risk of MI, these personality traits will confound the association between coffee and MI. If this is the actual explanation, then the association between coffee and MI does not represent cause–effect, although it is perfectly real: Coffee drinking is an innocent bystander in terms of causation.

*In order to be a confounder, a variable must be associated with the predictor of interest and also be a cause of the outcome.* Confounding can be even more complicated, and sometimes, yet another factor is involved. For example, work environment could cause people to drink coffee and to smoke cigarettes, which is a risk factor for MI. Appendix 9A gives a numeric example of how differences in cigarette smoking could lead to an apparent association between coffee drinking and MI.

What if coffee drinking caused smoking and smoking caused MI? In that case, smoking is called a **mediator** of the (causal) association between coffee drinking and MI, not a confounder. In general, it is best to avoid controlling for factors that lie along the causal path between a predictor and an outcome.

Aside from bias, confounding is often the only likely alternative explanation to cause–effect and the most important one to try to rule out. It is also the most challenging; much of the rest of this chapter is devoted to strategies for coping with confounders. It is worth noting, however, that all of these strategies involve judgments, and that no amount of epidemiologic or statistical sophistication can substitute for understanding the underlying biology.

# ■ COPING WITH CONFOUNDERS IN THE DESIGN PHASE

Most strategies for coping with confounding variables require that an investigator measure them, so it is helpful to begin by listing the variables (like age and sex) that may be associated with the predictor variable and also cause the outcome. The investigator must then choose among design and analysis strategies for controlling the influence of these potential confounding variables.

The first two design phase strategies (Table 9.4), **specification** and **matching**, involve changes in the sampling scheme. Cases and controls (in a case–control study) or exposed and unexposed subjects (in a cohort study) can be sampled in such a way that they have comparable values of the confounding variable. This removes the confounder as an explanation for any association that is observed between predictor and outcome. A third design phase strategy, using **opportunistic study designs**, is only applicable to selected research questions for which the right conditions exist. However, when applicable, these designs can resemble randomized trials in their ability to reduce or eliminate confounding not only by measured variables, but by unmeasured variables as well.

## Specification

The simplest strategy is to design inclusion criteria that **specify** a value of the potential confounding variable and exclude everyone with a different value. For example, the investigator studying coffee and MI could specify that only nonsmokers be included in the study. If an association were then observed between coffee and MI, it obviously could not be due to smoking.

Specification is an effective strategy, but, as with all restrictions in the sampling scheme, it has disadvantages. First, even if coffee does not cause MIs in nonsmokers, it may cause them in smokers. This phenomenon—an effect of coffee on MI that is different in smokers from that in nonsmokers—is called **effect modification** (also known as an **interaction**); see Appendix 9A.Thus, specification limits the generalizability of information available from a study, in this instance

**TABLE 9.4 DESIGN PHASE STRATEGIES FOR COPING WITH CONFOUNDERS**

| STRATEGY | ADVANTAGES | DISADVANTAGES |
|---|---|---|
| ***Specification*** | • Easily understood<br>• Focuses the sample of subjects for the research question at hand | • Limits generalizability and sample size |
| ***Matching*** | • Can eliminate influence of strong constitutional confounders like age and sex<br>• Can eliminate the influence of confounders that are difficult to measure<br>• Can increase power by balancing the number of cases and controls in each stratum<br>• May be a sampling convenience, making it easier to select the controls in a case–control study | • May be time-consuming and expensive; may be less efficient than increasing the number of subjects<br>• Decision to match must be made at outset of study and has an irreversible effect on analysis<br>• Requires early decision about which variables are predictors and which are confounders<br>• Eliminates the option of studying matched variables as predictors or as intervening variables<br>• Requires matched analysis<br>• Creates the danger of overmatching (i.e., matching on a factor that is not a confounder, thereby reducing power)<br>• Only feasible for case–control and multiple-cohort studies |
| ***"Opportunistic" study designs*** | • Can provide great strength of causal inference<br>• May be a lower cost and elegant alternative to a randomized trial | • Only possible in select circumstances where predictor variable is randomly or virtually randomly assigned, or instrumental variable exists |

compromising our ability to generalize to smokers. A second disadvantage is that if smoking is highly prevalent among the patients available for the study, the investigator may not be able to recruit a large enough sample of nonsmokers. These problems can become serious if specification is used to control too many confounders or to control them too narrowly. Sample size and generalizability would be major problems if a study were restricted to lower-income, nonsmoking, 70- to 74-year-old men.

## Matching

In a case–control study, **matching** can be used to prevent confounding by selecting cases and controls who have the same (matching) values of the confounding variable(s). Matching and specification both prevent confounding by allowing comparison only of cases and controls who share similar levels of the confounder. Matching differs from specification, however, in preserving generalizability, because subjects at all levels of the confounder can be studied.

Matching is usually done individually (**pair-wise matching**). To control for smoking in a study of coffee drinking as a predictor of MI, for example, each case (a subject with an MI) would be individually matched to one or more controls who smoked roughly the same amount as the case (e.g., 10 to 20 cigarettes/day). The coffee drinking of each case would then be compared with the coffee drinking of the matched control(s).

An alternative approach to pair-wise matching is to match in groups (**frequency matching**). For each level of smoking, the cases with that amount of smoking are counted, and an appropriate number of controls with the same level of smoking are selected. If the study called for two controls per case and there were 20 cases who smoked 10 to 20 cigarettes/day, the investigators would select 40 controls who smoked this amount, matched as a group to the 20 cases.

Matching is most commonly used in **case–control studies**, but it can also be used with **multiple-cohort designs**. For example, to investigate the effects of service in the 1990–1991 Gulf War on subsequent fertility in male veterans, Maconochie et al. (5) compared men deployed to the Gulf region during the war with men who were not deployed, but were frequency-matched by service, age, fitness to be deployed, and so on. They found a slightly higher risk of reported infertility (OR ~1.5) and a longer time to conception in the Gulf War veterans.

## Advantages to Matching (Table 9.4)

- Matching is an effective way to **prevent confounding by constitutional factors** like age, sex, and race that are strong determinants of outcome, not susceptible to intervention, and unlikely to be intermediaries on a causal path.
- Matching can be used to **control confounders that cannot be measured** and controlled in any other way. For example, matching siblings (or, better yet, twins) with one another can control for a whole range of genetic and familial factors that would be impossible to measure. Matching for clinical center in a multicenter study can control for unspecified differences among the populations or staff at geographically dispersed centers.
- Matching may **increase the precision** of comparisons between groups (and therefore the power of the study to find a real association) by balancing the *number* of cases and controls at each level of the confounder. This may be important if the available number of cases is limited or if the cost of studying the subjects is high. However, the effect of matching on precision is modest and not always favorable (see "overmatching," p. 125). In general, the desire to enhance precision is a less important reason to match than the need to control confounding.
- Finally, matching may be used primarily as a **sampling convenience**, to narrow down an otherwise impossibly large number of potential controls. For example, in a study of marijuana use as a risk factor for testicular germ cell tumors, investigators asked cases (men with testicular tumors) to suggest friends of similar age without tumors to be in the control group (6). This convenience, however, also runs the risk of overmatching.

## Disadvantages to Matching (Table 9.4)

- Matching requires additional **time and expense** to identify a match for each subject. In case–control studies, for example, the more matching criteria there are, the larger the pool of controls that must be searched to match each case. The possible increase in statistical power from matching must therefore be weighed against the increase in power that might be obtained by enrolling more cases.
- When matching is used as a sampling strategy, the decision to match must be made at the beginning of the study. It is therefore **irreversible**. This precludes further analysis of the effect of the matched variables on the outcome. It also can create a serious error if the matching variable is not a constitutional variable like age or sex, but an intermediary in the causal path between the predictor and outcome. For example, if an investigator wishing to investigate the effects of alcohol intake on risk of MI matched on serum high-density lipoprotein (HDL) levels, she would miss any beneficial effects of alcohol that are mediated through an increase in HDL. Although the same error can occur with the analysis phase strategies, matching builds the error into the study in a way that cannot be undone; with the analysis phase strategies the error can be avoided by altering the analysis.
- Correct analysis of pair-matched data requires special analytic techniques (**matched analyses**) that compare each subject only with her match, and not with other subjects who have differing levels of confounders. This means cases for whom a match cannot be found cannot be included. In the study of marijuana use and germ cell tumors, 39 of the 187 cases did not provide a friend control (6). The authors had to exclude these 39 cases from the

matched analysis. The use of unmatched analytic techniques on matched data can lead to incorrect results (generally biased toward no effect) because the assumption that the groups are sampled independently is violated.

- A final disadvantage of matching is the possibility of **overmatching**, which occurs when the matching variable is associated with the predictor but turns out not to be a confounder because it is not associated with the outcome. Overmatching can reduce the power of a case–control study, because the matched analysis discards matched case–control sets with the same level of exposure (Appendix 8A.3). In the marijuana and germ cell tumor study, for example, use of friend controls may have reduced the power by increasing the concordance in exposures between cases and their matched controls: Friends might tend to have similar patterns of marijuana use.

## Opportunistic Studies

Occasionally, there are opportunities to control for confounding variables in the design phase, even without measuring them; we call these "opportunistic" designs because they utilize unusual opportunities for controlling confounding. One example, useful when studying the immediate effects of short-term exposures, is the **case-crossover** study (Chapter 8)—all potential confounding variables that are constant over time (e.g., sex, race, social class, genetic factors) are controlled because each subject is compared only with herself in a different time period.

Another opportunistic design involves a **natural experiment**, in which subjects are either exposed or not exposed to a particular risk factor through a process that, in effect, acts randomly (7). For example, Lofgren et al. (8) studied the effects of discontinuity of in-hospital care by taking advantage of the fact that patients admitted after 5:00 PM to their institution were alternately assigned to senior residents who either maintained care of the patients or transferred them to another team the following morning. They found that patients whose care was transferred had 38% more laboratory tests ordered ($P = 0.01$) and 2-day longer median length of stay ($P = 0.06$) than those kept on the same team. Similarly, Bell and Redelmeier (9) studied the effects of nursing staffing by comparing outcomes for patients with selected diagnoses who were admitted on weekends to those admitted on weekdays. They found higher mortality from all three conditions they predicted would be affected by reduced weekend staffing ratios, but no increase in mortality for patients hospitalized for other conditions.

As genetic differences in susceptibility to an exposure are elucidated, a strategy called **Mendelian randomization** (10) becomes an option. This strategy works because, for common genetic polymorphisms, the allele a person receives is determined at random within families, and not linked to most confounding variables. For example, some farmers who dip sheep in insecticides (to kill ticks, lice, etc.) have health complaints, such as headache and fatigue, that might be due to that occupational insecticide exposure. Investigators (11) took advantage of a polymorphism in the paraoxonase-1 gene, which leads to enzymes with differing ability to hydrolyze the organophosphate insecticide (diazinonoxon) used in sheep dip. They found that exposed farmers with health complaints were more likely to have alleles associated with reduced paraoxonase-1 activity than similarly exposed but asymptomatic farmers. This finding provided strong evidence of a causal relationship between exposure to sheep dip and health problems.

Natural experiments and Mendelian randomization are examples of a more general approach to enhancing causal inference in observational studies, the use of **instrumental variables**. These are variables associated with the predictor of interest, but not independently associated with the outcome. Whether someone is admitted on a weekend, for example, is associated with staffing levels, but was thought not to be otherwise associated with mortality risk (for the diagnoses studied), so admission on a weekend can be considered an instrumental variable. Similarly, activity of the paraoxonase-1 enzyme is associated with possible toxicity due to dipping sheep, but not otherwise associated with ill health. Other examples of instrumental variables are draft

lottery numbers to investigate delayed effects on mortality of military service during the Vietnam War era (12); and whether long-term survival for early-stage kidney cancer depends on how far someone lives from a urologist who does partial nephrectomies versus one who only does radical nephrectomies (13).

## COPING WITH CONFOUNDERS IN THE ANALYSIS PHASE

The Design phase strategies specification and matching require deciding at the outset of the study which variables are confounders, and the investigators cannot subsequently estimate the effects of those confounders on an outcome. By contrast, analysis phase strategies keep the investigator's options open, so that she can change her mind about which variables to control for at the time of analysis.

Sometimes there are several predictor variables, each of which may act as a confounder to the others. For example, although coffee drinking, smoking, male sex, and personality type are associated with MI, they are also associated with each other. The goal is to determine which of these predictor variables are independently associated with MI and which are associated with MI only because they are associated with other (causal) risk factors. In this section, we discuss analytic methods for assessing the **independent** contribution of predictor variables in observational studies. These methods are summarized in Table 9.5.[1]

### Stratification

Like specification and matching, **stratification** ensures that only cases and controls (or exposed and unexposed subjects) with similar levels of a potential confounding variable are compared. It involves segregating the subjects into strata (**subgroups**) according to the level of a potential confounder and then examining the relation between the predictor and outcome separately in each stratum. Stratification is illustrated in Appendix 9A. By considering smokers and nonsmokers separately ("stratifying on smoking"), the confounding effects of smoking can be removed.

Appendix 9A also illustrates **effect modification**, in which stratification reveals that the association between predictor and outcome varies with (is modified by) the level of a third factor. Effect modification introduces additional complexity, because a single measure of association no longer can summarize the relationship between predictor and outcome. By chance alone, the estimates of association in different strata will rarely be precisely the same, and it is only when the estimates vary markedly that the findings suggest effect modification. Clinically significant effect modification is uncommon, and before concluding that it is present it is necessary to assess its statistical significance, and, especially if many subgroups have been examined (increasing the likelihood of at least one being statistically significant due to chance), to see if it can be replicated in another population. Biologic plausibility, or the lack thereof, may also contribute to the interpretation. The issue of effect modification also arises for subgroup analyses of clinical trials (Chapter 11), and for meta-analyses when homogeneity (similarity) of studies is being considered (Chapter 13).

Stratification has the advantage of **flexibility**: by performing several stratified analyses, the investigator can decide which variables appear to be confounders and ignore the remainder. This can be done by combining knowledge about the likely directions of causal relationships with analyses determining whether the results of stratified analyses substantially differ from those of unstratified analyses (see Appendix 9A). Stratification also has the advantage of being reversible: No choices need be made at the beginning of the study that might later be regretted.

[1]Similar questions arise in studies of diagnostic tests (Chapter 12), but in those situations the goal is not to determine a causal effect, but to determine whether the test being studied adds substantial predictive power to information already available at the time it was done.

**TABLE 9.5 ANALYSIS PHASE STRATEGIES FOR COPING WITH CONFOUNDERS**

| STRATEGY | ADVANTAGES | DISADVANTAGES |
|---|---|---|
| ***Stratification*** | • Easily understood<br>• Flexible and reversible; can choose which variables to stratify upon after data collection | • Number of strata limited by sample size needed for each stratum<br>  • Few covariables can be considered<br>  • Few strata per covariable leads to incomplete control of confounding<br>• Relevant covariables must have been measured |
| ***Statistical adjustment*** | • Multiple confounders can be controlled simultaneously<br>• Information in continuous variables can be fully used<br>• Flexible and reversible | • Model may not fit:<br>  • Incomplete control of confounding (if model does not fit confounder-outcome relationship)<br>  • Inaccurate estimates of strength of effect (if model does not fit predictor–outcome relationship)<br>• Results may be hard to understand. (Many people do not readily comprehend the meaning of a regression coefficient.)<br>• Relevant covariables must have been measured |
| ***Propensity scores*** | • Multiple confounders can be controlled simultaneously<br>• Information in continuous variables can be fully used<br>• Enhances power to control for confounding when more people receive the treatment than get the outcome<br>• If a stratified or matched analysis is used, does not require model assumptions<br>• Flexible and reversible<br>• Lack of overlap of propensity scores can highlight subgroups in whom control of confounding is difficult or impossible | • Results may be hard to understand<br>• Relevant covariables must have been measured<br>• Can only be done for exposed and unexposed subjects with overlapping propensity scores, reducing sample size |

The principal disadvantage of stratified analysis is the limited number of variables that can be controlled simultaneously. For example, possible confounders in the coffee and MI study might include age, personality type, systolic blood pressure, serum cholesterol, and cigarette smoking. To stratify on these five variables with only three strata for each would require $3^5 = 243$ strata! With this many strata there will be some strata with no cases or no controls, and these strata cannot be used.

To maintain a sufficient number of subjects in each stratum, a variable is often divided into broader strata. When the strata are too broad, however, the confounder may not be adequately controlled. For example, if the preceding study stratified age using only two strata (e.g., <50 and ≥50 years), some residual confounding would still be possible if within each age stratum the subjects drinking the most coffee were older and therefore at higher risk of MI.

## Adjustment

Several statistical techniques are available to adjust for confounders. These techniques *model* the nature of the associations among the variables to isolate the effects of predictor variables and confounders. For example, a study of the effect of lead levels on the intelligence quotient (IQ)

in children might examine parental education as a potential confounder. Statistical adjustment might model the relation between parents' years of schooling and the child's IQ as a straight line, in which each year of parent education is associated with a fixed increase in child IQ. The IQs of children with different lead levels could then be adjusted to remove the effect of parental education using the approach described in Appendix 9B.

Often, an investigator wants to adjust simultaneously for several potential confounders—such as age, sex, race, and education. This requires using multivariate adjustment techniques, such as multivariable linear or logistic regression, or Cox proportional hazards analysis. These techniques have another advantage: They enable the use of all the information in continuous variables. It is easy, for example, to adjust for a parent's education level in 1-year intervals, rather than stratifying into just a few categories. In addition, **interaction terms** can be used to model effect modification among the variables.

There are, however, disadvantages of multivariate adjustment. Most important, the model may not fit. Computerized statistical packages have made these models so accessible that the investigator may not stop to consider whether their use is appropriate for the predictor and outcome variables in the study.[3] Taking the example in Appendix 9B, the investigator should examine whether the relation between the parents' years of schooling and the child's IQ is actually linear. If the pattern is very different (e.g., the slope of the line becomes steeper with increasing education) then attempts to adjust IQ for parental education using a linear model will be imperfect and the estimate of the independent effect of lead will be incorrect.

Second, the resulting statistics are often difficult to understand. This is particularly a problem if transformations of variables (e.g., parental education squared) or interaction terms are used. Investigators should spend the necessary time with a statistician (or take the necessary courses) to make sure they can explain the meaning of coefficients or other highly derived statistics they plan to report. As a safety precaution, it is a good idea always to start with simple, stratified analyses, and to seek help understanding what is going on if more complicated analyses yield substantially different results.

## Propensity Scores

**Propensity scores** can be particularly useful for observational studies of treatment efficacy to control **confounding by indication**—the problem that patients for whom a treatment is indicated (and prescribed) are often at higher risk, or otherwise different, from those who do not get the treatment. Recall that in order to be a confounder, a variable must be associated with both the predictor and outcome. Instead of adjusting for all factors that predict *outcome*, use of propensity scores involves creating a multivariate model to predict receipt of the *treatment*. Each subject can then be assigned a predicted probability of treatment—a "propensity score." This single score can be used as the only confounding variable in a stratified or mutlivariable analysis.

Alternatively, subjects who did and did not receive the treatment can be matched by propensity score, and outcomes compared between matched pairs. Unlike use of matching as a design-phase (sampling) strategy, propensity matching resembles other analysis phase strategies in being reversible. However, matched propensity analyses fail for subjects who cannot be matched because their propensity scores are close to 0 or 1. While this reduces sample size, it may be an advantage because in these unmatchable subjects the propensity score analysis has identified a lack of comparability between groups and inability to control for confounding that might not have been apparent with other methods of multivariable analysis.

[3]One of our biostatistician colleagues has quipped that trying to design a user-friendly, intuitive statistical software package is like trying to design a car so that a child can reach the pedals.

### EXAMPLE 9.1 Propensity Analysis

Gum et al. (14) prospectively studied 6,174 consecutive adults undergoing stress echocardiography, 2,310 of whom (37%) were taking aspirin and 276 of whom died in the 3.1-year follow-up period. In unadjusted analyses, aspirin use was not associated with mortality (4.5% in both groups). However, when 1,351 patients who had received aspirin were matched to 1,351 patients with the same propensity to receive aspirin but who did not, mortality was 47% lower in those treated ($P = 0.002$).

Analyses using propensity scores have several advantages. The number of potential confounding variables that can be modeled as predictors of an intervention is usually greater than the number of variables that can be modeled as predictors of an outcome, because the number of people treated is generally much greater than the number who develop the outcome (2,310 compared with 276 in the Example 9.1). Another reason that more confounders can be included is that there is no danger of "overfitting" the propensity model—interaction terms, quadratic terms, and multiple indicator variables can all be included (15). Finally, investigators are usually more confident in identifying the determinants of treatment than the determinants of outcome, because the treatment decisions were made by clinicians based on a limited number of patient characteristics.

Of course, like other multivariate techniques, use of propensity scores still requires that potential confounding variables be identified and measured. A limitation of this technique is that it does not provide information about the relationship between any of the confounding variables and outcome—the only result is for the predictor (usually, a treatment) that was modeled. However, because this is an analysis phase strategy, it does not preclude doing more traditional multivariate analyses as well, and both types of analysis are usually done.

## ■ OTHER PITFALLS IN QUANTIFYING CAUSAL EFFECTS

### Conditioning on a Shared Effect

The bias caused by **conditioning on a shared effect** is kind of tricky, and it is sometimes skipped in introductory textbooks because most explanations of it use abstract diagrams and notation. By contrast, we will first give a few examples of how it might occur, and then try to explain what the name means.

Consider a study of people who have lost at least 15 pounds in the previous year. An investigator finds that the subjects who have been dieting have a lower risk of cancer than those who have not been dieting. Do you think dieting prevented cancer in these subjects?

If you stop and think, you'll probably answer no, because cancer also causes weight loss. You can imagine that if someone loses weight for no apparent reason it is much more likely to signify a cancer than if someone loses weight while dieting. *Among people who have lost weight*, if the weight loss was not caused by dieting, it is more likely to have been caused by something more ominous. The investigators created an inverse association between dieting and cancer by conditioning on (restricting attention to) a shared effect (weight loss, which is caused by both dieting and cancer).

Here's another example. Among low birth weight babies, those whose mothers smoked during pregnancy have lower infant mortality than those whose mothers did not smoke (16). Should we encourage more mothers to smoke during pregnancy? Definitely not! The reason for this observation is that smoking causes low birth weight, but so do other things, especially prematurity. So *among low birth weight babies*, if the low birth weight was not caused by smoking, it is more likely to have been caused by prematurity. The investigators created an inverse

association between smoking and prematurity (and its associated mortality risk) by conditioning on (restricting attention to) a shared effect (low birth weight, which is caused by both smoking and prematurity).

Now the phrase "conditioning on a shared effect" makes sense. **Conditioning** is an epidemiologic term that means looking at associations between predictor and outcome variables "conditioned on" (i.e., at specified levels of) some attribute. A **shared effect** refers to an attribute (like losing weight, or being a low birth weight baby) that has several causes. Bias due to conditioning on a shared effect can occur if the investigator treats something *caused* by the risk factor being studied as an inclusion criterion, a matching variable, or a possible confounding variable.

### Underestimation of Causal Effects

To this point, our emphasis has been on evaluating the likelihood of alternative explanations for an association, in order to avoid concluding that an association is real and causal when it is not. However, another type of error is also possible—*underestimation* of causal effects. Chance, bias, and confounding can also be reasons why a real association might be missed or underestimated.

We discussed **chance** as a reason for missing an association in Chapter 5, when we reviewed type II errors and the need to make sure the sample size will provide adequate **power** to find real associations. After a study has been completed, however, the power calculation is no longer a good way to quantify uncertainty due to random error. At this stage a study's hypothetical power to detect an effect of a specified size is less relevant than the actual findings, expressed as the observed estimate of association (e.g., risk ratio) and its 95% **confidence interval** (17).

**Bias** can also distort estimates of association toward no effect. In Chapter 8, the need for blinding in ascertaining risk factor status among cases and controls was to avoid **differential measurement bias**, for example, differences between the cases and controls in the way questions were asked or answers interpreted that might lead observers to get the answers they desire. Because observers might desire results in either direction, differential measurement bias can bias results to either overestimate or underestimate causal effects. Non-differential bias, on the other hand, will generally lead to underestimation of associations.

**Confounding** can also lead to attenuation of real associations. For example, suppose coffee drinking actually protected against MI, but was more common in smokers. If smoking were not controlled for, the beneficial effects of coffee might be missed—coffee drinkers might appear to have the same risk of MI as those who did not drink coffee, when their higher prevalence of smoking should have caused their risk to be higher. This type of confounding, in which the effects of a beneficial factor are hidden by its association with a cause of the outcome, is sometimes called **suppression** (18). It is a common problem for observational studies of treatments, because treatments are often most indicated in those at higher risk of a bad outcome. The result, noted earlier, is that a beneficial treatment can appear to be useless (as aspirin did in Example 9.1) or even harmful until the confounding by indication is controlled.

## CHOOSING A STRATEGY

What general guidelines can be offered for deciding whether to cope with confounders during the design or analysis phases, and how best to do it? The use of **specification** to control confounding is most appropriate for situations in which the investigator is chiefly interested in specific subgroups of the population; this is really just a special form of the general process of establishing criteria for selecting the study subjects (Chapter 3). However, for studies in which causal inference is the goal, there's the additional caution to avoid inclusion criteria that could be caused by predictor variables you wish to study (i.e., conditioning on a shared effect).

An important decision to make in the design phase of the study is whether to **match**. Matching is most appropriate for case–control studies and fixed constitutional factors such as age,

race, and sex. Matching may also be helpful when the sample size is small compared with the number of strata necessary to control for known confounders, and when the confounders are more easily matched than measured. However, because matching can permanently compromise the investigator's ability to observe real associations, it should be used sparingly, particularly for variables that may be in the causal chain. In many situations the analysis phase strategies (stratification, adjustment, and propensity scores) are just as good for controlling confounding, and have the advantage of being **reversible**—they allow the investigator to add or subtract covariates to explore different causal models.

Although not available for all research questions, it is always worth considering the possibility of an **opportunistic** study design. If you don't stop and consider (and ask your colleagues about) these studies, you might miss a great opportunity to do one.

The final decision to **stratify**, **adjust**, or use **propensity scores** need not be made until after the data are collected; in many cases the investigator may wish to do all of the above. However, it is important during study design to consider which factors may later be used for adjustment, in order to know which variables to measure. In addition, because different analysis phase strategies for controlling confounding do not always yield the same results, it is best to specify a primary analysis plan in advance. This may help investigators resist the temptation of selecting the strategy that provides the most desired results.

## Evidence Favoring Causality

The approach to enhancing causal inference has largely been a negative one thus far—how to rule out the four rival explanations in Table 9.1. A complementary strategy is to seek characteristics of associations that provide positive evidence for causality, of which the most important are the consistency and strength of the association, the presence of a dose–response relation, and biologic plausibility.

When the results are **consistent** in studies of various designs, it is less likely that chance or bias is the cause of an association. Real associations that represent effect–cause or confounding, however, will also be consistently observed. For example, if cigarette smokers drink more coffee and have more MIs in the population, studies will consistently observe an association between coffee drinking and MI.

The **strength** of the association is also important. For one thing, stronger associations give more significant *P* values, making chance a less likely explanation. Stronger associations also provide better evidence for causality by reducing the likelihood of confounding. Associations due to confounding are indirect (i.e., via the confounder) and therefore are generally weaker than direct cause–effect associations. This is illustrated in Appendix 9A: The strong associations between coffee and smoking (odds ratio = 16) and between smoking and MI (odds ratio = 4) led to a much weaker association between coffee and MI (odds ratio = 2.25).

A **dose–response** relation provides positive evidence for causality. The association between cigarette smoking and lung cancer is an example: Moderate smokers have higher rates of cancer than do nonsmokers, and heavy smokers have even higher rates. Whenever possible, predictor variables should be measured continuously or in several categories, so that any dose–response relation that is present can be observed. Once again, however, a dose–response relation can be observed with effect–cause associations or with confounding.

Finally, **biologic plausibility** is an important consideration for drawing causal inference—if a causal mechanism that makes sense biologically can be proposed, evidence for causality is enhanced, whereas associations that do not make sense given our current understanding of biology are less likely to represent cause–effect. For example, in the study of marijuana use as a risk factor for germ cell tumors, use of marijuana less than once a day was associated with lower risk than no use (6). It is hard to explain this biologically.

It is important not to overemphasize biologic plausibility, however. Investigators seem to be able to come up with a plausible mechanism for virtually any association and some associations

originally dismissed as biologically implausible, such as a bacterial etiology for peptic ulcer disease, have turned out to be real.

## SUMMARY

1. The design of **observational studies** should anticipate the need to interpret **associations**. The inference that the association represents a **cause–effect** relationship (often the goal of the study) is strengthened by strategies that reduce the likelihood of the **four rival explanations—chance**, **bias**, **effect–cause**, and **confounding**.
2. The role of **chance (random error)** can be minimized by designing a study with **adequate sample size and precision** to assure low **type I and type II error** rates. Once the study is completed, the effect of random error can be judged from the width of the **95% confidence interval** and the consistency of the results with **previous evidence**.
3. **Bias (systematic error)** arises from differences between the population and phenomena addressed by the research question and the actual subjects and measurements in the study. Bias can be minimized by basing design decisions on a **judgment** as to whether these differences will lead to a wrong answer to the research question.
4. **Effect–cause** is made less likely by designing a study that permits assessment of **temporal sequence**, and by considering **biologic plausibility**.
5. **Confounding**, which may be present when a third variable is associated with the predictor of interest and is a cause of the outcome, is made less likely by the following strategies, most of which require potential confounders to be anticipated and measured:
   a. **Specification** or **matching** in the **design phase**, which alters the sampling strategy to ensure that only groups with similar levels of the confounder are compared. These strategies **should be used judiciously** because they can **irreversibly** limit the information available from the study.
   b. **Analysis phase** strategies that accomplish the same goal and preserve options for investigating causal paths:
      - **Stratification**, which in addition to controlling for confounding can reveal **effect modification** ("**interaction**"), a different magnitude of predictor–outcome association at different levels of a third variable.
      - **Adjustment**, which can permit the impact of many predictor variables to be controlled simultaneously.
      - **Propensity scores**, which enhance the power for addressing **confounding by indication** in observational studies of treatment efficacy.
6. Investigators should be on the lookout for **opportunistic** observational designs, including **natural experiments**, **Mendelian randomization**, and other **instrumental variable** designs, that offer a strength of causal inference that can approach that of a randomized clinical trial.
7. Investigators should avoid **conditioning on shared effects** in the design phase by not selecting subjects based on covariates that might be caused by the predictor, and in the analysis phase by not controlling for these covariates.
8. Causal inference can be enhanced by positive evidence, notably the **consistency and strength of the association**, the presence of a **dose–response** relation, and **biologic plausibility**.

# APPENDIX 9A
# Hypothetical Example of Confounding and Effect Modification

The entries in these tables are numbers of subjects in this hypothetical case–control study

Panel 1. If we look at the entire group of study subjects, there appears to be an association between coffee drinking and MI:

| | Smokers and Nonsmokers Combined | |
|---|---|---|
| | MI | No MI |
| Coffee | 90 | 60 |
| No coffee | 60 | 90 |

Odds ratios (OR) for MI associated with coffee in smokers and nonsmokers combined $= \dfrac{90 \times 90}{60 \times 60} = 2.25$

Panel 2. However, this could be due to **confounding**, as shown by the tables stratified on smoking which show that coffee drinking is not associated with MI in either smokers or nonsmokers:

| | Smokers | |
|---|---|---|
| | MI | No MI |
| Coffee | 80 | 40 |
| No coffee | 20 | 10 |

| | Nonsmokers | |
|---|---|---|
| | MI | No MI |
| Coffee | 10 | 20 |
| No coffee | 40 | 80 |

Odds ratios for MI associated with coffee:

$$\text{OR in smokers} = \frac{80 \times 10}{20 \times 40} = 1 \qquad \text{OR in nonsmokers} = \frac{10 \times 80}{40 \times 20} = 1$$

Smoking is a confounder because it is strongly associated with coffee drinking (below, left panel) and with MI (below, right panel): These tables were obtained by rearranging numbers in Panel 2.

| | MI and No MI Combined | |
|---|---|---|
| | Coffee | No Coffee |
| Smokers | 120 | 30 |
| Nonsmokers | 30 | 120 |

| | Coffee and No Coffee Combined | |
|---|---|---|
| | MI | No MI |
| Smokers | 100 | 50 |
| Nonsmokers | 50 | 100 |

Odds ratio for coffee drinking associated with smoking $= \dfrac{120 \times 120}{30 \times 30} = 16$

Odds ratio for MI associated with smoking $= \dfrac{100 \times 100}{50 \times 50} = 4$

Panel 3. The association between coffee drinking and MI in Panel 1 could also represent **effect modification**, if stratification on smoking revealed that the association between coffee drinking and MI differs in smokers and nonsmokers. In the table below, the OR of 2.25 for the association between coffee drinking and MI in smokers and nonsmokers combined is due entirely to a strong association in smokers. When effect modification is present, the odds ratios in different strata are different, and must be reported separately:

| | Smokers | |
|---|---|---|
| | MI | No MI |
| Coffee | 50 | 15 |
| No coffee | 10 | 33 |

| | Nonsmokers | |
|---|---|---|
| | MI | No MI |
| Coffee | 40 | 45 |
| No coffee | 50 | 57 |

Odds ratios for MI associated with coffee:

$$\text{OR in smokers} = \frac{50 \times 33}{15 \times 10} = 11 \qquad \text{OR in nonsmokers} = \frac{40 \times 57}{45 \times 50} = 1$$

Bottom Line: The overall association between coffee drinking and MI in Panel 1 could be hiding the presence of confounding by smoking, which would be revealed by stratification on smoking (Panel 2). Or it could be hiding the presence of effect modification by smoking, which would also be revealed by stratification on smoking (Panel 3). It could also represent cause-effect, which would be supported (though not proven) if stratification on smoking did not alter the association between coffee drinking and MI. Finally (and most realistically), if could be a result of some mixture of all of the above.

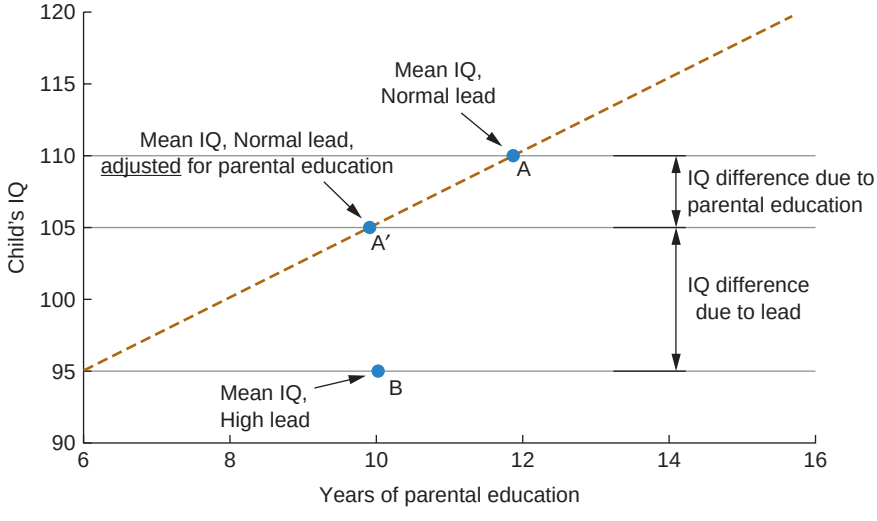
# APPENDIX 9B
# A Simplified Example of Adjustment

Suppose that a study finds two major predictors of the intelligence quotient (IQ) of children: the parental education level and the child's blood lead level. Consider the following hypothetical data on children with normal and high lead levels:

| | Average Years of Parental Education | Average IQ of Child |
|---|---|---|
| High lead level | 10.0 | 95 |
| Normal lead level | 12.0 | 110 |

Note that the parental education level is also associated with the child's blood lead level. The question is, "Is the difference in IQ between children with normal and high lead levels more than can be accounted for on the basis of the difference in parental education?" To answer this question we look at how much difference in IQ the difference in parental education levels would be expected to produce. We do this by plotting parental educational level versus IQ in the children with normal lead levels (Figure 9.2).[4]

The diagonal dashed line in Figure 9.2 shows the relationship between the child's IQ and parental education in children with normal lead levels; there is an increase in the child's IQ of 5 points for each 2 years of parental education. Therefore, we can adjust the IQ of the normal lead group to account for the difference in mean parental education by sliding down the line from




■ **FIGURE 9.2** Hypothetical graph of child's IQ as a linear function (*dashed line*) of years of parental education.

[4]This description of analysis of covariance (ANCOVA) is simplified. Actually, parental education is plotted against the child's IQ in both the normal and high lead groups, and the single slope that fits both plots the best is used. The model for this form of adjustment therefore assumes linear relationships between education and IQ in both groups, and that the slopes of the lines in the two groups are the same.

point *A* to point *A′*. (Because the group with normal lead levels had 2 more years of parental education on the average, we adjust their IQs downward by 5 points to make them comparable in mean parental education to the high lead group.) This still leaves a 10-point difference in IQ between points *A* and *B*, suggesting that lead has an independent effect on IQ of this magnitude. Therefore, of the 15-point difference in IQ of children with low and high lead levels, 5 points can be accounted for by their parents' different education levels and the remaining 10 are attributable to the lead exposure.

## REFERENCES


1. McEvoy SP, Stevenson MR, McCartt AT, et al. Role of mobile phones in motor vehicle crashes resulting in hospital attendance: a case-crossover study. *BMJ* 2005;331(7514):428.
2. Magruder JT, Elahi D, Andersen DK. Diabetes and pancreatic cancer: chicken or egg? *Pancreas* 2011;40(3):339–351.
3. Huxley R, Ansary-Moghaddam A, Berrington de Gonzalez A, et al. Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies. *Br J Cancer* 2005;92(11):2076–2083.
4. Bosetti C, Rosato V, Polesel J, et al. Diabetes mellitus and cancer risk in a network of case-control studies. *Nutr Cancer* 2012;64(5):643–651.
5. Maconochie N, Doyle P, Carson C. Infertility among male UK veterans of the 1990-1 Gulf war: reproductive cohort study. *BMJ* 2004;329(7459):196–201.
6. Trabert B, Sigurdson AJ, Sweeney AM, et al. Marijuana use and testicular germ cell tumors. *Cancer* 2011; 117(4):848–853.
7. Newman TB, Kohn M. *Evidence-based diagnosis*. New York: Cambridge University Press, 2009. Chapter 10.
8. Lofgren RP, Gottlieb D, Williams RA, et al. Post-call transfer of resident responsibility: its effect on patient care [see comments]. *J Gen Intern Med* 1990;5(6):501–505.
9. Bell CM, Redelmeier DA. Mortality among patients admitted to hospitals on weekends as compared with weekdays. *N Engl J Med* 2001;345(9):663–668.
10. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;32(1):1–22.
11. Cherry N, Mackness M, Durrington P, et al. Paraoxonase (PON1) polymorphisms in farmers attributing ill health to sheep dip. *Lancet* 2002;359(9308):763–764.
12. Hearst N, Newman TB, Hulley SB. Delayed effects of the military draft on mortality. A randomized natural experiment. *N Engl J Med* 1986;314(10):620–624.
13. Tan HJ, Norton EC, Ye Z, et al. Long-term survival following partial vs radical nephrectomy among older patients with early-stage kidney cancer. *JAMA* 2012;307(15):1629–1635.
14. Gum PA, Thamilarasan M, Watanabe J, et al. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: A propensity analysis. *JAMA* 2001;286(10):1187–1194.
15. Klungel OH, Martens EP, Psaty BM, et al. Methods to assess intended effects of drug treatment in observational studies are reviewed. *J Clin Epidemiol* 2004;57(12):1223–1231.
16. Hernandez-Diaz S, Schisterman EF, Hernan MA. The birth weight "paradox" uncovered? *Am J Epidemiol* 2006;164(11):1115–1120.
17. Bacchetti P. Current sample size conventions: flaws, harms, and alternatives. *BMC Med* 2010;8:17.
18. Katz MH. *Multivariable analysis: a practical guide for clinicians*, 2nd ed. Cambridge, UK; New York: Cambridge University Press, 2006.