9 Enhancing Causal Inference in Observational Studies

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

One of the most important aspects of clinical research is the inference that an association represents a **cause-effect** relation. In this chapter we discuss ways to strengthen causal inferences based on associations in observational studies. We begin with a discussion of how to avoid **spurious associations** and then concentrate on ruling out **real associations** that do not represent cause-effect, especially those due to confounding.

Suppose that a study reveals an association between coffee drinking and myocardial infarction (MI). One possibility is that coffee drinking is a cause of MI. Before reaching this conclusion, however, four rival explanations must be considered (Table 9.1). The first two of these, **chance** (random error) and **bias** (systematic error), represent spurious associations: Coffee drinking and MI are not really associated in the population, only in the study findings.

Even if the association is real, however, it may not represent a cause-effect relation. Two rival explanations must be considered. One is the possibility of an **effect-cause** relation—that having an MI makes people drink more coffee. (This is just cause and effect in reverse.) The other is the possibility of **confounding**—that some third factor (such as cigarette smoking) is a cause of MI and is also associated with coffee drinking.

■ SPURIOUS ASSOCIATIONS

Ruling Out Spurious Associations Due to Chance

Imagine that there is no association between coffee drinking and MI in the population, and that 60% of the entire population drinks coffee, whether or not they have had an MI. If we were to select a random sample of 20 MI patients, we would expect about 12 of them to drink coffee. But by chance alone we might happen to get 19 coffee drinkers in a sample of 20 MI patients. In that case, unless we were lucky enough to get a similar chance excess of coffee drinkers among the controls, a spurious association between coffee consumption and MI would be observed. Such an association due to **random error** (chance) is called a Type I error (Chapter 5).

Strategies for minimizing random errors are available in both the design and analysis phases of research (Table 9.2). The design strategies of increasing the

■ **TABLE 9.1**
The Five Explanations When an Association Between Coffee Drinking and Myocardial Infarction (MI) Is Observed in a Sample

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	Coffee drinking → MI
4. Confounding	Real	Coffee drinking is associated with a third, extrinsic factor that is a cause of MI	Factor X; Coffee drinking; MI
5. Cause-Effect	Real	Coffee drinking is a cause of MI	Coffee drinking → MI

precision of measurements and increasing the sample size are important ways to reduce random error that are discussed in Chapters 4 and 6. 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 alone. For example, a *P* value of 0.10 indicates that the observed difference between the two groups was as large a difference as would occur by chance alone about one time in 10.

Ruling Out Spurious Associations Due to Bias

Associations that are spurious because of bias are trickier. To understand bias it is important to distinguish between the research question and the question actually answered by the study (Chapter 1). The research question is the uncertainty in the universe the investigator really wishes to settle, and the question answered

■ **TABLE 9.2**
Strengthening the Inference That an Association Has a Cause-Effect Basis: Ruling Out Spurious Associations

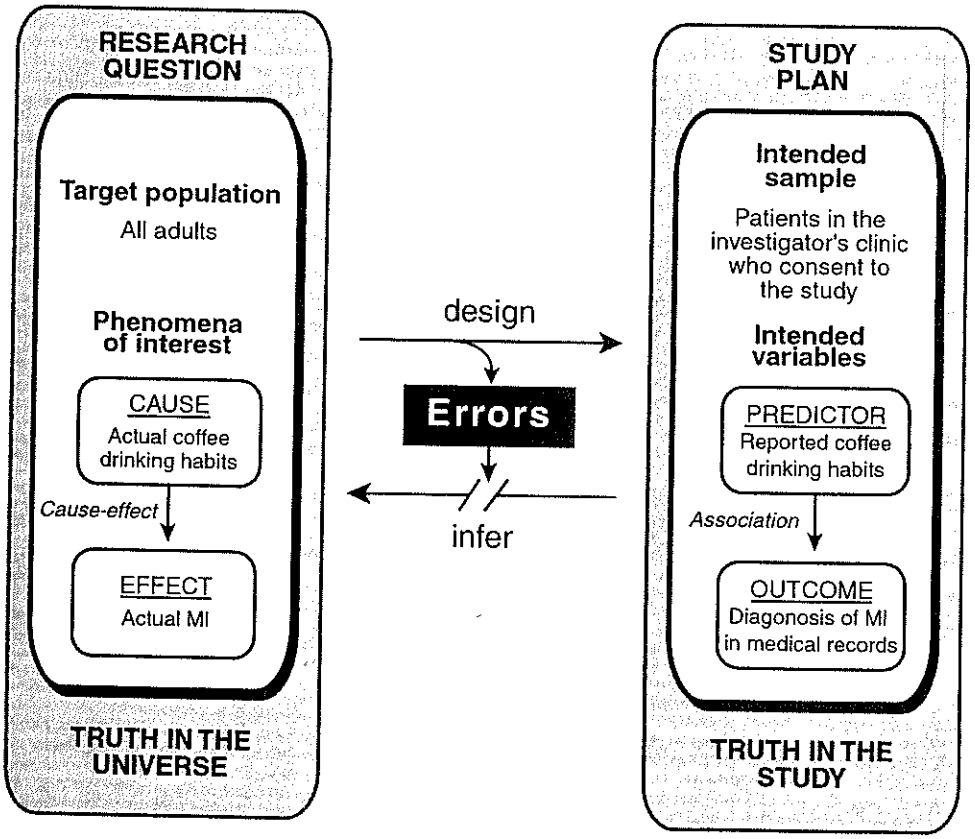
Type of Spurious Association	Design Phase (How to prevent the rival explanation)	Analysis Phase (How to evaluate the rival explanation)
Chance (due to random error)	Increase sample size and other strategies (Chapters 4 and 6).	Interpret *P* value in context of prior evidence (Chapter 5).
Bias (due to systematic error)	Carefully consider the potential consequences of each difference between the research question and the study plan: Subjects Predictor Outcome	Obtain additional data to see if potential biases have actually occurred. Check consistency with other studies (especially those using different methods).

by the study reflects the compromises the investigator needed to make for the study to be feasible. Bias can be thought of as a systematic difference between the research question and the actual question answered by the study that may cause the study to give the wrong answer to the research question. Strategies for minimizing these **systematic errors** are available in both the design and analysis phases of research (Table 9.2).

Design Phase. Many kinds of bias have been identified, and dealing with some of them has been a major topic of this book. To the specific strategies noted in Chapters 3, 4, 7, and 8 we now add a general approach to minimizing sources of bias. Write down the research question and the study plan side by side, as in Fig. 9.1. Then carefully think through the following three concerns:

1. Do the **samples** of study subjects (e.g., cases and controls) sufficiently represent the population(s) of interest?
2. Does the **measurement of the predictor variable** sufficiently represent the predictor of interest?
3. Does the **measurement of the outcome variable** sufficiently represent the outcome of interest?

For each question answered "No" or "Maybe not," consider whether the bias is large enough that the study could give the wrong answer to the research question.



■ **FIGURE 9.1**
Minimizing bias by comparing the research question and the study plan.

To illustrate this with our coffee and MI example, consider the implications of drawing the sample of control subjects from a population of hospitalized patients. If many of these patients have chronic illnesses that have caused them to reduce their coffee intake, the sample of controls will not represent the target population from which the MI patients arose; there will be a shortage of coffee drinkers. Furthermore, if coffee drinking is measured by questionnaire, the answers on the questionnaire may not accurately represent actual coffee drinking, the predictor of interest. 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. 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 making a judgment on the likelihood of the potential bias and the degree to which it will compromise the conclusions.

Analysis Phase. The investigator is often faced with one or more potential biases after the data have been collected. Some may have been anticipated but too difficult to prevent, and others may not have been suspected until it was too late to avoid them.

In either situation, one approach is to obtain additional information to estimate the magnitude of the potential bias. Suppose, for example, the investigator is concerned that the hospitalized control subjects do not represent the target population of people free of MI because they have decreased their coffee intake due to chronic illness. The magnitude of this sampling bias could be estimated by reviewing the diagnoses of the control subjects and separating them into two groups: those with an illness like peptic ulcer that might alter coffee habits and those with illnesses that would not. If both types of controls drank less coffee than the MI cases, then sampling bias would be a less likely explanation for the findings. Similarly, 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. Finally, if the outcome measure is in doubt, the investigator could specify objective electrocardiographic and serum enzyme changes needed for the diagnosis, and reanalyze the data excluding the subset of cases that do not meet these criteria.

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 methods and are thus unlikely to share the same biases. In many cases, 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

Once spurious associations have been determined to be unlikely, the two types of associations that are real but do not represent cause-effect must be considered (Table 9.3).

■ 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 Obtain data on the historic sequence of the variables	Consider biologic plausibility Consider findings of other studies with different designs
Confounding (another variable is associated with the predictor and a cause of the outcome)	See Table 9.4	See Table 9.5

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, especially when the predictor variable is a laboratory test for which no previous values are available. Suppose, for example, that a study finds high serum C-reactive protein levels (a marker for inflammation) in men recovering from myocardial infarction. The MI may have caused the high C-reactive protein levels rather than vice versa.

Effect-cause is less commonly a problem in cohort studies because risk factor measurements can be made in a group of patients 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. A good example is the association between low serum cholesterol levels and excess cancer mortality that has been observed in many cohort studies (1). Excess cancer mortality risk associated with low serum cholesterol decreases over time (whereas the excess heart disease mortality in those with high cholesterol continues undiminished) (2). This suggests that preexisting but hidden cancer caused low cholesterol levels at baseline in some subjects. Effect-cause is further supported by the observation that cancer death is associated with *falling* cholesteral levels (3).

This example illustrates a general approach to ruling out effect-cause: drawing inferences from assessments of the variables at different points in time. In addition, effect-cause is often unlikely on the grounds of biologic implausibility. For example, it is unlikely that predisposition to lung cancer causes cigarette smoking.

Confounding

The other rival explanation in Table 9.3 is confounding, which occurs when there is an extrinsic factor involved in the association that is the real cause of the outcome. The general connotation of confounding is something that confuses interpretation, but in epidemiologic research the term has a more specific statistical meaning. A **confounding variable** is one that is **associated with the predictor variable and is a cause of the outcome variable.**

Cigarette smoking is a likely confounder in the coffee and MI example because smoking is associated with coffee drinking and is a cause of MI. If this is the

actual explanation, then the association between coffee and MI does not represent cause-effect even though it is perfectly real; the coffee is an innocent bystander. Appendix 9.A gives a numeric example of how cigarette smoking could cause an apparent association between coffee drinking and MI.

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; the rest of this chapter is devoted to strategies for coping with confounders.

■ COPING WITH CONFOUNDERS IN THE DESIGN PHASE

In observational studies, most strategies for coping with confounding variables require that an investigator be aware of and able to measure them. (This is not true of experiments, which can control unmeasured confounders by randomization. See Chapter 10.) The first step is to list the variables (like age and sex) that may be associated with the predictor variable of interest and that may also be a cause of the outcome. The investigator must then choose between the design and analysis strategies for controlling the influence of these potential confounding variables.

The 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) are sampled in such a way

■ **TABLE 9.4**
Design Phase Strategies for Coping with Confounders

Strategy	Advantages	Disadvantages
Specification	Easily understood	Limits generalizability
	Focuses the sample of subjects for the research question at hand	May make it difficult to acquire an adequate sample size
Matching	Can eliminate influence of strong constitutional confounders like age and sex	May be time-consuming and expensive, less efficient than increasing the number of subjects (e.g., the number of controls per case)
	Can eliminate influence of confounders that are difficult to measure	Decision to match must be made at outset of study and can have irreversible adverse effect on analysis and conclusions
	Can increase precision (power) by balancing the number of cases and controls in each stratum	Requires early decision about which variables are predictors and which are confounders
	May be a sampling convenience, making it easier to select the controls in a case-control study	Removes the option of studying matched variables as predictors or as intervening variables
		Requires matched analysis
		Creates the danger of overmatching (i.e., matching on a factor that is not a confounder, thereby reducing power)

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.

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 would 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 (Chapter 3), it has disadvantages. First, even if coffee does not cause MI 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** or **interaction.**) Thus specification limits the generalizability of information available from a study, in this instance compromising our ability to generalize to smokers. Second, 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 75-year-old men.

Matching

In a case-control study, matching involves selecting cases and controls with matching values of the confounding variable(s). Matching and specification are both sampling strategies that prevent confounding by allowing comparison only of cases and controls that share comparable 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 (**pairwise matching**). In the study of coffee drinking as a predictor of MI, for example, each case (a patient with an MI) could be individually matched to one or more controls that smoked roughly the same amount as the case (e.g., 10 to 20 cigarettes per day). The coffee drinking of each case then would be compared with the coffee drinking of the matched control(s) of that case.

An alternative approach to matching is to do it in groups (**frequency matching**). For each level of smoking, the number of cases with that amount of smoking could be counted, and an appropriate number of controls with the same level of smoking could be selected. If the study called for two controls per case and there were 20 cases that had smoked 10 to 20 cigarettes per day, the investigators would select 40 controls that 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 other designs. For example, to investigate the effects of acute respiratory distress syndrome (ARDS) on subsequent quality of life, 73 survivors of ARDS were matched with controls with comparable severity of underlying illness or injury but no ARDS (4). The matched pairs were then interviewed 2 years after their initial hospitalization about their current quality of life. The subjects who had survived ARDS were faring worse than their paired controls in most domains. This illustrates the use of matching in a double cohort design.

Advantages to Matching. There are four main advantages to matching (Table 9.4). The first three relate to the control of confounding variables; the last is a matter of logistics.

- Matching is an effective way to prevent confounding by **constitutional factors** like age and sex that are strong determinants of outcome, are not susceptible to intervention, and are unlikely to be an intermediary in a causal pathway.
- 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, and matching for clinical center in a multicenter study can control for unspecified differences among the populations seen at the centers.
- Matching may increase the **precision** of comparisons between groups (and thus 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. 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 nationwide study of toxic shock syndrome, victims were asked to identify friends to serve as controls (5). This convenience, however, runs the risk of "overmatching" (discussed later).

Disadvantages to Matching. There are a number of disadvantages to matching (Table 9.4).

- Matching sometimes 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. Cases for which no match can be found will need to be discarded. The possible increase in statistical power from matching must thus be weighed against the potential loss of otherwise eligible cases or controls.
- Because matching is a sampling strategy, the decision to match must be made at the beginning of the study and is **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 fixed (constitutional) variable like age or sex, but a variable intermediate in the causal pathway 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 beneficial effects of alcohol mediated through an increase in HDL. Although the same error can occur with the analysis phase strategies discussed later, matching builds the error into the study in a way that cannot be undone; with the analysis phase strategies the error can be avoided simply by appropriately altering the analysis.
- Correct analysis of matched data requires **special analytic techniques** that compare each subject only with the individual(s) with whom she has been matched, and not with subjects who have differing levels of confounders. The use of

ordinary statistical analysis 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. This sometimes creates a problem because the appropriate matched analyses, especially multivariate techniques, are less familiar to most investigators and less readily available in packaged statistical programs than are the usual unmatched techniques.

A final disadvantage of matching is the possibility of **overmatching,** which occurs when the matching variable is not a confounder because it is not associated with the outcome. Overmatching can reduce the power of a case-control study, making it more difficult to find an association that really exists in the population. In the study of toxic shock syndrome that used friends for controls, for example, matching may have inappropriately controlled for regional differences in tampon marketing, making it more probable that cases and controls would use the same brand of tampon. It is important to note, however, that overmatching will not distort the estimated relative risk (provided that a matched analysis is used); it will only reduce its statistical significance.[*] Thus when the findings of the study are statistically significant (as was the case in the toxic shock example), overmatching is not a problem.

■ COPING WITH CONFOUNDERS IN THE ANALYSIS PHASE

Both design-phase strategies (Table 9.5) require deciding at the outset of the study which variables are predictors and which are confounders. **Stratification** and **adjustment,** two analysis-phase strategies, allow the investigator to defer that decision until she has looked at the data and seen which variables may be confounders.

Sometimes there are several predictor variables, each of which may act as a confounder to the others. For example, although coffee drinking, smoking, sex, and personality type are associated with MI, they are also associated with each other. The goal may be 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.

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)

[*]The reason that overmatching reduces power can be seen with a matched pairs analysis of a case-control study. In the matched analysis, only case-control pairs that are discordant for exposure to the risk factor are analyzed. Matching on a variable associated with the risk factor will lead to fewer discordant pairs and hence smaller effective sample size and less power. Of course, this happens to some extent any time matching is used, not just with overmatching. The difference with overmatching is that this cost comes with no benefit, because the matching was not necessary to control confounding. If a matched analysis is not used, then the estimate of the effect size will be distorted, because the matching causes the cases and controls to be more likely to have the same value of the risk factor.

■ **TABLE 9.5**
Analysis Phase Strategies for Coping with Confounders

Strategy	Advantages	Disadvantages
Stratification	Easily understood	Number of strata limited by sample size needed for each stratum:
	Flexible and reversible; can choose which variables to stratify upon after data collection	Few covariables can be considered
		Few strata per covariable leads to less complete control of confounding
		Relevant covariables must have been measured
Statistical adjustment	Multiple confounders can be controlled simultaneously	Model may not fit:
	Information in continuous variables can be fully used	Incomplete control of confounding (if model does not fit confounder-outcome relationship)
	Flexible and reversible	Inaccurate estimates of strength of effect (if model does not fit predictor-outcome relationship)
		Results may be hard to understand
		Relevant covariables must have been measured

according to the level of the potential confounder and then examining the relation between the predictor and outcome separately in each stratum. Stratification is illustrated in Appendix 9.A. By considering smokers and nonsmokers separately ("stratifying on smoking"), the confounding effects of smoking can be removed.

Like matching, stratification is easily understood. An advantage of stratification is its flexibility: by performing several stratified analyses, the investigators can decide which variables appear to be confounders and ignore the remainder. (This may be done by determining whether the results of stratified analyses substantially differ from those of unstratified analyses; see Appendix 9.A.) No choices need be made at the beginning of the study that might later be regretted.

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, systolic blood pressure, serum cholesterol, cigarette smoking, and alcohol intake. To stratify on these five variables, even if there were only three strata for each, would require 3^5 (= 243) strata! With this many strata there will be some that have zeroes in the margins (e.g., strata with cases but no controls), and data in these strata cannot be used.

To maintain a sufficient number of subjects in each stratum, a variable is often divided into just two strata. When the strata are too broad, however, the confounder may not be adequately controlled. For example, if the preceding study stratified using only two age strata (e.g., age < 50 and age > 50), confounding would still be possible if within each stratum the subjects drinking the most coffee were also the oldest and thus at highest 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 ingestion on 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. 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 9.B. Similar adjustments can be made for several confounders simultaneously, using software for multivariate analysis.

One of the **great advantages of multivariate adjustment** techniques is the capacity to control the influence of many confounders simultaneously. Another advantage is their 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 two or three categories.

There are, however, two **disadvantages of multivariate adjustment.** First, 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 to the particular study. Taking the example in Appendix 9.B, the investigator should examine the data to see 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., quadratic), 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 highly derived statistics are difficult to understand intuitively. This is particularly a problem if a simple model does not fit and transformations (e.g., parental education squared) or interaction terms (e.g., child sex times parental education) are needed.

■ CHOOSING A STRATEGY

What general guidelines can be offered for when to use each of these strategies? 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 in every study of establishing criteria for selecting the study subjects (Chapter 3).

An important decision to make in the design phase of the study is whether to match. Matching is most appropriate for 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, and in situations where it is clear that analysis-phase strategies are not as good.

The decision to stratify or adjust can wait until after the data are collected and the investigator can analyze the data to see which factors are potential confounders (i.e., associated with both the predictor of interest and the outcome). However, it is important to consider which factors may be used for adjustment at the time the study is designed, in order to know which variables to measure. Also, since

strategies for controlling the influence of a specific confounding variable can only succeed to the degree that the confounder is well measured, it is important to design measurement approaches that have adequate precision and accuracy (Chapter 4).

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 9.A: the very 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 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. For example, if heavier coffee drinkers also were heavier smokers, their MI risk would be greater than that of moderate coffee drinkers.

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. It is important not to overemphasize biologic plausibility, however. Investigators can come up with a suggested mechanism for virtually any association.

■ SUMMARY

1. The design of observational studies should anticipate the need to interpret **associations.** The inference that the association represents a **cause-effect** relation is strengthened by strategies that reduce the likelihood of the **four rival explanations—chance, bias, effect-cause,** and **confounding.**
2. The role of **chance** can be minimized by designing a study with **adequate sample size and precision** to assure a low Type I error rate. Once the study is completed, the likelihood that chance is the basis of the association can be

judged from the P value and the consistency of the results with previous evidence.

3. **Bias** 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 avoided 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** is made less likely by the following strategies:
   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 sparingly because they can irreversibly limit the information available from the study.
   b. **Stratification** or **adjustment** in the analysis phase, which accomplishes the same goal statistically and preserves more options for coping with confounders. Adjustment permits many factors to be controlled simultaneously, but the data may not fit the statistical model and the resulting measures of association may not be easy to grasp intuitively.
6. Causal inference is further enhanced by positive evidence: the **consistency and strength of the association,** the presence of a **dose-response** relation, and **biologic plausibility.**

EXERCISES

1. You are planning a case-control study to address the research question, "Does eating more fruits and vegetables reduce the risk of coronary heart disease (CHD) in the elderly?" Suppose that your study shows that people in the control group report a higher intake of fruits and vegetables than people with CHD.

What are the possible explanations for this inverse association between intake of fruits and vegetables and CHD? How could each of these possibilities be altered in the design phase of the study? How could they be addressed in the analysis phase?

Give special attention to the possibility that the association between eating fruits and vegetables and CHD may be confounded by exercise (if people who eat more fruits and vegetables also exercise more, and this is the cause of their lower CHD rates). What approaches could you use to cope with exercise as a possible confounder, and what are the advantages and disadvantages of each plan?

References

1. Jacobs D, Blackburn H, Higgins M, et al. Report of the Conference on Low Blood Cholesterol: mortality associations. *Circulation* 1992;86:1046–60.
2. Pekkanen J, Nissinen A, Punsar S, Karvonen MJ. Short- and long-term association of serum cholesterol with mortality: the 25-year follow-up of the Finnish cohorts of the seven countries study. *Am J Epidemiol* 1992;135:1251–8.
3. Sharp SJ, Pocock SJ. Time trends in serum cholesterol before cancer death. *Epidemiology* 1997;8:132–6.

4. Davidson T, Caldwell E, Curtis J, et al. Reduced quality of life in survivors of acute respiratory distress syndrome compared with critically ill control patients. *JAMA* 1999;281:354–60.
5. Shands KN, Schmid GP, Dan BB, et al. Toxic-shock syndrome in menstruating women: association with tampon use and *Staphylococcus aureus* and clinical features in 52 cases. *N Engl J Med* 1980;303:1436–42.

■ APPENDIX 9.A

Hypothetical Example of Confounding

Confounding by cigarette smoking can be the cause of an apparent association between coffee drinking and MI. The entries in these tables are numbers of subjects. Thus the top left entry of Panel 1 means that, among the smokers, 80 MI cases drank coffee (out of 80 + 20 = 100 total MI cases).

1. Both in smokers and in nonsmokers, coffee drinking is not associated with MI.

	Smokers		Nonsmokers	
	MI	No MI	MI	No MI
Coffee	80	40	10	20
No Coffee	20	10	40	80

Odds ratio for MI associated with coffee:

$$\text{in smokers} \quad \frac{80 \times 10}{20 \times 40} = 1$$

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

2. However, if we did not stratify on smoking (i.e., if we did not consider smokers or nonsmokers separately), coffee drinking and MI would appear to be related. Combining the two preceding tables gives

	Smokers and Nonsmokers Combined	
	MI	No MI
Coffee	90	60
No Coffee	60	90

Odds ratio for MI associated with coffee:

$$\text{in smokers and nonsmokers combined} = \frac{90 \times 90}{60 \times 60} = 2.25$$

3. Smoking is a confounder because it is strongly associated with coffee drinking (below, left) and with MI (below, right):*

	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

$$\text{Odds ratio for coffee drinking associated with smoking} = \frac{120 \times 120}{30 \times 30} = 16 \qquad \text{Odds ratio for MI associated with smoking} = \frac{100 \times 100}{50 \times 50} = 4$$

*These tables were obtained by rearranging numbers in Panel 1 and then combining tables

APPENDIX 9.B

A Simplified Example of Adjustment

Suppose that a study finds two major predictors of the 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 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 (Fig. 9.2).*

The dotted line in Fig. 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 five points for each 2 years of parental education. Thus, 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 point *A* to point *A'*. (Because

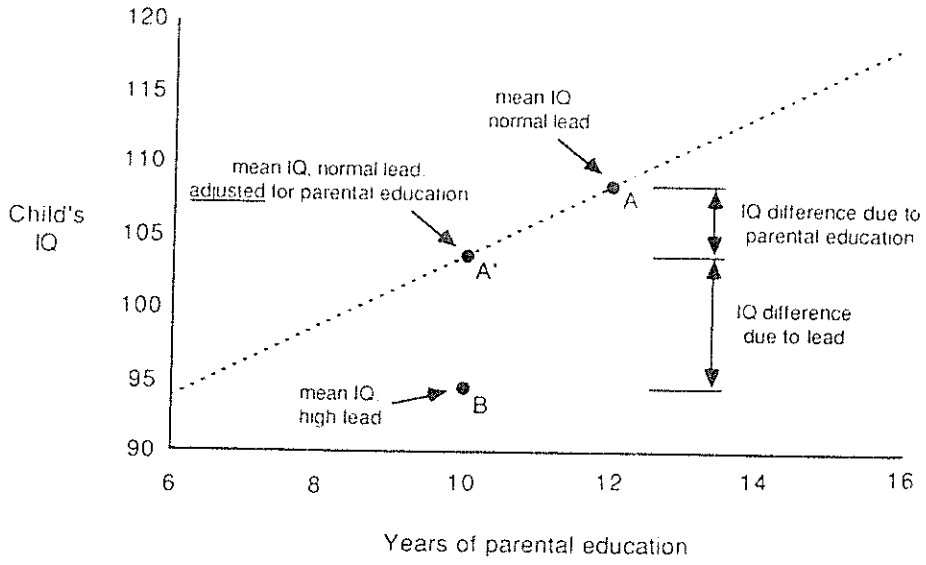


FIGURE 9.2

Hypothetical graph of child's IQ as a linear function (*dotted line*) of years of parental education.

* 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 thus assumes linear relationships between education and IQ in both groups, and that the slopes of the lines in the two groups are the same.

the group with normal lead levels had 2 more years of parental education on the average, we adjust their IQs downward by five 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. Thus of the 15-point difference in IQ of children with low and high lead levels, five points can be accounted for by their parents' different education levels and the remaining 10 are attributable to the lead exposure.