Chapter 7
CONFOUNDING

Suppose an investigator conducted an observational study to answer the causal question "does one's looking up to the sky make other pedestrians look up too?" She found an association between a first pedestrian's looking up and a second one's looking up. However, she also found that pedestrians tend to look up when they hear a thunderous noise above. Thus it was unclear what was making the second pedestrian look up, the first pedestrian's looking up or the thunderous noise? She concluded the effect of one's looking up was confounded by the presence of a thunderous noise.

In randomized experiments treatment is assigned by the flip of a coin, but in observational studies treatment (e.g., a person's looking up) may be determined by many factors (e.g., a thunderous noise). If those factors affect the risk of developing the outcome (e.g., another person's looking up), then the effects of those factors become entangled with the effect of treatment. We then say that there is confounding, which is just a form of lack of exchangeability between the treated and the untreated. Confounding is often viewed as the main shortcoming of observational studies. In the presence of confounding, the old adage "association is not causation" holds even if the study population is arbitrarily large. This chapter provides a definition of confounding and reviews the methods to adjust for it.

7.1 The structure of confounding

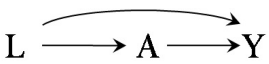

Figure 7.1

In a causal DAG, a backdoor path is a noncausal path between treatment and outcome that remains even if all arrows pointing from treatment to other variables (the descendants of treatment) are removed. That is, the path has an arrow pointing into treatment.

The structure of confounding, the bias due to common causes of treatment and outcome, can be represented by using causal diagrams. For example, the diagram in Figure 7.1 (same as Figure 6.1) depicts a treatment A, an outcome Y, and their shared (or common) cause L. This diagram shows two sources of association between treatment and outcome: 1) the path $A \rightarrow Y$ that represents the causal effect of A on Y, and 2) the path $A \leftarrow L \rightarrow Y$ between A and Y that includes the common cause L. The path $A \leftarrow L \rightarrow Y$ that links A and Y through their common cause L is an example of a *backdoor path*.

If the common cause L did not exist in Figure 7.1, then the only path between treatment and outcome would be $A \rightarrow Y$, and thus the entire association between A and Y would be due to the causal effect of A on Y. That is, the associational risk ratio $\Pr[Y = 1|A = 1] / \Pr[Y = 1|A = 0]$ would equal the causal risk ratio $\Pr\left[Y^{a=1} = 1\right] / \Pr\left[Y^{a=0} = 1\right]$; association would be causation. But the presence of the common cause L creates an additional source of association between the treatment A and the outcome Y, which we refer to as confounding for the effect of A on Y. Because of confounding, the associational risk ratio does not equal the causal risk ratio; association is not causation.

Examples of confounding abound in observational research. Consider the following examples of confounding for the effect of various kinds of treatments on health outcomes:

- Occupational factors: The effect of working as a firefighter A on the risk of death Y will be confounded if "being physically fit" L is a cause of both being an active firefighter and having a lower mortality risk. This

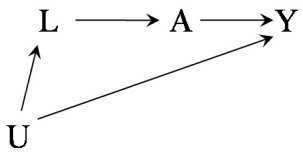

Figure 7.2

Some authors prefer to replace the unmeasured common cause U (and the two arrows leaving it) by a bidirectional edge between the measured variables that U causes.

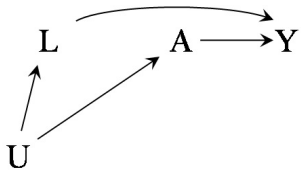

Figure 7.3

Early statistical descriptions of confounding were provided by Yule (1903) for discrete variables and by Pearson et al. (1899) for continuous variables. Yule described the association due to confounding as "fictitious", "illusory", and "apparent". Pearson et al. (1899) referred to it as a "spurious" correlation. However, there is nothing fictitious, illusory, apparent, or spurious about these associations. Associations due to common causes are quite real associations, though they cannot be causally interpreted as treatment effects. Or, in Yule's words, they are associations "to which the most obvious physical meaning must not be assigned."

bias, depicted in the causal diagram in Figure 7.1, is often referred to as a *healthy worker bias*.

- Clinical decisions: The effect of drug A (say, aspirin) on the risk of disease Y (say, stroke) will be confounded if the drug is more likely to be prescribed to individuals with certain condition L (say, heart disease) that is both an indication for treatment and a risk factor for the disease. Heart disease L is a risk factor for stroke Y because L has a direct causal effect on Y as in Figure 7.1 or, as in Figure 7.2, because both L and Y are caused by atherosclerosis U, an unmeasured variable. This bias is known as *confounding by indication* or *channeling*, the last term often being reserved to describe the bias created by patient-specific risk factors L that encourage doctors to use certain drug A within a class of drugs.

- Lifestyle: The effect of behavior A (say, exercise) on the risk of Y (say, death) will be confounded if the behavior is associated with another behavior L (say, cigarette smoking) that has a causal effect on Y and tends to co-occur with A. The structure of the variables L, A, and Y is depicted in the causal diagram in Figure 7.3, in which the unmeasured variable U represents the sort of personality and social factors that lead to both lack of exercise and smoking. Another frequent problem: subclinical disease U results both in lack of exercise A and an increased risk of clinical disease Y. This form of confounding is often referred to as *reverse causation* when L is unknown.

- Genetic factors: The effect of a DNA sequence A on the risk of developing certain trait Y will be confounded if there exists a DNA sequence L that has a causal effect on Y and is more frequent among people carrying A. This bias, also represented by the causal diagram in Figure 7.3, is known as *linkage disequilibrium* or *population stratification*, the last term often being reserved to describe the bias arising from conducting studies in a mixture of individuals from different ethnic groups. Thus the variable U can stand for ethnicity or other factors that result in linkage of DNA sequences.

- Social factors: The effect of income at age 65 A on the level of disability at age 75 Y will be confounded if the level of disability at age 55 L affects both future income and disability level. This bias may be depicted by the causal diagram in Figure 7.1.

- Environmental exposures: The effect of airborne particulate matter A on the risk of coronary heart disease Y will be confounded if other pollutants L whose levels co-vary with those of A cause coronary heart disease. This bias is also represented by the causal diagram in Figure 7.3, in which the unmeasured variable U represent weather conditions that affect the levels of all types of air pollution.

In all these cases, the bias has the same structure: it is due to the presence of a cause (L or U) that is shared by the treatment A and the outcome Y, which results in an open backdoor path between A and Y. We refer to the bias caused by shared causes of treatment and outcome as confounding, and we use other names to refer to biases caused by structural reasons other than the presence of shared causes of treatment and outcome. For simplicity of presentation, we assume throughout this chapter that positivity and consistency hold, that all nodes in the causal diagrams are perfectly measured, that

there are no selection nodes S with a box around them (that is, the data are a random sample from the population of interest), and that random variability is absent. Causal diagrams with selection nodes will be discussed in Chapter 8, and causal diagrams with mismeasured nodes in Chapter 9. Random variability is discussed in Chapter 10.

7.2 Confounding and exchangeability

See Greenland and Robins (1986, 2009) for a detailed discussion on the relations between confounding and exchangeability.

We now link the concept of confounding, which we have defined using causal diagrams, with the concept of exchangeability, which we have defined using counterfactuals in earlier chapters.

When exchangeability $Y^a \perp\!\!\!\perp A$ holds, as in a marginally randomized experiment in which all individuals have the same probability of receiving treatment, the average causal effect can be identified without adjustment for any variables. For a binary treatment A, the average causal effect $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$ is calculated as the difference of conditional means $\mathrm{E}[Y|A=1] - \mathrm{E}[Y|A=0]$.

When exchangeability $Y^a \perp\!\!\!\perp A$ does not hold but conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ does, as in a conditionally randomized experiment in which the probability of receiving treatment varies across values of L, the average causal effect can also be identified. However, as we described in Chapter 2, identification of the causal effect $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$ in the population requires adjustment for the variables L via standardization or IP weighting. Also, as we described in Chapter 4, conditional exchangeability also allows the identification of the conditional causal effects $\mathrm{E}[Y^{a=1}|L=l] - \mathrm{E}[Y^{a=0}|L=l]$ for any value l via stratification.

Under conditional exchangeability, $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}] = \sum_l \mathrm{E}[Y|L=l, A=1]\Pr[L=l] - \sum_l \mathrm{E}[Y|L=l, A=0]\Pr[L=l]$.

In practice, if we believe confounding is likely, a key question arises: can we determine whether there exists a set of measured covariates L for which conditional exchangeability holds? Answering this question is difficult because thinking in terms of conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ is often not intuitive in complex causal systems.

In this chapter, we will see that answering this question is possible if one knows the causal DAG that generated the data. To do so, suppose that we know the true causal DAG (for now, it doesn't matter how we know it: perhaps we have sufficient subject-matter knowledge, or perhaps an omniscient god gave it to us). How does the causal DAG allow us to determine whether there exists a set of variables L for which conditional exchangeability holds? There are two main approaches: (i) the backdoor criterion applied to the causal DAG and (ii) the transformation of the causal DAG into a SWIG. Though the use of SWIGs is a more direct approach, it also requires a bit more machinery so we are going to first explain the backdoor criterion; we will describe the SWIG approach in Section 7.5.

Pearl (1995, 2009) proposed the backdoor criterion for nonparametric identification of causal effects.

A set of covariates L satisfies the *backdoor criterion* if all backdoor paths between A and Y are blocked by conditioning on L and L contains no variables that are descendants of treatment A. Under faithfulness and a further condition discussed in Technical Point 7.1, conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ holds if and only if L satisfies the backdoor criterion. (A simple proof of this fact will be given below based on SWIGs.) Hence, we can now answer any query we may have about whether, for a given set of covariates L, conditional exchangeability given L holds. Thus, by trying every subset of measured non-descendants of treatment, we can answer the question of whether conditional exchangeability holds for any subset. (In fact, algorithms exist that can greatly reduce the

Technical Point 7.1

Does conditional exchangeability imply the backdoor criterion? That L satisfies the backdoor criterion always implies conditional exchangeability given L, even in the absence of faithfulness. In the main text we also said that, given faithfulness, conditional exchangeability given L implies that L satisfies the backdoor criterion. This last sentence is true under an FFRCISTG model (see Technical Point 6.2). In contrast, under an NPSEM-IE model, conditional exchangeability can hold even if the backdoor criterion does not, as is the case in a causal DAG with nodes A, L, Y and arrows $A \rightarrow L$, $A \rightarrow Y$. In this book we always assume an FFRCISTG model and faithfulness, unless stated otherwise.

This difference between causal models is due to the fact that the NPSEM-IE, unlike an FFRCISTG model, assumes cross-world independencies between counterfactuals. However a cross-world independence can never be verified, even in principle, by any randomized experiment, which was the very reason that Robins (1986, 1987) did not assume cross-world independencies in his FFRCISTG model. We will return to this issue in Chapter 23.

number of subsets that must be tried in order to answer the question.)

Let us now relate the backdoor criterion (i.e., exchangeability) to confounding. The two settings in which the backdoor criterion is satisfied are

1. *No common causes of treatment and outcome.* In Figure 6.2, there are no common causes of treatment and outcome, and hence no backdoor paths that need to be blocked. Then the set of variables that satisfies the backdoor criterion is the empty set and we say that there is no confounding.

2. *Common causes of treatment and outcome but a subset L of measured non-descendants of A suffices to block all backdoor paths.* In Figure 7.1, the set of variables that satisfies the backdoor criterion is L. Thus, we say that there is confounding, but that there is no residual confounding whose elimination would require adjustment for unmeasured variables (which, of course, is not possible). For brevity, we say that there is *no unmeasured confounding*.

The first setting describes a marginally randomized experiment in which confounding is not expected because treatment assignment is solely determined by the flip of a coin—or its computerized upgrade: the random number generator—and the flip of the coin cannot cause the outcome. That is, when the treatment is unconditionally randomly assigned, the treated and the untreated are expected to be exchangeable because no common causes exist or, equivalently, because there are no open backdoor paths. Marginal exchangeability, i.e., $Y^a \perp\!\!\!\perp A$, is equivalent to no common causes of treatment and outcome.

The second setting describes a conditionally randomized experiment in which the probability of receiving treatment is the same for all individuals with the same value of L but, by design, this probability varies across values of L, that is there is an arrow $L \rightarrow A$. This experimental design guarantees confounding if L is also either a cause of the outcome (as in Figure 7.1) or the descendant of an unmeasured cause of the outcome as in Figure 7.2. Hence, there are open backdoor paths. However, conditioning on the covariates L will block all backdoor paths and therefore conditional exchangeability, i.e., $Y^a \perp\!\!\!\perp A|L$, will hold. We say that a set L of measured non-descendants of A is a *sufficient set for confounding adjustment* when conditioning on L blocks all backdoor paths—that is, the treated and the untreated are exchangeable within levels of L.

Take our heart transplant study, a conditionally randomized experiment, as an example. Individuals who received a transplant ($A = 1$) are different from the others ($A = 0$) because, had the treated remained untreated, their risk of death Y would have been higher than that of those that were actually untreated—the treated had a higher frequency of severe heart disease L, a common cause of A and Y. The presence of common causes of treatment and outcome implies that the treated and the untreated are not marginally exchangeable but are conditionally exchangeable given L. This second setting is also what one hopes for in observational studies in which many variables L have been measured.

The backdoor criterion does not answer questions regarding the magnitude or direction of confounding. It is logically possible that some unblocked backdoor paths are weak (e.g., if L does not have a large effect on either A or Y) and thus induce little bias, or that several strong backdoor paths induce bias in opposite directions and thus result in a weak net bias. Because unmeasured confounding is not an "all or nothing" issue, in practice, it is important to consider the expected direction and magnitude of the bias (see Fine Point 7.1).

7.3 Confounding and the backdoor criterion

We now describe several examples of the application of the backdoor criterion to determine whether the causal effect of A on Y is identifiable and, if so, which variables are required to ensure conditional exchangeability. Remember that all causal DAGs in this chapter include perfectly measured nodes that are not conditioned on.

In Figure 7.1 there is confounding because the treatment A and the outcome Y share the cause L, i.e., because there is an open backdoor path between A and Y through L. However, this backdoor path can be blocked by conditioning on L. Thus, if the investigators collected data on L for all individuals, there is no unmeasured confounding given L.

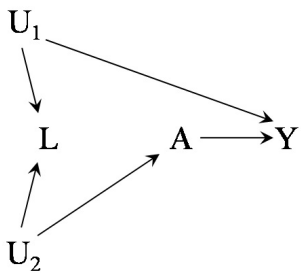

Figure 7.4

In Figure 7.2 there is confounding because the treatment A and the outcome Y share the unmeasured cause U, i.e., there is a backdoor path between A and Y through U. (Unlike the variables L, A, and Y, the variable U was not measured by the investigators.) This backdoor path could be theoretically blocked, and thus confounding eliminated, by conditioning on U, had data on this variable been collected. However, this backdoor path can also be blocked by conditioning on L. Thus, there is no unmeasured confounding given L.

In Figure 7.3 there is also confounding because the treatment A and the outcome Y share the cause U, and the backdoor path can also be blocked by conditioning on L. Therefore there is no unmeasured confounding given L.

Now consider Figure 7.4. In this causal diagram there are no common causes of treatment A and outcome Y, and therefore there is no confounding. The backdoor path between A and Y through L ($A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$) is blocked because L is a collider on that path. Thus all the association between A and Y is due to the effect of A on Y: association is causation. For example, suppose A represents physical activity, Y cervical cancer, U_1 a pre-cancer lesion, L a diagnostic test (Pap smear) for pre-cancer, and U_2 a health-conscious personality (more physically active, more visits to the doctor). Then, under the causal diagram in Figure 7.4, the effect of physical activity A on cancer Y is unconfounded and there is no need to adjust for L to compute either $\Pr[Y^{a=1}]$ or $\Pr[Y^{a=0}]$ and thus to compute the causal effect in the population.

Fine Point 7.1

The strength and direction of confounding bias. Suppose you conducted an observational study to identify the effect of heart transplant A on death Y and that you assumed no unmeasured confounding. A thoughtful critic says "the inferences from this observational study may be incorrect because of potential confounding due to cigarette smoking L." A crucial question is whether the bias results in an attenuated or an exaggerated estimate of the effect of heart transplant. For example, suppose that the risk ratio from your study was 0.6 (heart transplant was estimated to reduce mortality during the follow-up by 40%) and that, as the reviewer suspected, cigarette smoking L is a common cause of A (cigarette smokers are less likely to receive a heart transplant) and Y (cigarette smokers are more likely to die). Because there are fewer cigarette smokers ($L = 1$) in the heart transplant group ($A = 1$) than in the other group ($A = 0$), one would have expected to find a lower mortality risk in the group $A = 1$ even under the null hypothesis of no effect of treatment A on Y. Adjustment for cigarette smoking will therefore move the effect estimate upwards (say, from 0.6 to 0.7). In other words, lack of adjustment for cigarette smoking resulted in an exaggeration of the beneficial average causal effect of heart transplant.

An approach to predict the direction of confounding bias is the use of *signed causal diagrams*. Consider the causal diagram in Figure 7.1 with dichotomous L, A, and Y variables. A positive sign over the arrow from L to A is added if L has a positive average causal effect on A (i.e., if the probability of $A = 1$ is greater among those with $L = 1$ than among those with $L = 0$), otherwise a negative sign is added if L has a negative average causal effect on A (i.e., if the probability of $A = 1$ is greater among those with $L = 0$ than among those with $L = 1$). Similarly a positive or negative sign is added over the arrow from L to Y. If both arrows are positive or both arrows are negative, then the confounding bias is said to be positive, which implies that effect estimate will be biased upwards in the absence of adjustment for L. If one arrow is positive and the other one is negative, then the confounding is said to be negative, which implies that the effect estimate will be biased downwards in the absence of adjustment for L. Unfortunately, this simple rule may fail in more complex causal diagrams or when the variables are not dichotomous. See VanderWeele, Hernán, and Robins (2008) for a more detailed discussion of signed diagrams in the context of average causal effects.

Regardless of the sign of confounding, another key issue is the magnitude of the bias. Biases that are not large enough to affect the conclusions of the study may be safely ignored in practice, whether the bias is upwards or downwards. A large confounding bias requires a strong confounder-treatment association and a strong confounder-outcome association (conditional on the treatment). For discrete confounders, the magnitude of the bias depends also on prevalence of the confounder (Cornfield et al. 1959, Walker 1991). If the confounders are unknown, one can only guess what the magnitude of the bias is. Educated guesses can be organized by conducting sensitivity analyses (i.e., repeating the analyses under several assumptions regarding the magnitude of the bias), which may help quantify the maximum bias that is reasonably expected. See Rosenbaum (2005), Greenland (1996a), Robins, Rotnitzky, and Scharfstein (1999), Greenland and Lash (2008), and VanderWeele and Arah (2011) for detailed descriptions of sensitivity analyses for unmeasured confounding.

An informal definition for Figures 7.1 to 7.4: 'A confounder is any variable that can be used to adjust for confounding.' Note this definition is not circular because we have previously provided a definition of confounding. Another example of a non-circular definition: "A musician is a person who plays music," stated after we have defined what music is.

Suppose, as in the last four examples, that data on L, A, and Y suffice to identify the causal effect. In such setting we define L to be a *confounder* if the data on A and Y do not suffice for identification (i.e., we have structural confounding). We define L to be a non-confounder if data on A, Y alone suffice for identification. These definitions are equivalent to defining L as a confounder if there is conditional exchangeability but not unconditional exchangeability (i.e., structural confounding) and as a non-confounder if there is unconditional exchangeability.

Thus, in Figures 7.1-7.3, L is a confounder because $\Pr[Y^a = 1]$ is identified by the standardized risk $\sum_l \Pr[Y = 1|A = a, L = l] \Pr[L = l]$. In Figures 7.2 and 7.3, L is not a common cause of A and Y, yet we still say that L is a confounder because it is needed to block the open backdoor path attributable to the unmeasured common cause U of A and Y. In Figure 7.4, L is a non-confounder and the identifying formula for $\Pr[Y^a = 1]$ is just the conditional

mean $\Pr[Y = 1|A = a]$.

Interestingly, in Figure 7.4, conditional exchangeability given L does not hold and thus the counterfactual risks $\Pr[Y^a = 1|L = l]$ are not equal to the stratum-specific risks $\Pr[Y = 1|A = a, L = l]$, and the conditional treatment effects with strata of L are not identified. Further, adjustment for L via standardization $\sum_l \Pr[Y = 1|A = a, L = l] \Pr[L = l]$ gives a biased estimate of $\Pr[Y^a]$. This follows from the fact that adjustment for L would induce bias because conditioning on the collider L opens the backdoor path between A and Y ($A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$), which was previously blocked by the collider itself. Thus the association between A and Y would be a mixture of the association due to the effect of A on Y and the association due to the open backdoor path. Association would not be causation any more. This is the first example we have seen for which unconditional exchangeability holds but conditional exchangeability does not: the average causal effect is identified, but generally not the conditional causal effects within levels of L. We refer to the resulting bias in the conditional effect as selection bias because it it arises from selecting (conditioning) on the common effect L of two marginally independent variables U_1 and U_2, one of which is associated with A and the other with Y (see Chapter 8).

The possibility of identification of unconditional effects without identification of conditional effects was non-graphically demonstrated by Greenland and Robins (1986). The conditional bias in Figure 7.4 was described by Greenland, Pearl, and Robins (1999) and referred to as M-bias (Greenland 2003) because the structure of the variables involved in it—U_2, L, U_1—resembles a letter M lying on its side.

If U_1 caused U_2, or U_2 caused U_1, or an unmeasured U_3 caused both, there would exist a common cause of A and Y, and we would have neither unconditional nor conditional exchangeability given L.

The definition of collider is path-specific: L is a collider on the path $A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$, but not on the path $A \leftarrow L \leftarrow U_1 \rightarrow Y$.

The causal diagram in Figure 7.5 is a variation of the one in Figure 7.4. The difference is that, in Figure 7.5, there is an arrow $L \rightarrow A$. The presence of this arrow creates an open backdoor path $A \leftarrow L \leftarrow U_1 \rightarrow Y$ because U_1 is a common cause of A and Y, and so confounding exists. Conditioning on L would block that backdoor path but would simultaneously open a backdoor path on which L is a collider ($A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$).

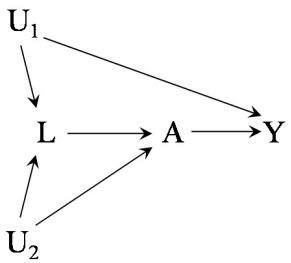

Figure 7.5

Therefore, in Figure 7.5, the bias is intractable: attempting to block the confounding path opens a selection bias path. There is neither unconditional exchangeability nor conditional exchangeability given L. A solution to the bias in Figure 7.5 would be to measure either (i) a variable L_1 between U_1 and either A or Y, or (ii) a variable L_2 between U_2 and either A or L. In the first case we would have conditional exchangeability given L_1. In the second case we would have conditional exchangeability given both L_2 and L. For example, Figure 7.6 includes the variable L_1 between U_1 and Y and the variable L_2 between U_2 and A. See Fine Point 7.2 for a discussion of identification of causal effects depending on what variables are measured in Figure 7.6.

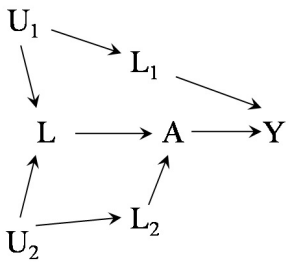

Figure 7.6

The causal diagrams in this section depict two structural sources of lack of exchangeability that are due to the presence of open backdoor paths between treatment and outcome. The first source is the presence of common causes of treatment and outcome—which creates an open backdoor path. The second source is conditioning on a common effect—which may open a previously blocked backdoor path. For pedagogic purposes, we have reserved the term "confounding" for the first and "selection bias" for the latter. An alternative way to structurally define confounding could be the "bias due to an open backdoor path between A and Y." This alternative definition is identical to ours except that it labels the bias due to conditioning on L in Figure 7.4 as confounding rather than as selection bias. The alternative definition can be equivalently expressed as follows: confounding is "any systematic bias that would be eliminated by randomized assignment of A". To see this, note that the bias induced in Figure 7.4 by conditioning on L could not occur in an experiment in which treatment A is randomly assigned because the random assignment ensures the absence of an unmeasured U_2 that is a common cause of A and L and thus conditioning on L would no longer open a backdoor path.

One interesting distinction between these two definitions is the following.

Fine Point 7.2

Identification of conditional and unconditional effects. Under any causal diagram, the causal effects that can be identified depend on the variables that are measured in addition to the treatment and the outcome. Take Figure 7.6 as an example. If we measure only L_2 (but not L and L_1), we have neither unconditional nor conditional exchangeability given L_2, and no causal effects can be identified. If we measure L_2 and L, we have conditional exchangeability given L_2 and L, but we do not have conditional exchangeability given either L_2 alone or L alone. However, we can identify:

- The conditional causal effects within joint strata of L_2 and L. The identifying formula for each of the counterfactual means is $\mathrm{E}[Y|A=a, L=l, L_2=l_2]$.
- The unconditional causal effect. The identifying formula for each of the counterfactual means is $\sum_{l,l_2} \mathrm{E}[Y|A=a, L=l, L_2=l_2] \Pr[L=l, L_2=l_2]$.
- The conditional causal effects within strata of L. The identifying formula for each of the counterfactual means is $\sum_{l_2} \mathrm{E}[Y|A=a, L=l, L_2=l_2] \Pr[L_2=l_2|L=l]$.
- The conditional causal effects within strata of L_2. The identifying formula for each of the counterfactual means is $\sum_{l} \mathrm{E}[Y|A=a, L=l, L_2=l_2] \Pr[L=l|L_2=l_2]$.

If we only measure L_1, then we have conditional exchangeability given L_1 so we can identify the conditional causal effects within strata of L_1 and the unconditional causal effect. If we measure L_1 and L, then we can also identify the conditional causal effects within joint strata of L_1 and L, and within strata of L alone. If we measure L, L_1, and L_2, then we can also identify the conditional effects within joint strata of all three variables.

The existence of a common cause of treatment and the outcome (the structural definition of confounding) is a substantive fact about the study population and the world, independent of the method chosen to analyze the data. On the other hand, the definition of confounding as any bias that would have been eliminated by randomization implies that the existence of confounding depends on the method of analysis. In Figure 7.4, we have no confounding if we do not adjust for L, but we introduce confounding if we do adjust.

Nonetheless, the choice of one definition over the other is just a matter of taste with no practical implications as all our conclusions regarding identifiability are based solely on whether conditional or unconditional exchangeability holds and not on our definition of confounding. The next chapter provides more detail on the distinction between confounding and selection bias.

7.4 Confounding and confounders

In the previous section, we have described how to use causal diagrams to decide whether confounding exists and, if so, to identify whether a given set of measured variables L is a sufficient set for confounding adjustment. The procedure requires a priori knowledge of the causal DAG that includes all causes—both measured and unmeasured—shared by the treatment A and the outcome Y. Once the causal diagram is known, we simply need to apply the backdoor criterion to determine what variables need to be adjusted for.

In contrast, the traditional approach to handle confounding was based mostly on observed associations rather than on prior causal knowledge. The traditional approach first labels variables that meet certain (mostly) associa-

Technically, investigators do not need structural knowledge. They only need to know a set of variables that guarantees conditional exchangeability.

tional conditions as confounders and then mandates that these so-called confounders are adjusted for in the analysis. Confounding is said to exist when the adjusted estimate differs from the unadjusted estimate.

Under the traditional approach, a confounder was defined as a variable that meets the following three conditions: (1) it is associated with the treatment, (2) it is associated with the outcome conditional on the treatment (with "conditional on the treatment" often replaced by "in the untreated"), and (3) it does not lie on a causal pathway between treatment and outcome. However, this traditional approach may lead to inappropriate adjustment. To see why, let us revisit Figures 7.1-7.4.

In Figure 7.1, the variable L is associated with the treatment (because it has a causal effect on A), is associated with the outcome conditional on the treatment (because it has a direct causal effect on Y), and it does not lie on the causal pathway between treatment and outcome. In Figure 7.2, the variable L is associated with the treatment (because it has a causal effect on A), is associated with the outcome conditional on the treatment (because it shares the cause U with Y), and it does not lie on the causal pathway between treatment and outcome. In Figure 7.3, L is associated with the treatment (it shares the cause U with A), is associated with the outcome conditional on the treatment (it has a causal effect on Y), and it does not lie on the causal pathway between treatment and outcome.

Therefore, according to the traditional approach, L is a confounder in the settings represented by Figures 7.1-7.3 and it needs be adjusted for. That was also our conclusion when using the backdoor criterion in the previous section. For Figures 7.1-7.3, there is no discrepancy between the traditional, mostly associational approach and the application of the backdoor criterion to the causal diagram.

Now consider Figure 7.4 again in which there is no confounding and L is a non-confounder by the definition given in Section 7.3. However, L meets the criteria for a traditional confounder: it is associated with the treatment (it shares the cause U_2 with A), it is associated with the outcome conditional on the treatment (it shares the cause U_1 with Y), and it does not lie on the causal pathway between treatment and outcome. Hence, according to the traditional approach, L is a confounder that should be adjusted for, even in the absence of confounding! But, as we saw above, adjustment for L results in a biased estimator of the causal effect in the population due to selection bias. Figure 7.7 is another example in which the traditional approach leads to inappropriate adjustment for L by inducing selection bias.

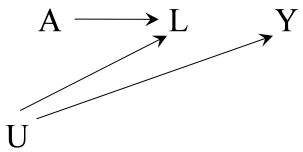

Figure 7.7

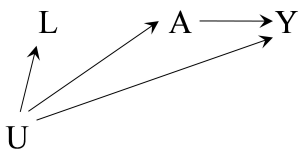

Figure 7.8

These examples show that associational or statistical criteria are insufficient to characterize confounding. An approach based on a definition of confounder that relies almost exclusively on statistical considerations may lead, as shown by Figures 7.4 and 7.7, to the wrong advice: adjust for a "confounder" even when structural confounding does not exist. To eliminate this problem for Figure 7.4, a follower of the traditional approach might replace the associational condition "(2) it is associated with the outcome conditional on the treatment" by the structural condition "(2) it is a cause of the outcome." This modified definition of confounder prevents inappropriate adjustment for L in Figure 7.4, but only to create a new problem by not considering L a confounder—that needs to be adjusted for—in Figure 7.2. See Technical Point 7.2.

The traditional approach misleads investigators into adjusting for variables when adjustment is harmful. The problem arises because the traditional approach starts by defining confounders in the absence of sufficient causal knowledge about the sources of confounding, and then mandates adjustment for

Fine Point 7.3

Surrogate confounders. Under the causal DAG in Figure 7.8, there is confounding for the effect of A on Y because of the presence of the unmeasured common cause U. The measured variable L is a proxy or surrogate for U. For example, the unmeasured variable socioeconomic status U may confound the effect of physical activity A on the risk of cardiovascular disease Y. Income L is a surrogate for the often ill-defined variable socioeconomic status. Should we adjust for the variable L? On the one hand, it can be said that L is not a confounder because it does not lie on a backdoor path between A and Y. On the other hand, adjusting for the measured L, which is associated with the unmeasured U, may indirectly adjust for some of the confounding caused by U. In the extreme, if L were perfectly correlated with U then it would make no difference whether one conditions on L or on U. Indeed if L is binary and is a nondifferentially misclassified (see Chapter 9) version of U, conditioning on L will result in a partial blockage of the backdoor path $A \leftarrow U \rightarrow Y$ under some weak conditions (Greenland 1980, Ogburn and VanderWeele 2012). Therefore we will typically prefer to adjust, rather than not to adjust, for L.

We refer to variables that can be used to reduce confounding bias even though they are not on a backdoor path (and so could never completely eliminate confounding) as *surrogate confounders*. A possible strategy to fight confounding is to measure as many surrogate confounders as possible and adjust for all of them. See Chapter 18 for discussion.

those so-called confounders. If the adjusted and unadjusted estimates differ, the traditional approach declares the existence of confounding. However, change in estimates may occur for reasons other than confounding, including selection bias when adjusting for non-confounders (see Chapter 8) and the use of noncollapsible effect measures (see Fine Point 4.3). Attempts to define confounding based on change in estimates have been long abandoned because of these problems.

In contrast, a structural approach starts by explicitly identifying the sources of confounding—the common causes of treatment and outcome that, were they all measured, would be sufficient to adjust for confounding—and then identifies a sufficient set of adjustment variables.

The structural approach makes clear that including a particular variable in a sufficient set depends on the variables already included in the set. For example, in Figures 7.2 and 7.3 the set of variables L is needed to block a backdoor path because the set of variables U is not measured. We could then say that the variables in L are confounders. However, if the variables U had been measured and used to block the backdoor path, then the variables L would not be confounders given U (see also Fine Point 7.3). Given a causal DAG, confounding is an absolute concept whereas confounder is a relative one.

VanderWeele and Shpitser (2013) also proposed a formal definition of confounder.

A structural approach to confounding emphasizes that causal inference from observational data requires a priori causal knowledge. This causal knowledge is summarized in a causal DAG that encodes the researchers' beliefs or assumptions about the causal network. Of course, there is no guarantee that the researchers' causal DAG is correct and thus it is possible that, contrary to the researchers' beliefs, their chosen set of adjustment variables fails to eliminate confounding or introduces selection bias. However, the structural approach to confounding has two important advantages. First, it prevents inconsistencies between beliefs and actions. For example, if you believe Figure 7.4 is the true causal diagram—and therefore that there is no confounding for the effect of A on Y—then you will not adjust for the variable L, regardless of what non-structural definitions of confounder may say. Second, the researchers' assumptions about confounding become explicit and therefore can be explicitly criticized by other investigators.

Technical Point 7.2

Fixing the traditional definition of confounder. Figures 7.4 and 7.7 depict two graphical examples in which the traditional non-graphical definition of confounder and confounding misleads investigators into adjusting for a variable when adjustment for such variable is not only superfluous but also harmful. The traditional definition fails because it relies on two incorrect statistical criteria—conditions (1) and (2)—and one incorrect causal criterion—condition (3). To "fix" the traditional definition one needs to do two things:

1. Replace condition (3) by the condition that "there exist variables L and U such that there is conditional exchangeability within their joint levels $Y^a \perp\!\!\!\perp A|L, U$. This new condition is stronger than the earlier condition because it effectively implies that L is not on a causal pathway between A and Y and that $\mathrm{E}[Y^a|L = l, U = u]$ is identified by $\mathrm{E}[Y|L = l, U = u, A = a]$.
2. Replace conditions (1) and (2) by the following condition: U can be decomposed into two disjoint subsets U_1 and U_2 (i.e., $U = U_1 \cup U_2$ and $U_1 \cap U_2$ is empty) such that (i) U_1 and A are not associated within strata of L, and (ii) U_2 and Y are not associated within joint strata of A, L, and U_1. The variables in U_1 may be associated with the variables in U_2. U_1 can always be chosen to be the largest subset of U that is unassociated with treatment.

If these two new conditions are met we say U is a non-confounder given data on L. These conditions were proposed by Robins (1997a, Theorem 4.3) and further discussed by Greenland, Pearl, and Robins (1999, pp. 45-46, note the condition that $U = U_1 \cup U_2$ was inadvertently left out). These conditions overcome the difficulties found in Figures 7.4 and 7.7 because they allow us to dismiss variables as non-confounders (Robins 1997a). For example, Greenland, Pearl, and Robins applied these conditions to Figure 7.4 to show that there is no confounding.

7.5 Single-world intervention graphs

Exchangeability is translated into graph language as the lack of open paths between the treatment A and outcome Y nodes—other than those originating from A—that would result in an association between A and Y. Chapters 7–9 describe different ways in which lack of exchangeability can be represented in causal diagrams. For example, in this chapter we discuss confounding, a violation of exchangeability due to the presence of an open backdoor path between treatment and outcome.

The equivalence between unconditional exchangeability $Y^a \perp\!\!\!\perp A$ and the backdoor criterion seems rather magical: there appears to be no obvious relationship between counterfactual independence and the absence of backdoor paths because counterfactuals are not included as variables on causal diagrams. Since graphs are so useful for evaluating independencies via d-separation, it seems natural to want to construct graphs that include counterfactuals as nodes, so that unconditional and conditional exchangeability can be directly read off the graph.

A new type of graph—Single-world intervention graphs (SWIGs)— unify the counterfactual and graphical approaches by explicitly including the counterfactual variables on the graph. A SWIG depicts the variables and causal relations that would be observed in a hypothetical world in which all individuals received treatment level a. That is, a SWIG is a *graph* that represents a counterfactual *world* created by a *single intervention*. In contrast, the variables on a standard causal diagram represent the actual world. A SWIG can then be viewed as a function that transforms a given causal diagram under a given intervention. The following examples describe this transformation.

Richardson and Robins (2013) showed that SWIGs overcome some of the shortcomings of previously proposed twin causal diagrams (Balke and Pearl 1994).

Suppose the causal diagram in Figure 7.2 represents the observed study

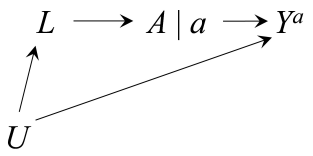

Figure 7.9

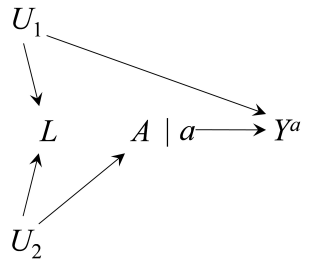

Figure 7.10

Under an FFRCISTG model, it can be shown that d-separation also implies statistical independence on the SWIG.

In the single intervention world, a is a constant and thus cannot affect other variables. When drawing SWIGs, however, we include arrows from a as a convenient way to keep track of the variables directly affected by A in the original DAG.

data. The SWIG in Figure 7.9 is a transformation of Figure 7.2 that represents a world in which all individuals have received an intervention that sets their treatment to the fixed value a.

In the SWIG, the treatment node is split into left and right sides which are to be regarded as separate nodes (variables) once split. The right side encodes the treatment value a under the intervention and inherits all the arrows that were out of A in the original causal DAG. The left side encodes the value of treatment A that would have been observed in the absence of intervention, i.e., *the natural value of treatment*. It inherits all nodes that were into A on the causal DAG because its causal inputs are the same in the intervened on (counterfactual) world as in the actual world. Note that A does not have an arrow into a because the value a is the same for all individuals, i.e., is a constant in the intervened on world.

We assume that the natural value of treatment A is well defined even though we are generally unable to measure it under intervention a. In some settings, though, A may be measurable: recent experiments suggest that electroencephalogram recordings can detect the choice individuals will make up to 1/2 second before individuals becomes conscious of their decision. If so, A could actually be measured via electroencephalogram, while still leaving 1/2 second to intervene and give treatment a.

In the SWIG, the outcome is Y^a, the value of Y in the intervened on world. Because the remaining variables are temporally prior to A, they are not affected by the intervention and therefore take the same value as in the observed world. i.e., they are not labeled as a counterfactual variable. In fact, any variable that is a non-descendant of A need not be labeled as a counterfactual because, under the faithfulness assumption (which we make), treatment has no causal effect on its non-descendants for any individual. Under our causal model, conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ holds because all paths between Y^a and A are blocked after conditioning on L, i.e., Y^a and A are d-separated given L.

Consider now the causal diagram in Figure 7.4 and the SWIG in Figure 7.10. Marginal exchangeability $Y^a \perp\!\!\!\perp A$ holds because, on the SWIG, all paths between Y^a and A are blocked (without conditioning on L). In contrast, conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ does not hold because, on the SWIG, the path $Y^a \longleftarrow U_1 \longrightarrow L \longleftarrow U_2 \longrightarrow A$ is open when the collider L is conditioned on. This is why the marginal A-Y association is causal, but the conditional A-Y association given L is not, and thus any method that adjusts for L results in bias. These examples show how SWIGs unify the counterfactual and graphical approaches. In fact it is straightforward to see that, on the SWIG, Y^a is d-separated from A given L if and only if L is a non-descendant of A that blocks all backdoor paths from A to Y (see also Fine Point 7.4).

7.6 Confounding adjustment

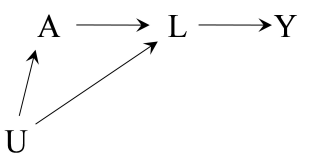

Figure 7.11

In the absence of randomization, causal inference relies on the uncheckable assumption that we have measured a set of variables L that is a *sufficient set for confounding adjustment*, i.e., a set of non-descendants of treatment A that includes enough variables to block all backdoor paths from A to Y. Under this assumption of conditional exchangeability given L, standardization and IP weighting can be used to compute the average causal effect in the population. But, as discussed in Section 4.6, standardization and IP weighting are not the only available methods to adjust for confounding in observational

Fine Point 7.4

Confounders cannot be descendants of treatment, but can be in the future of treatment. Consider the causal DAG in Figure 7.11. L is a descendant of treatment A that blocks all backdoor paths from A to Y. Unlike in Figures 7.4 and 7.7, conditioning on L does not cause selection bias because no collider path is opened. Rather, because the causal effect of A on Y is solely through the intermediate variable L, conditioning on L completely blocks this pathway. This example shows that adjusting for a variable L that blocks all backdoor paths does not eliminate bias when L is a descendant of A.

Since conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ implies that the adjustment for L eliminates all bias, it must be the case that conditional exchangeability fails to hold and the average treatment effect $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$ cannot be identified in this example. This failure can be verified by analyzing the SWIG in Figure 7.12, which depicts a counterfactual world in which A has been set to the value a. In this world, the factual variable L is replaced by the counterfactual variable L^a, i.e., the value of L that would have been observed if all individuals had received treatment value a. Since L^a blocks all paths from Y^a to A we conclude that $Y^a \perp\!\!\!\perp A|L^a$ holds, but we cannot conclude that conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ holds as L is not even on the graph. (Under an FFRCISTG, any independence that cannot be read off the SWIG cannot be assumed to hold.) Therefore, we cannot ensure that the average treatment effect $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$ is identified from data on (L, A, Y).

The problem arises because L is a descendant of A, not because L is in the future of A. If, in Figure 7.11, the arrow from A to L did not exist, then L would be a non-descendant of A that blocks all the backdoor paths. Analogously, on the SWIG in Figure 7.12, we can replace L^a by L as A is no longer a cause of L (note Y^a and A are now d-separated by L). Therefore adjusting for L would eliminate all bias, even if L were still in the future of A. What matters is the topology of the causal diagram (which variables cause which variables), not the time sequence of the nodes. Rosenbaum (1984) and Robins (1986, section 11) give non-graphical discussions of the control of confounding by temporally post-treatment variables.

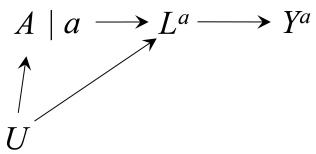

Figure 7.12

studies. Methods that adjust for confounders L can be classified into two broad categories:

- G-methods: Standardization, IP weighting, and g-estimation. These methods (the 'g' stands for 'generalized') exploit conditional exchangeability given L to estimate the causal effect of A on Y in the entire population or in any subset of the population. In our heart transplant study, we used g-methods to adjust for confounding by disease severity L in Sections 2.4 (standardization) and 2.5 (IP weighting). Part II describes model-based extensions of g-methods: the parametric g-formula (standardization), IP weighting of marginal structural models, and g-estimation of nested structural models.

- Conventional methods for stratification-based adjustment: Stratification (including restriction) and matching. These methods exploit conditional exchangeability given L to estimate the association between A and Y in subsets defined by L. In our heart transplant study, we used stratification-based methods to adjust for confounding by disease severity L in Sections 4.4 (stratification) and 4.5 (matching). Part II describes the model-based extension of conventional stratification: outcome regression.

A common variation of stratification and matching replaces each individual's variables L by the individual's estimated probability of receiving treatment $\Pr[A = 1|L]$: the *propensity score* (Rosenbaum and Rubin 1983). See Chapter 15.

Standardization and IP weighting simulate the A-Y association in the population if backdoor paths involving the measured variables L did not exist. For example, IP weighting achieves this by creating a pseudo-population in which treatment A is independent of the measured confounders L, i.e., by "deleting" the arrow from L to A. In contrast, conventional methods based on stratification do not delete the arrow from L to A but rather compute the conditional

effect in a subset of the observed population, which is represented by adding a selection box. In Part III, focused on time-varying treatments, we describe why "deleting" the arrow $L \rightarrow A$ is advantageous when using standardization or IP weighting, and why g-estimation is the only generally valid stratification-based method. The bias of conventional stratification-based methods is described in Chapter 20. In settings with time-varying treatments, and therefore time-varying confounders, g-methods are the methods of choice to adjust for confounding because conventional stratification-based methods may result in selection bias.

All the above methods require conditional exchangeability given L. However, confounding can sometimes be handled by methods that do not require conditional exchangeability. Some examples of these methods are difference-in-differences (Technical Point 7.3), instrumental variable estimation (Chapter 16), proximal inference (Technical Point 7.3), the front door criterion (Technical Point 7.4), and others. Unfortunately, these methods require alternative assumptions that, like conditional exchangeability, are unverifiable. Therefore, in practice, the validity of the resulting effect estimates is not guaranteed. The choice of adjustment method will depend on which unverifiable assumptions—either conditional exchangeability or the alternative conditions—are believed more likely to hold in a particular setting.

Achieving conditional exchangeability may be an unrealistic goal in many observational studies but, as discussed in Section 3.2, expert knowledge about the causal structure can be used to get as close as possible to that goal. Therefore, in observational studies, investigators measure many variables L (which are non-descendants of treatment) in an attempt to ensure that the treated and the untreated are conditionally exchangeable. The hope is that, even though common causes may exist (confounding), the measured variables L are sufficient to block all backdoor paths (no unmeasured confounding). However, there is no guarantee that this attempt will be successful, which makes causal inference from observational data a risky undertaking.

A practical example of the application of expert knowledge of the causal structure to confounding evaluation was described by Hernán et al (2002).

In addition, expert knowledge can be used to avoid adjusting for variables that may introduce bias. At the very least, investigators should generally avoid adjustment for variables affected by either the treatment or the outcome. Of course, thoughtful and knowledgeable investigators could believe that two or more causal structures, possibly leading to different conclusions regarding confounding and confounders, are equally plausible. In that case they would perform multiple analyses and explicitly state the assumptions about causal structure required for the validity of each. Unfortunately, one can never be certain that the set of causal structures under consideration includes the true one; this uncertainty is unavoidable with observational data.

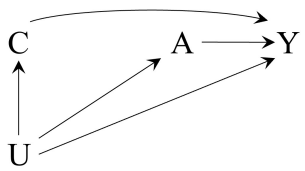

Figure 7.13

There is a scientific consequence to the always present threat of confounding in observational studies. Suppose you conducted an observational study to quantify the effect of heart transplant A on death Y. You did your best (e.g., consulting subject-matter experts) to identify and measure confounders L, and assumed no unmeasured confounding after adjusting for L. A critic of your study says "the inferences from this observational study may be incorrect because of potential confounding." The critic is not making a scientific statement, but a logical one. Since the findings from *any* observational study may be confounded, it is obviously true that those of your study can be confounded. If the critic's intent was to provide evidence about the shortcomings of your particular study, he failed. His criticism is noninformative because he simply restated a characteristic of observational research that you and the critic already knew before the study was conducted.

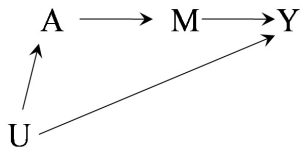

Figure 7.14

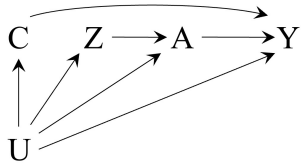

Figure 7.15

To appropriately criticize your study, the critic needs to engage in a scientific conversation. For example, the critic may cite experimental or observational evidence that contradict your findings, or he can say something along the lines of "the inferences from this observational study may be incorrect because of potential confounding due to cigarette smoking, a common cause through which a backdoor path may remain open". This latter option provides you with a testable challenge to your assumption of no unmeasured confounding. The burden of the proof is again yours. Your next move is to try and adjust for smoking or, if data on smoking could not be obtained, to conduct a sensitivity analysis to investigate the possible bias induced by smoking.

Though the above discussion was restricted to bias due to confounding, the absence of biases due to selection and measurement is also needed for valid causal inference from observational data. But, unlike confounding, these other biases may arise in *both* randomized experiments and observational studies. After having explored confounding in this chapter, the next chapter presents another potential source of lack of exchangeability between the treated and the untreated: selection of individuals into the analysis.

Technical Point 7.3

Difference-in-differences and negative outcome controls. Suppose we want to compute the average causal effect of aspirin A (1: yes; 0: no) on blood pressure Y, but there are unmeasured common causes U of A and Y such as history of heart disease. Then we cannot compute the effect via standardization or IP weighting because there is unmeasured confounding. But there is an alternative method that, under some conditions, may adjust for the unmeasured confounding: the use of negative outcome controls (also known as "placebo tests").

Suppose further that, for each individual in the population, we have also measured the value of the outcome right before treatment was available. We refer to this pre-treatment outcome C as a negative outcome control (also referred to as negative control outcome). As depicted in Figure 7.13, U is a cause of both Y and C, and treatment A is obviously not a cause of the pre-treatment C. Now, even though the causal effect of A on C is known to be zero, the contrast $\mathrm{E}[C|A=1] - \mathrm{E}[C|A=0]$ is not zero because of confounding by U. In fact, $\mathrm{E}[C|A=1] - \mathrm{E}[C|A=0]$ measures the magnitude of confounding for the effect of A on C on the additive scale. If the magnitude of additive confounding for the effect of A on the negative control outcome C is the same as for the effect of A on the true outcome Y, then we can compute the effect of A on Y in the treated. Specifically, under the assumption of additive equi-confounding $\mathrm{E}\left[Y^0|A=1\right] - \mathrm{E}\left[Y^0|A=0\right] = \mathrm{E}[C|A=1] - \mathrm{E}[C|A=0]$, the effect is

$$\mathrm{E}\left[Y^1 - Y^0|A=1\right] = (\mathrm{E}[Y|A=1] - \mathrm{E}[Y|A=0]) - (\mathrm{E}[C|A=1] - \mathrm{E}[C|A=0])$$

That is, the effect in the treated is equal to the association between treatment A and outcome Y (which is a mixture of the causal effect and confounding) minus the confounding as measured by the association between A and C. Note that the direct arrow from C to Y in Figure 7.13 is not necessary for C to be a negative outcome control.

This method for confounding adjustment is known as difference-in-differences (Card 1990, Meyer 1995, Angrist and Krueger 1999). In practice, the method is often combined with adjustment for measured covariates using parametric or semiparametric approaches (Abadie 2005). However, difference-in-differences is a somewhat restrictive approach to negative outcome controls (Sofer et al. 2016): it requires measurement of the outcome both pre- and post-treatment (or at least that the true outcome Y and the negative control outcome C are measured on the same scale) and it requires additive equi-confounding. Sofer et al. (2016) describe more general methods that allow for Y and C to be on different scales, rely on weaker versions of equi-confounding, and incorporate adjustment for measured covariates. For a general introduction to the use of negative outcome controls to detect confounding, see Lipsitch et al. (2010) and Flanders et al. (2011).

Surprisingly, when one has both a negative outcome control C and a negative treatment control Z, the causal effect can be nonparametrically identified even in the presence of unmeasured confounders U under additional assumptions. In fact, if U, C, and Z are discrete and C and Z have at least as many levels as does U, then the causal effect of A on Y will quite generally be identified (Miao et al. 2018). This identification approach is referred to as *proximal causal inference* (Cui et al. 2024). Figure 7.15 is one example in which C is a negative outcome control and Z is a negative treatment control.

Technical Point 7.4

The front door criterion. The causal diagram in Figure 7.14 depicts a setting in which the treatment A and the binary outcome Y share an unmeasured cause U, and in which there is a variable M that fully mediates the effect of A on Y and that shares no unmeasured causes with either A or Y. Under this causal structure, a data analyst cannot directly use standardization (nor IP weighting) to compute the counterfactual risks $\Pr\left[Y^{a=1}=1\right]$ and $\Pr\left[Y^{a=0}=1\right]$ because the variable U, which is necessary to block the backdoor path between A and Y, is not available. Therefore, the average causal effect of A on Y cannot be identified using the methods described in previous chapters. However, Pearl (1995) showed that $\Pr[Y^a=1]$ is identified by the so-called *front door formula*

$$\sum_m \Pr[M=m|A=a] \sum_{a'} \Pr[Y=1|M=m,A=a'] \Pr[A=a']$$

Pearl refers to this identification formula as front door adjustment because it relies on the existence of a path from A and Y that, contrary to a backdoor path, goes through a descendant M of A that completely mediates the effect of A on Y. Pearl often uses the term backdoor formula to refer to the identification formula that we refer to as standardization or the point treatment g-formula (Robins 1986). A proof of the front door identification formula follows.

Note that $\Pr[Y^a=1] = \sum_m \Pr[M^a=m] \Pr[Y^a=1|M^a=m]$ and that, under Figure 7.14, $\Pr[M^a=m] = \Pr[M=m|A=a]$ because there is no confounding for the effect of A on M (i.e., $A \perp\!\!\!\perp M^a$), and $\Pr[Y^a=1|M^a=m] = \sum_{a'} \Pr[Y=1|M=m,A=a'] \Pr[A=a']$. To prove the last equality, first note that $\Pr[Y^a=1|M^a=m] = \Pr[Y^m=1]$ because (i) $Y^a = Y^m$ when $M^a = m$ (A affects Y only through M in Figure 7.14) and (ii) $Y^m \perp\!\!\!\perp M^a$ by d-separation on a SWIG under the joint intervention in which M is set to m and A to a. Finally, by conditional exchangeability $Y^m \perp\!\!\!\perp M|A$ on the SWIG where we intervene on M alone, $\Pr[Y^m=1] = \sum_{a'} \Pr[Y=1|M=m,A=a'] \Pr[A=a']$.

The above proof requires well-defined counterfactual outcomes Y^m under interventions on M. In Technical Points 21.11 and 21.12 we present alternative proofs of the front door formula that do not require this condition.