Chapter 6
GRAPHICAL REPRESENTATION OF CAUSAL EFFECTS

Causal inference generally requires expert knowledge and untestable assumptions about the causal network linking treatment, outcome, and other variables. Earlier chapters focused on the conditions and methods to compute causal effects in oversimplified scenarios (e.g., the causal effect of your looking up on other pedestrians' behavior, an idealized heart transplant study). The goal was to provide a gentle introduction to the ideas underlying the more sophisticated approaches that are required in realistic settings. Because the scenarios we considered were so simple, there was really no need to make the causal network explicit. As we start to turn our attention towards more complex situations, however, it will become crucial to be explicit about what we know and what we assume about the variables relevant to our particular causal inference problem.

This chapter introduces a graphical tool to represent our qualitative expert knowledge and a priori assumptions about the causal structure of interest. By summarizing knowledge and assumptions in an intuitive way, graphs help clarify conceptual problems and enhance communication among investigators. The use of graphs in causal inference problems makes it easier to follow a sensible advice: draw your assumptions before your conclusions.

6.1 Causal diagrams

Comprehensive books on this subject have been written by Pearl (2009) and Spirtes, Glymour and Scheines (2000).

This chapter describes graphs, which we will refer to as causal diagrams, to represent key causal concepts. The modern theory of diagrams for causal inference arose within the disciplines of computer science and artificial intelligence. This and the next three chapters are focused on problem conceptualization via causal diagrams.

Take a look at the graph in Figure 6.1. It comprises three nodes representing random variables (L, A, Y) and three edges (the arrows). We adopt the convention that time flows from left to right, and thus L is temporally prior to A and Y, and A is temporally prior to Y. As in previous chapters, L, A, and Y represent disease severity, heart transplant, and death, respectively.

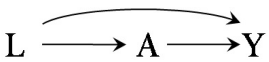

Figure 6.1

The presence of an arrow pointing from a particular variable V to another variable W indicates that we know there is a direct causal effect (i.e., an effect not mediated through any other variables on the graph) for at least one individual. Alternatively, the lack of an arrow means that we know that V has no direct causal effect on W for any individual in the population. For example, in Figure 6.1, the arrow from L to A means that disease severity affects the probability of receiving a heart transplant. A standard causal diagram does not distinguish whether an arrow represents a harmful effect or a protective effect. Furthermore, if, as in Figure 6.1, a variable (here, Y) has two causes, the diagram does not encode how the two causes interact.

Causal diagrams like the one in Figure 6.1 are known as *directed acyclic graphs*, which is commonly abbreviated as DAGs. "Directed" because the edges imply a direction: because the arrow from L to A is into A, L may cause A, but not the other way around. "Acyclic" because there are no cycles: a variable cannot cause itself, either directly or through another variable.

Directed acyclic graphs have applications other than causal inference. Here

Technical Point 6.1

Causal directed acyclic graphs. We define a directed acyclic graph (DAG) G to be a graph whose nodes (vertices) are random variables $V = (V_1, \ldots, V_M)$ with directed edges (arrows) and no directed cycles. We use PA_m to denote the parents of V_m, i.e., the set of nodes from which there is a direct arrow into V_m. The variable V_m is a descendant of V_j (and V_j is an ancestor of V_m) if there is a sequence of nodes connected by edges between V_j and V_m such that, following the direction indicated by the arrows, one can reach V_m by starting at V_j. For example, consider the DAG in Figure 6.1. In this DAG, $M = 3$ and we can choose $V_1 = L$, $V_2 = A$, and $V_3 = Y$; the parents PA_3 of $V_3 = Y$ are (L, A). We will adopt the ordering convention that if $m > j$, V_m is not an ancestor of V_j. We define the distribution of V to be Markov with respect to a DAG G (equivalently, the distribution factors according to a DAG G) if, for each j, V_j is independent of its non-descendants conditional on its parents. This latter statement is mathematically equivalent to the statement that the density $f(V)$ of the variables V in DAG G satisfies the Markov factorization

$$f(v) = \prod_{j=1}^{M} f(v_j \mid pa_j) \ .$$

A causal DAG is a DAG in which 1) the lack of an arrow from node V_j to V_m (i.e., V_j is not a parent of V_m) can be interpreted as the absence of a direct causal effect of V_j on V_m relative to the other variables on the graph, 2) all common causes, even if unmeasured, of any pair of variables on the graph are themselves on the graph, and 3) any variable is a cause of its descendants. Causal DAGs are of no practical use unless we make an assumption linking the causal structure represented by the DAG to the data obtained in a study. This assumption, referred to as the causal Markov assumption, states that, conditional on its direct causes, a variable V_j is independent of any variable for which it is not a cause. That is, conditional on its parents, V_j is independent of its non-descendants; hence, a causal DAG is Markov with respect to the DAG G.

we focus on *causal* directed acyclic graphs. A defining property of causal DAGs is that, conditional on its direct causes, any variable on the DAG is independent of any other variable for which it is not a cause. This assumption, referred to as the causal Markov assumption, implies that in a causal DAG the common causes of any pair of variables in the graph must be also in the graph. For a formal definition of causal DAGs, see Technical Point 6.1.

For example, suppose in our study individuals are randomly assigned to heart transplant A with a probability that depends on the severity of their disease L. Then L is a common cause of A and Y, and needs to be included in the graph, as shown in the causal diagram in Figure 6.1. Now suppose in our study all individuals are randomly assigned to heart transplant with the same probability regardless of their disease severity. Then L is not a common cause of A and Y and need not be included in the causal diagram. Figure 6.1 represents a conditionally randomized experiment, whereas Figure 6.2 represents a marginally randomized experiment.

Figure 6.2

Figure 6.1 may also represent an observational study. Specifically, Figure 6.1 represents an observational study in which we are willing to assume that the assignment of heart transplant A has as parent disease severity L *and no other causes of* Y. Otherwise, those causes of Y, even if unmeasured, would need to be included in the diagram, as they would be common causes of A and Y. In the next chapter we will describe how the willingness to consider Figure 6.1 as the causal diagram for an observational study is the graphic translation of the assumption of conditional exchangeability given L, $Y^a \perp\!\!\!\perp A|L$ for all a.

Many people find the graphical approach to causal inference easier to use and more intuitive than the counterfactual approach. However, the two ap-

Technical Point 6.2

Counterfactual models associated with a causal DAG. In this book, a causal DAG G represents an underlying counterfactual model. To provide a formal definition of the counterfactual model represented by a DAG G, we use the following notation. For any random variable W, let $\mathcal{W}$ denote the support (i.e., the set of possible values w) of W. For any set of ordered variables $W_1, \ldots, W_m$, define $\overline{w}_m = (w_1, \ldots, w_m)$. Let R denote any subset of variables in V and let r be a value of R. Then V_m^r denotes the counterfactual value of V_m when R is set to r.

A nonparametric structural equation model (NPSEM) represented by a DAG G with vertex set $V = (V_1, V_2, \ldots, V_M)$ (ordered such that if $i < j$ then V_i is not a descendant of V_j) assumes the existence of unobserved random variables (errors) ϵ_m and deterministic unknown functions $f_m(pa_m, \epsilon_m)$ such that $V_1 = f_1(\epsilon_1)$ and the one-step ahead counterfactual $V_m^{\overline{v}_{m-1}} \equiv V_m^{pa_m}$ is given by $f_m(pa_m, \epsilon_m)$. That is, only the parents of V_m have a direct effect on V_m relative to the other variables on G. An NPSEM implies that any variable V_j on the graph can be intervened on, as counterfactuals in which V_j has been set to a specific value v_j are assumed to exist. Both the factual variable V_m and the counterfactuals V_m^r for any $R \subset V$ are obtained recursively from V_1 and $V_j^{\overline{v}_{j-1}}$, $M \geq j > 1$. For example, $V_3^{v_1} = V_3^{v_1, V_2^{v_1}}$, i.e., the counterfactual value $V_3^{v_1}$ of V_3 when V_1 is set to v_1 is the one-step ahead counterfactual $V_3^{v_1, v_2}$ with v_2 equal to the counterfactual value $V_2^{v_1}$ of V_2. Similarly, $V_3 = V_3^{V_1, V_2^{V_1}}$ and $V_3^{v_1, v_4} = V_3^{v_1}$ because V_4 is not a direct cause of V_3. The absence of an arrow from V_j to V_k implies that V_j is not a direct cause of V_k for any individual.

Robins (1986) introduced this NPSEM, referred to it as a finest causally interpreted structural tree graph (FCISTG) "as detailed as the data", and referred to the parents PA_m of V_m as causal risk factors for V_m controlling for the earlier variables in the ordering. Pearl (2009) showed how to represent this model with a DAG. Robins (1986) also proposed often more realistic causally interpreted structural tree graphs in which only a subset of the variables are subject to intervention. For expositional purposes, we will generally assume that every variable can be intervened on, even though the statistical methods considered here do not actually require this assumption.

Richardson and Robins (2013) developed the Single World Intervention Graph (SWIG).

proaches are intimately linked. Specifically, associated with each graph is an underlying counterfactual model (see Technical Points 6.2 and 6.3). It is this model that provides the mathematical justification for the heuristic, intuitive graphical methods we now describe. However, conventional causal diagrams do not include the underlying counterfactual variables on the graph. Therefore the link between graphs and counterfactuals has traditionally remained hidden. A recently developed type of causal directed acyclic graph—the Single World Intervention Graph (SWIG)—seamlessly unifies the counterfactual and graphical approaches to causal inference by explicitly including the counterfactual variables on the graph. We defer the introduction of SWIGs until Chapter 7 as the material covered in this chapter serves as a necessary prerequisite.

Causal diagrams are a simple way to encode our subject-matter knowledge, and our assumptions, about the qualitative causal structure of a problem. But, as described in the next sections, causal diagrams also encode information about potential associations between the variables in the causal network. It is precisely this simultaneous representation of association and causation that makes causal diagrams such an attractive tool. What follows is an informal introduction to graphic rules to infer associations from causal diagrams. Our emphasis is on conceptual insight rather than on formal rigor.

6.2 Causal diagrams and marginal independence

Consider the following two examples. First, suppose you know that aspirin use A has a preventive causal effect on the risk of heart disease Y, i.e., $\Pr[Y^{a=1} =$

Technical Point 6.3

Independencies associated with counterfactual models. An FCISTG model does not imply that the causal Markov assumption of Technical Point 6.1 holds; additional statistical independence assumptions are needed. For example, Pearl (2000) usually assumed an NPSEM in which all error terms ϵ_m are mutually independent. We refer to Pearl's model with independent errors as an NPSEM-IE. In contrast, Robins (1986) only assumed that, given any $\overline{v}_M$, the one-step ahead counterfactuals $V_m^{\overline{v}_{m-1}} = f_m(pa_m, \epsilon_m)$ for $m = 1, \ldots, M$ are jointly independent where $\overline{v}_{m-1}$ is a subvector of the $\overline{v}_M$, and referred to this as the finest fully randomized causally interpreted structured tree graph (FFRCISTG) model as detailed as the data.

More precisely, Robins (1986) made the assumption that for each m, conditional on the factual past $\overline{V}_{m-1} = \overline{v}_{m-1}$, any future evolution from $m+1$ of one-step ahead counterfactuals (consistent with $\overline{v}_{m-1}$) is independent of the factual variable V_m. Robins and Richardson (2010) showed that this assumption is equivalent to the assumption of the previous paragraph for a positive distribution. In the absence of positivity, we define the model as in the last paragraph.

Robins (1986) showed his independence assumption implies that the causal Markov assumption holds. An NPSEM-IE is an FFRCISTG but not vice-versa because an NPSEM-IE makes many more independence assumptions than an FFRCISTG (Robins and Richardson 2010).

Unless stated otherwise, a DAG represents an NPSEM but we may need to specify which type. For example, the DAG in Figure 6.2 may correspond to either an NPSEM-IE that implies full exchangeability $\left(Y^{a=0}, Y^{a=1}\right) \perp\!\!\!\perp A$, or to an FFRCISTG that only implies marginal exchangeability $Y^a \perp\!\!\!\perp A$ for both $a = 0$ and $a = 1$. We will assume that a causal DAG represents an FFRCISTG as detailed as the data whenever we do not mention the underlying model.

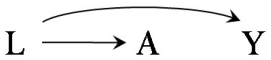

Figure 6.3

A path between two variables R and S in a DAG is a route that connects R and S by following a sequence of edges such that the route visits no variable more than once. A path is causal if it consists entirely of edges with their arrows pointing in the same direction. Otherwise it is noncausal.

$1] \neq \Pr[Y^{a=0} = 1]$. The causal diagram in Figure 6.2 is the graphical translation of this knowledge for an experiment in which aspirin A is randomly, and unconditionally, assigned. Second, suppose you know that carrying a lighter A has no causal effect (causative or preventive) on anyone's risk of lung cancer Y, i.e., $\Pr[Y^{a=1} = 1] = \Pr[Y^{a=0} = 1]$, and that cigarette smoking L has a causal effect on both carrying a lighter A and lung cancer Y. The causal diagram in Figure 6.3 is the graphical translation of this knowledge. The lack of an arrow between A and Y indicates that carrying a lighter does not have a causal effect on lung cancer; L is depicted as a common cause of A and Y.

To draw Figures 6.2 and 6.3 we only used your knowledge about the causal relations among the variables in the diagram but, interestingly, these causal diagrams also encode information about the expected associations (or, more exactly, the lack of them) among the variables in the diagram. We now argue heuristically that, in general, the variables A and Y will be associated in both Figure 6.2 and 6.3, and describe key related results from causal graphs theory.

Take first the randomized experiment represented in Figure 6.2. Intuitively one would expect that two variables A and Y linked only by a causal arrow would be associated. And that is exactly what causal graphs theory shows: when one knows that A has a causal effect on Y, as in Figure 6.2, then one should also generally expect A and Y to be associated. This is of course consistent with the fact that, in an ideal randomized experiment with unconditional exchangeability, causation $\Pr[Y^{a=1} = 1] \neq \Pr[Y^{a=0} = 1]$ implies association $\Pr[Y = 1|A = 1] \neq \Pr[Y = 1|A = 0]$, and vice versa. A heuristic that captures the causation-association correspondence in causal diagrams is the visualization of the paths between two variables as pipes or wires through which association flows. Association, unlike causation, is a symmetric relationship between two variables; thus, when present, association flows between two variables regardless of the direction of the causal arrows. In Figure 6.2 one could equivalently say that the association flows from A to Y or from Y to A.

Now let us consider the observational study represented in Figure 6.3. We know that carrying a lighter A has no causal effect on lung cancer Y. The question now is whether carrying a lighter A is associated with lung cancer Y. That is, we know that $\Pr[Y^{a=1} = 1] = \Pr[Y^{a=0} = 1]$ but is it also true that $\Pr[Y = 1|A = 1] = \Pr[Y = 1|A = 0]$? To answer this question, imagine that a naive investigator decides to study the effect of carrying a lighter A on the risk of lung cancer Y (we do know that there is no effect but this is unknown to the investigator). He asks a large number of people whether they are carrying lighters and then records whether they are diagnosed with lung cancer during the next 5 years. Hera is one of the study participants. We learn that Hera is carrying a lighter. But if Hera is carrying a lighter ($A = 1$), then it is more likely that she is a smoker ($L = 1$), and therefore she has a greater than average risk of developing lung cancer ($Y = 1$). We then intuitively conclude that A and Y are expected to be associated because the cancer risk in those carrying a lighter ($A = 1$) is different from the cancer risk in those not carrying a lighter ($A = 0$), or $\Pr[Y = 1|A = 1] \neq \Pr[Y = 1|A = 0]$. In other words, having information about the treatment A improves our ability to predict the outcome Y, even though A does not have a causal effect on Y. The investigator will make a mistake if he concludes that A has a causal effect on Y just because A and Y are associated. Causal graphs theory again confirms our intuition. In graphic terms, A and Y are associated because there is a flow of association from A to Y (or, equivalently, from Y to A) through the common cause L.

Let us now consider a third example. Suppose you know that certain genetic haplotype A has no causal effect on anyone's risk of becoming a cigarette smoker Y, i.e., $\Pr[Y^{a=1} = 1] = \Pr[Y^{a=0} = 1]$, and that both the haplotype A and cigarette smoking Y have a causal effect on the risk of heart disease L. The causal diagram in Figure 6.4 is the graphical translation of this knowledge. The lack of an arrow between A and Y indicates that the haplotype does not have a causal effect on cigarette smoking, and L is depicted as a common effect of A and Y. The common effect L is referred to as a *collider* on the path $A \rightarrow L \leftarrow Y$ because two arrowheads collide on this node.

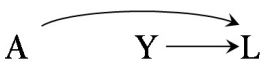

Figure 6.4

Again the question is whether A and Y are associated. To answer this question, imagine that another investigator decides to study the effect of haplotype A on the risk of becoming a cigarette smoker Y (we do know that there is no effect but this is unknown to the investigator). She makes genetic determinations on a large number of children, and then records whether they end up becoming smokers. Apollo is one of the study participants. We learn that Apollo does not have the haplotype ($A = 0$). Is he more or less likely to become a cigarette smoker ($Y = 1$) than the average person? Learning about the haplotype A does not improve our ability to predict the outcome Y because the risk in those with ($A = 1$) and without ($A = 0$) the haplotype is the same, or $\Pr[Y = 1|A = 1] = \Pr[Y = 1|A = 0]$. In other words, we would intuitively conclude that A and Y are not associated, i.e., A and Y are independent or $A \perp\!\!\!\perp Y$. The knowledge that both A and Y cause heart disease L is irrelevant when considering the association between A and Y. Causal graphs theory again confirms our intuition because it says that colliders, unlike other variables, block the flow of association along the path on which they lie. Thus A and Y are independent because the only path between them, $A \rightarrow L \leftarrow Y$, is blocked by the collider L.

In summary, two variables are (marginally) associated if one causes the other, or if they share common causes. Otherwise they will be (marginally) independent. The next section explores the conditions under which two variables A and Y may be independent conditionally on a third variable L.

6.3 Causal diagrams and conditional independence

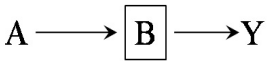

Figure 6.5

Because no conditional independences are expected in complete causal diagrams (those in which all possible arrows are present), it is often said that information about associations is in the missing arrows.

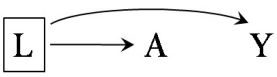

Figure 6.6

Blocking the flow of association between treatment and outcome through the common cause is the graph-based justification to use stratification as a method to achieve exchangeability.

We now revisit the settings depicted in Figures 6.2, 6.3, and 6.4 to discuss the concept of conditional independence in causal diagrams.

According to Figure 6.2, we expect aspirin A and heart disease Y to be associated because aspirin has a causal effect on heart disease. Now suppose we obtain an additional piece of information: aspirin A affects the risk of heart disease Y because it reduces platelet aggregation B. This new knowledge is translated into the causal diagram of Figure 6.5 that shows platelet aggregation B (1: high, 0: low) as a mediator of the effect of A on Y.

Once a third variable is introduced in the causal diagram we can ask a new question: is there an association between A and Y within levels of (conditional on) B? Or, equivalently: when we already have information on B, does information about A improve our ability to predict Y? To answer this question, suppose data were collected on A, B, and Y in a large number of individuals, and that we restrict the analysis to the subset of individuals with low platelet aggregation ($B = 0$). The square box placed around the node B in Figure 6.5 represents this restriction. (We would also draw a box around B if the analysis were restricted to the subset of individuals with $B = 1$.)

Individuals with low platelet aggregation ($B = 0$) have a lower than average risk of heart disease. Now take one of these individuals. Regardless of whether the individual was treated ($A = 1$) or untreated ($A = 0$), we already knew that he has a lower than average risk because of his low platelet aggregation. In fact, because aspirin use affects heart disease risk *only* through platelet aggregation, learning an individual's treatment status does not contribute any additional information to predict his risk of heart disease. Thus, in the subset of individuals with $B = 0$, treatment A and outcome Y are not associated. (The same informal argument can be made for individuals in the group with $B = 1$.) Even though A and Y are marginally associated, A and Y are *conditionally independent* (unassociated) given B because the risk of heart disease is the same in the treated and the untreated within levels of B: $\Pr[Y = 1|A = 1, B = b] = \Pr[Y = 1|A = 0, B = b]$ for all b. That is, $A \perp\!\!\!\perp Y|B$. Graphically, we say that a box placed around variable B blocks the flow of association through the path $A \rightarrow B \rightarrow Y$.

Let us now return to Figure 6.3. We concluded in the previous section that carrying a lighter A was associated with the risk of lung cancer Y because the path $A \leftarrow L \rightarrow Y$ was open to the flow of association from A to Y. The question we ask now is whether A is associated with Y conditional on L. This new question is represented by the box around L in Figure 6.6. Suppose the investigator restricts the study to nonsmokers ($L = 0$). In that case, learning that an individual carries a lighter ($A = 1$) does not help predict his risk of lung cancer ($Y = 1$) because the entire argument for better prediction relied on the fact that people carrying lighters are more likely to be smokers. This argument is irrelevant when the study is restricted to nonsmokers or, more generally, to people who smoke with a particular intensity. Even though A and Y are marginally associated, A and Y are conditionally independent given L because the risk of lung cancer is the same in the treated and the untreated within levels of L: $\Pr[Y = 1|A = 1, L = l] = \Pr[Y = 1|A = 0, L = l]$ for all l. That is, $A \perp\!\!\!\perp Y|L$. Graphically, we say that the flow of association between A and Y is interrupted because the path $A \leftarrow L \rightarrow Y$ is blocked by the box around L.

Finally, consider Figure 6.4 again. We concluded in the previous section that having the haplotype A was independent of being a cigarette smoker

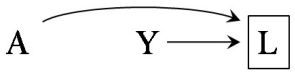

Figure 6.7

Y because the path between A and Y, $A \rightarrow L \leftarrow Y$, was blocked by the collider L. We now argue heuristically that, in general, A and Y will be conditionally associated within levels of their common effect L. Suppose that the investigators, who are interested in estimating the effect of haplotype A on smoking status Y, restricted the study population to individuals with heart disease ($L = 1$). The square around L in Figure 6.7 indicates that they are conditioning on a particular value of L. Knowing that an individual with heart disease lacks haplotype A provides some information about her smoking status because, in the absence of A, it is more likely that another cause of L such as Y is present. That is, among people with heart disease, the proportion of smokers is increased among those without the haplotype A. Therefore, A and Y are inversely associated conditionally on $L = 1$. The investigator will make a mistake if he concludes that A has a causal effect on Y just because A and Y are associated within levels of L. In the extreme, if A and Y were the only causes of L, then among people with heart disease the absence of one of them would perfectly predict the presence of the other. Causal graphs theory shows that indeed conditioning on a collider like L opens the path $A \rightarrow L \leftarrow Y$, which was blocked when the collider was not conditioned on. Intuitively, whether two variables (the causes) are associated cannot be influenced by an event in the future (their effect), but two causes of a given effect generally become associated once we stratify on the common effect.

See Chapter 8 for more on associations due to conditioning on common effects.

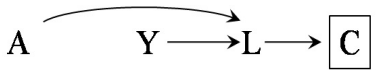

Figure 6.8

As another example, the causal diagram in Figure 6.8 adds to that in Figure 6.7 a diuretic medication C whose use is a consequence of a diagnosis of heart disease. A and Y are also associated within levels of C because C is a common effect of A and Y. Causal graphs theory shows that conditioning on a variable C affected by a collider L also opens the path $A \rightarrow L \leftarrow Y$. This path is blocked in the absence of conditioning on either the collider L or its consequence C.

This and the previous section review three structural reasons why two variables may be associated: one causes the other, they share common causes, or they share a common effect and the analysis is restricted to certain level of that common effect (or of its descendants). Along the way we introduced a number of graphical rules that can be applied to any causal diagram to determine whether two variables are (conditionally) independent. The arguments we used to support these graphical rules were heuristic and relied on our causal intuitions. These arguments, however, have been formalized and mathematically proven. See Fine Point 6.1 for a systematic summary of the graphical rules, and Fine Point 6.2 for an introduction to the concept of faithfulness.

The mathematical theory underlying the graphical rules is known as "d-separation" (Pearl 1995).

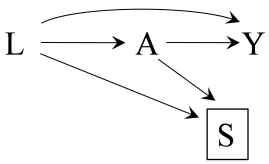

Figure 6.9

There is another possible source of association between two variables that we have not discussed yet: chance or random variability. Unlike the structural reasons for an association between two variables—causal effect of one on the other, shared common causes, conditioning on common effects—random variability results in chance associations that become smaller when the size of the study population increases.

To focus our discussion on structural associations rather than chance associations, we continue to assume until Chapter 10 that we have recorded data on every individual in a very large (perhaps hypothetical) population of interest.

6.4 Positivity and consistency in causal diagrams

Because causal diagrams encode our qualitative expert knowledge about the causal structure, they can be used as a visual aid to help conceptualize causal

Fine Point 6.1

D-separation. We define a path to be either blocked or open according to the following graphical rules.

1. If there are no variables being conditioned on, a path is blocked if and only if two arrowheads on the path collide at some variable on the path. In Figure 6.1, the path $L \rightarrow A \rightarrow Y$ is open, whereas the path $A \rightarrow Y \leftarrow L$ is blocked because two arrowheads on the path collide at Y. We call Y a collider on the path $A \rightarrow Y \leftarrow L$.

2. Any path that contains a non-collider that has been conditioned on is blocked. In Figure 6.5, the path between A and Y is blocked after conditioning on B. We use a square box around a variable to indicate that we are conditioning on it.

3. A collider that has been conditioned on does not block a path. In Figure 6.7, the path between A and Y is open after conditioning on L.

4. A collider that has a descendant that has been conditioned on does not block a path. In Figure 6.8, the path between A and Y is open after conditioning on C, a descendant of the collider L.

Rules 1–4 can be summarized as follows. A path is blocked if and only if it contains a non-collider that has been conditioned on, or it contains a collider that has not been conditioned on and has no descendants that have been conditioned on. Two variables are d-separated if all paths between them are blocked (otherwise they are d-connected). Two sets of variables are d-separated if each variable in the first set is d-separated from every variable in the second set. Thus, A and L are not d-separated in Figure 6.1 because there is one open path between them ($L \rightarrow A$), despite the other path ($A \rightarrow Y \leftarrow L$)'s being blocked by the collider Y. In Figure 6.4, however, A and Y are d-separated because the only path between them is blocked by the collider L.

The relationship between statistical independence and the purely graphical concept of d-separation relies on the causal Markov assumption (Technical Point 6.1): In a causal DAG, any variable is independent of its non-descendants conditional on its parents. Pearl (1988) proved the following fundamental theorem: The causal Markov assumption implies that, given any three disjoint sets A, B, C of variables, if A is d-separated from B conditional on C, then A is statistically independent of B given C. The assumption that the converse holds, i.e., that A is d-separated from B conditional on C if A is statistically independent of B given C, is a separate assumption—the faithfulness assumption described in Fine Point 6.2. Under faithfulness, A is conditionally independent of Y given B in Figure 6.5, A is not conditionally independent of Y given L in Figure 6.7, and A is not conditionally independent of Y given C in Figure 6.8. The d-separation rules ('d-' stands for directional) to infer associational statements from causal diagrams were formalized by Pearl (1995). An equivalent set of graphical rules, known as "moralization", was developed by Lauritzen et al. (1990).

Pearl (2009) reviews quantitative methods for causal inference that are derived from graph theory.

problems and guide data analyses. In fact, the formulas that we described in Chapter 2 to quantify treatment effects—standardization and IP weighting—can also be derived using causal graphs theory, as part of what is sometimes referred to as the do-calculus. Therefore, our choice of counterfactual theory in Chapters 1-5 did not really privilege one particular approach but only one particular notation.

Regardless of the notation used (counterfactuals or graphs), exchangeability, positivity, and consistency are conditions required for causal inference via standardization or IP weighting. If any of these conditions does not hold, the numbers arising from the data analysis may not be appropriately interpreted as measures of causal effect. In the next section (and in Chapters 7 and 8) we discuss how the exchangeability condition is translated into graph language. Here we focus on positivity and consistency.

Unfortunately, causal graphs cannot encode violations of positivity except in special cases, e.g, when positivity is violated because a treatment A hap-

Fine Point 6.2

Faithfulness. In a causal DAG the absence of an arrow from A to Y indicates that the sharp null hypothesis of no causal effect of A on any individual's Y holds, and an arrow $A \rightarrow Y$ (as in Figure 6.2) indicates that A has a causal effect on the outcome Y of at least one individual in the population. Thus, we would generally expect that, under Figure 6.2, the average causal effect of A on Y, $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$, and the association between A and Y, $\Pr[Y = 1|A = 1] - \Pr[Y = 1|A = 0]$, are not null. However, that is not necessarily true: a setting represented by Figure 6.2 may be one in which there is neither an average causal effect nor an association. For an example, remember the data in Table 4.1. Heart transplant A increases the risk of death Y in women (half of the population) and decreases the risk of death in men (the other half). Because the beneficial and harmful effects of A perfectly cancel out, the average causal effect is null, $\Pr[Y^{a=1} = 1] = \Pr[Y^{a=0} = 1]$. Yet Figure 6.2 is the correct causal diagram because treatment A affects the outcome Y of some individuals—in fact, of all individuals—in the population.

Formally, faithfulness is the assumption that, for three disjoint sets A, B, C on a causal DAG (where C may be the empty set), A independent of B given C implies A is d-separated from B given C. When, as in our example, the causal diagram makes us expect a non-null association that does not actually exist in the data, we say that the joint distribution of the data is not faithful to the causal DAG. In our example the unfaithfulness was the result of effect modification (by sex) with opposite effects of exactly equal magnitude in each half of the population. Such perfect cancellation of effects is rare, and thus we will assume faithfulness throughout this book. Because unfaithful distributions are rare, in practice lack of d-separation (See Fine Point 6.1) can be almost always equated to non-zero association.

There are, however, instances in which faithfulness is violated by design. For example, consider the prospective study in Section 4.5. The average causal effect of A on Y was computed after matching on L. In the matched population, L and A are not associated because the distribution of L is the same in the treated and the untreated. That is, individuals are selected into the matched population because they have a particular combination of values of L and A. The causal diagram in Figure 6.9 represents the setting of a matched study in which selection S (1: yes, 0: no) is determined by both A and L. The box around S indicates that the analysis is restricted to those selected into the matched cohort ($S = 1$). According to d-separation rules, there are two open paths between A and L when conditioning on S: $L \rightarrow A$ and $L \rightarrow S \leftarrow A$. Thus one would expect L and A to be associated conditionally on S. However, matching ensures that L and A are not associated (see Chapter 4). Why the discrepancy? Matching creates an association via the path $L \rightarrow S \leftarrow A$ that is of equal magnitude, but opposite direction, as the association via the path $L \rightarrow A$. The net result is a perfect cancellation of the associations. Matching leads to unfaithfulness.

Finally, faithfulness may be violated when there exist deterministic relations between variables on the graph. Specifically, when two variables are linked by paths that include deterministic arrows, then the two variables are independent if all paths between them are blocked, but might also be independent even if some paths are open. In this book we will assume faithfulness unless we say otherwise. Faithfulness is also assumed when the goal of the data analysis is discovering the causal structure (see Fine Point 6.3)

pens to be a deterministic function of a pretreatment variable L. In this case, we bold the $L \rightarrow A$ arrow to indicate determinism. The first component of consistency—well-defined interventions—means that the arrow from treatment A to outcome Y corresponds to a possibly hypothetical but relatively unambiguous intervention. In the causal diagrams discussed in this book, positivity is implicit unless otherwise specified, and consistency is embedded in the notation because we only consider treatment nodes with relatively well-defined interventions. Note that positivity is concerned with arrows into the treatment nodes, and well-defined interventions are only concerned with arrows leaving the treatment nodes.

Thus, the treatment nodes are implicitly given a different status compared with all other nodes. Some authors make this difference explicit by including *decision nodes* in causal diagrams. Though this decision-theoretic approach largely leads to the same methods described here, we do not include decision

nodes in the causal diagrams presented in this chapter. Because we are always explicit about the potential interventions on the variable A, the additional nodes (to represent the potential interventions) would be somewhat redundant. However, we will give a different status to treatment nodes when using SWIGs—causal diagrams with nodes representing counterfactual variables—in subsequent chapters.

Influence diagrams are causal diagrams augmented with decision nodes to represent the interventions of interest (Dawid 2000, 2002).

The different status of treatment nodes compared with other nodes was also graphically explicit in the causal trees introduced in Chapter 2, in which non-treatment branches corresponding to non-treatment variables L and Y were enclosed in circles, and in the "pies" representing sufficient causes in Chapter 5, which distinguish between potential treatments A and E and background factors U. Also, our discussion on sufficiently well-defined interventions of treatment in Chapter 3 emphasizes the requirements imposed on the treatment variables A that do not apply to other variables.

In contrast, the causal diagrams in this chapter apparently assign the same status to all variables in the diagram—this is indeed the case when causal diagrams are considered as representations of nonparametric structural equations models with independent errors (see Technical Point 6.2). The apparently equal status of all variables in causal diagrams may be misleading because some of those variables correspond to ill-defined interventions. It may be okay to draw a causal diagram that includes a node for "obesity" as the outcome Y or even as a covariate L (more about this on Section 9.5). However, for the reasons discussed in Chapter 3, it is generally not okay to draw a causal diagram that includes a node for "obesity" as a treatment A. In causal diagrams, nodes for treatment variables need to correspond to sufficiently well-defined interventions.

Recently, Pearl (2018, 2019) has suggested a concept of causation based on variables that "listen to others," which continues to assume that for every variable there are well-defined counterfactuals.

For example, suppose that we are interested in the potential causal effect of "weight loss" A on mortality Y, as discussed in Chapter 3. The causal diagram in Figure 6.10 includes nodes for A and Y as well as nodes for factors that affect body weight. For simplicity, the causal diagram includes only 3 of those factors: caloric intake Z which (let us assume) can only affect mortality through weight loss, exercise L which can affect mortality through pathways other than weight loss, and genetic traits U which can affect mortality through other pathways that are also independent of weight loss.

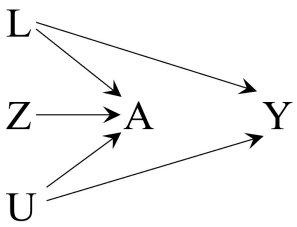

Figure 6.10

Identifying and interpreting the effect of a treatment A on an outcome Y requires knowledge about *how* to intervene on A. When there are several potential ways to intervene on A and some of those potential interventions have direct effects on the outcome Y as in Figure 6.10, it becomes unclear what "the effect of A on Y" means. In our example, reducing weight via caloric restriction Z would result in a different risk of mortality than reducing weight via increased exercise L or via genetic manipulation U. Even if one were willing to disregard the ill-defined causal effect, identifying the variables needed to achieve exchangeability would be a formidable challenge, as discussed in Chapter 3.

Being explicit about the interventions of interest is an important step towards having a well-defined causal effect, identifying relevant data, and choosing adjustment variables.

Fine Point 6.3

Discovery of causal structure. In this book we use causal diagrams as a way to represent our expert knowledge—or assumptions—about the causal structure of the problem at hand. That is, the causal diagram guides the data analysis. How about going in the opposite direction? Can we learn the causal structure by conducting data analyses without making assumptions about the causal structure? The process of learning components of the causal structure through data analysis is referred to as discovery. See the books by Spirtes et al. (2000) and by Peters et al. (2017) for descriptions of approaches to causal discovery.

We now briefly discuss causal discovery under the assumption that the observed data arose from an unknown causal DAG that includes, in addition to the observed variables, an unknown number of unobserved variables U. Causal discovery is sometimes possible if we assume faithfulness, so that statistical independencies in the observed data distribution imply missing causal arrows on the DAG. Even assuming faithfulness, discovery is often impossible. For example, suppose that we find a strong association between two variables B and C in our data. We cannot learn the causal structure involving B and C because their association is consistent with many causal diagrams: B causes C ($B \to C$), C causes B, ($C \to B$), B and C share an unmeasured cause U ($B \longleftarrow U \to C$), B and C have an unobserved common effect U that has been conditioned on, and various combinations. If we knew the time sequence of B and C, we could only rule out causal diagrams with either $B \to C$ (if C predates B) or $C \to B$ (if B predates C).

There are, however, some settings in which learning causal structure from data appears possible. For example, consider 3 variables Z, A, Y and we know that their time sequence is Z first, A second, and Y last. In the absence of prior knowledge and empirical data, Figure 6.11 represents a potential causal DAG as we cannot rule out that Z has a direct effect on A and Y, that A has an effect on Y, or that there exists an unmeasured common cause between any pair of variables. We also cannot rule out any subgraph with one or more arrows removed.

Now suppose, hypothetically, that data on variables Z, A, and Y become available for an infinite number of individuals. Our data analysis finds that all 3 variables are marginally associated with each other, and that the only conditional independence that holds is $Z \perp\!\!\!\perp Y|A$. Then, if we are willing to assume that faithfulness holds, the only possible causal DAG consistent with our analysis is $Z \to A \to Y$ with perhaps a common cause U_1 of Z and A in addition to (or in place of) the arrow from Z to A. This is because, if either Z was a parent of Y or shared a cause U_2 with Y, or an unmeasured common cause U_3 of A and Y was present, then Z and Y could not have been statistically independent given A (assuming faithfulness). Thus, to explain the marginal dependency of Y and A, there must be a causal arrow from A to Y. (Note that, if an unmeasured common cause of A and Y existed, no conditional independence would be found and then, even assuming faithfulness, we could not determine whether A causes Y.)

In summary, the causal DAG learned implies that Z is not a direct cause (parent) of Y, that no unmeasured common cause of A and Y exists, and that, in fact, the average causal effect of A on Y is identified by $\mathrm{E}[Y|A=1] - \mathrm{E}[Y|A=0]$ because exchangeability holds. Of course, we do not have an infinite sample size. We postpone a discussion about the implications of having a finite sample for causal discovery until Technical Point 10.7.

6.5 A structural classification of bias

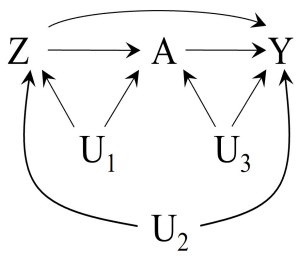

Figure 6.11

The word "bias" is frequently used by investigators making causal inferences. There are several related, but technically different, uses of the term "bias" (see Chapter 10). We say that there is *systematic bias* when the data are insufficient to identify—compute—the causal effect even with an infinite sample size. (In this chapter, due to the assumption of an infinite sample size, bias refers to systematic bias.) Informally, we often refer to systematic bias as any structural association between treatment and outcome that does not arise from the causal effect of treatment on outcome in the population of interest. Because causal diagrams are helpful to represent different sources of association, we can use causal diagrams to classify systematic bias according to its source, and thus to sharpen discussions about bias.

Take the crucial source of bias that we have discussed in previous chapters:

lack of exchangeability between the treated and the untreated. For the average causal effect in the entire population, we say that there is (unconditional) bias when $\Pr[Y^{a=1}=1] - \Pr[Y^{a=0}=1] \neq \Pr[Y=1|A=1] - \Pr[Y=1|A=0]$, which is the case when (unconditional) exchangeability $Y^a \perp\!\!\!\perp A$ does not hold. Absence of (unconditional) bias implies that the association measure (e.g., associational risk ratio or difference) in the population is a consistent estimate of the corresponding effect measure (e.g., causal risk ratio or difference) in the population.

When there is systematic bias, no estimator can be consistent. Review Chapter 1 for a definition of consistent estimator.

Lack of exchangeability results in bias even when the null hypothesis of no causal effect of treatment on the outcome holds. That is, even if the treatment had no causal effect on the outcome, treatment and outcome would be associated in the data. We then say that lack of exchangeability leads to *bias under the null*. In the observational study summarized in Table 3.1, there was bias under the null because the causal risk ratio was 1 whereas the associational risk ratio was 1.26. Any causal structure that results in bias under the null will also cause bias under the alternative (i.e., when treatment does have a non-null effect on the outcome). However, the converse is not true.

For example, conditioning on some variables may cause *selection bias under the alternative* (i.e., off the null) but not under the null, as described by Greenland (1977) and Hernán (2017). See also Chapter 18.

For the average causal effects within levels of L, we say that there is conditional bias whenever $\Pr[Y^{a=1}=1|L=l] - \Pr[Y^{a=0}=1|L=l]$ differs from $\Pr[Y=1|L=l, A=1] - \Pr[Y=1|L=l, A=0]$ for at least one stratum l, which is generally the case when conditional exchangeability $Y^a \perp\!\!\!\perp A|L=l$ does not hold for all a and l.

So far in this book we have referred to lack of exchangeability multiple times. However, we have yet to explore the causal structures that generate lack of exchangeability. With causal diagrams added to our methodological arsenal, we will be able to describe how lack of exchangeability can result from two different causal structures:

1. Common causes: When the treatment and outcome share a common cause, the association measure generally differs from the effect measure. Many epidemiologists use the term *confounding* to refer to this bias.

2. Conditioning on common effects: This structure is the source of bias that many epidemiologists refer to as *selection bias under the null*.

Chapter 7 will focus on confounding bias due to the presence of common causes, and Chapter 8 on selection bias due to conditioning on common effects. Again, both are examples of bias under the null due to lack of exchangeability.

Chapter 9 will focus on another source of bias: measurement error. So far we have assumed that all variables—treatment A , outcome Y, and covariates L— are perfectly measured. In practice, however, some degree of measurement error is expected. The bias due to measurement error is referred to as *measurement bias* or information bias. As we will see, some types of measurement bias also cause bias under the null.

Another form of bias may also result from (nonstructural) random variability. See Chapter 10.

Therefore, in the next three chapters we turn our attention to the three types of systematic bias—confounding, selection, and measurement. These biases may arise both in observational studies *and* in randomized experiments. The susceptibility to bias of randomized experiments may not be obvious from previous chapters, in which we conceptualized observational studies as some sort of imperfect randomized experiments, while only considering ideal randomized experiments with no participants lost during the follow-up, all participants adhering to their assigned treatment, and unknown treatment assignment for both study participants and investigators. While our quasi-mythological characterization of randomized experiments was helpful for teaching purposes, real

randomized experiments rarely look like that. The remaining chapters of Part I will elaborate on the sometimes fuzzy boundary between experimenting and observing.

Before that, we take a brief detour to describe causal diagrams in the presence of effect modification.

6.6 The structure of effect modification

Identifying potential sources of bias is a key use of causal diagrams: we can use our causal expert knowledge to draw graphs and then search for sources of association between treatment and outcome. Causal diagrams are less helpful to illustrate the concept of effect modification that we discussed in Chapter 4.

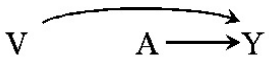

Figure 6.12

Suppose heart transplant A was randomly assigned in an experiment to identify the average causal effect of A on death Y. For simplicity, let us assume that there is no bias, and thus Figure 6.2 adequately represents this study. Computing the effect of A on the risk of Y presents no challenge. Because association is causation, the associational risk difference $\Pr[Y = 1|A = 1] - \Pr[Y = 1|A = 0]$ can be interpreted as the causal risk difference $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$. The investigators, however, want to go further because they suspect that the causal effect of heart transplant varies by the quality of medical care offered in each hospital participating in the study. Thus, the investigators classify all individuals as receiving high ($V = 1$) or normal ($V = 0$) quality of care, compute the stratified risk differences in each level of V as described in Chapter 4, and indeed confirm that there is effect modification by V on the additive scale. The causal diagram in Figure 6.12 includes the effect modifier V with an arrow into the outcome Y but no arrow into treatment A (which is randomly assigned and thus independent of V). Two important caveats.

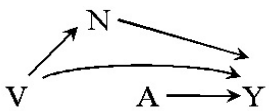

Figure 6.13

First, the causal diagram in Figure 6.12 would still be a valid causal diagram if it did not include V because V is not a common cause of A and Y. It is only because the causal question makes reference to V (i.e., what is the average causal effect of A on Y *within levels of* V?), that V needs to be included on the causal diagram. Other variables measured along the path between "quality of care" V and the outcome Y could also qualify as effect modifiers. For example, Figure 6.13 shows the effect modifier "therapy complications" N, which partly mediates the effect of V on Y.

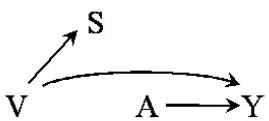

Figure 6.14

Second, the causal diagram in Figure 6.12 does not necessarily indicate the presence of effect modification by V. The causal diagram implies that both A and V affect death Y, but it does not distinguish among the following three qualitatively distinct ways that V could modify the effect of A on Y:

1. The causal effect of treatment A on mortality Y is in the same direction (i.e., harmful or beneficial) in both stratum $V = 1$ and stratum $V = 0$.

2. The direction of the causal effect of treatment A on mortality Y in stratum $V = 1$ is the opposite of that in stratum $V = 0$ (i.e., there is qualitative effect modification).

3. Treatment A has a causal effect on Y in one stratum of V but no causal effect in the other stratum A only kills individuals with $V = 0$.

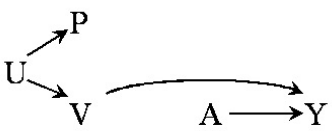

Figure 6.15

That is, valid causal graphs such as Figure 6.12 fail to distinguish between the above three different qualitative types of effect modification by V.

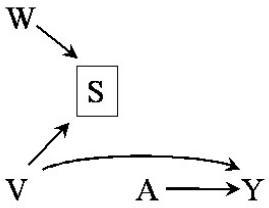

Figure 6.16

For a finer classification of effect modification via causal diagrams, see VanderWeele and Robins (2007b)

In the above example, the effect modifier V had a causal effect on the outcome. Many effect modifiers, however, do not have a causal effect on the outcome. Rather, they are surrogates for variables that have a causal effect on the outcome. Figure 6.14 includes the variable "cost of the treatment" S (1: high, 0: low), which is affected by "quality of care" V but has itself no effect on mortality Y. An analysis stratified by S (but not by V) will generally detect effect modification by S even though the variable that truly modifies the effect of A on Y is V. The variable S is a *surrogate effect modifier* whereas the variable V is a *causal effect modifier* (see Section 4.2). Because causal and surrogate effect modifiers are often indistinguishable in practice, the concept of effect modification comprises both. As discussed in Section 4.2, some prefer to use the neutral term "heterogeneity of causal effects," rather than "effect modification," to avoid confusion. For example, someone might be tempted to interpret the statement "cost modifies the effect of heart transplant on mortality because the effect is more beneficial when the cost is higher" as an argument to increase the price of medical care without necessarily increasing its quality.

A surrogate effect modifier is simply a variable associated with the causal effect modifier. Figure 6.14 depicts the setting in which such association is due to the effect of the causal effect modifier on the surrogate effect modifier. However, such association may also be due to shared common causes or conditioning on common effects. For example, Figure 6.15 includes the variables "place of residence" (1: Greece, 0: Rome) U and "passport-defined nationality" P (1: Greece, 0: Rome). Place of residence U is a common cause of both quality of care V and nationality P. Thus P will behave as a surrogate effect modifier because P is associated with the causal effect modifier V. Another (admittedly silly) example to illustrate this issue: Figure 6.16 includes the variables "cost of care" S and "use of bottled mineral water (rather than tap water) for drinking at the hospital" W. Use of mineral water W affects cost S but not mortality Y in developed countries. If the study were restricted to low-cost hospitals ($S = 0$), then use of mineral water W would be generally associated with medical care V, and thus W would behave as a surrogate effect modifier. In summary, surrogate effect modifiers can be associated with the causal effect modifier by structures including common causes, conditioning on common effects, or cause and effect.

Some intuition for the association between W and V in low-cost hospitals $S = 0$: suppose that low-cost hospitals that use mineral water need to offset the extra cost of mineral water by spending less on components of medical care that decrease mortality. Then use of mineral water would be inversely associated with quality of medical care in low-cost hospitals.

Causal diagrams are in principle agnostic about the presence of interaction between two treatments A and E. However, causal diagrams can encode information about interaction when augmented with nodes that represent sufficient-component causes (see Chapter 5), i.e., nodes with deterministic arrows from the treatments to the sufficient-component causes. Because the presence of interaction affects the magnitude and direction of the association due to conditioning on common effects, these augmented causal diagrams are discussed in Chapter 8.