Chapter 3
OBSERVATIONAL STUDIES

Consider again the causal question "does one's looking up at the sky make other pedestrians look up too?" After considering a randomized experiment as in the previous chapter, you concluded that looking up so many times was too time-consuming and unhealthy for your neck bones. Hence you decided to conduct the following study: Find a nearby pedestrian who is standing in a corner and not looking up. Then find a second pedestrian who is walking towards the first one and not looking up either. Observe and record their behavior during the next 10 seconds. Repeat this process a few thousand times. You could now compare the proportion of second pedestrians who looked up after the first pedestrian did, and compare it with the proportion of second pedestrians who looked up before the first pedestrian did. Such a scientific study in which the investigator observes and records the relevant data is referred to as an observational study.

If you had conducted the observational study described above, critics could argue that two pedestrians may both look up not because the first pedestrian's looking up causes the other's looking up, but because they both heard a thunderous noise above or some rain drops started to fall, and thus your study findings are inconclusive as to whether one's looking up makes others look up. These criticisms do not apply to randomized experiments, which is one of the reasons why randomized experiments are central to the theory of causal inference. However, in practice, the importance of randomized experiments for the estimation of causal effects is more limited. Many scientific studies are not experiments. Much human knowledge is derived from observational studies. Think of evolution, tectonic plates, global warming, or astrophysics. Think of how humans learned that hot coffee may cause burns. This chapter reviews some conditions under which observational studies lead to valid causal inferences.

3.1 Identifiability conditions

For simplicity, this chapter considers only randomized experiments in which all participants remain under follow-up and adhere to their assigned treatment throughout the entire study. Chapters 8 and 9 discuss alternative scenarios.

Ideal randomized experiments can be used to identify and quantify average causal effects because the randomized assignment of treatment leads to exchangeability. Take a marginally randomized experiment of heart transplant and mortality as an example: if those who received a transplant had not received it, they would have been expected to have the same death risk as those who did not actually receive the heart transplant. As a consequence, an associational risk ratio of 0.7 from the randomized experiment is expected to equal the causal risk ratio.

Observational studies, on the other hand, may be much less convincing (for an example, see the introduction to this chapter). A key reason for our hesitation to endow observational associations with a causal interpretation is the lack of randomized treatment assignment. As an example, take an observational study of heart transplant and mortality in which those who received the heart transplant were more likely to have a severe heart condition. Then, if those who received a transplant had not received it, they would have been expected to have a greater death risk than those who did not actually receive the heart transplant. As a consequence, an associational risk ratio of 1.1 from the observational study would be a compromise between the truly beneficial effect of transplant on mortality (which pushes the associational risk ratio to be under 1) and the underlying greater mortality risk in those who received transplant

(which pushes the associational risk ratio to be over 1). The best explanation for an association between treatment and outcome in an observational study is not necessarily a causal effect of the treatment on the outcome.

While recognizing that randomized experiments have intrinsic advantages for causal inference, sometimes we are stuck with observational studies to answer causal questions. What do we do? A common strategy is to analyze our data as if treatment had been randomly assigned conditional on measured covariates L—though we often know this is at best an approximation. Causal inference from observational data then revolves around the hope that the observational study can be viewed as a conditionally randomized experiment.

Informally, an observational study can be conceptualized as a conditionally randomized experiment if the following conditions hold:

1. the values of treatment under comparison correspond to well-defined interventions that, in turn, correspond to the versions of treatment in the data
2. the conditional probability of receiving every value of treatment, though not decided by the investigators, depends only on measured covariates L
3. the probability of receiving every value of treatment conditional on L is greater than zero, i.e., positive

In this chapter we describe these three conditions in the context of observational studies. Condition 1 was referred to as consistency in Chapter 1, condition 2 was referred to as exchangeability in the previous chapters, and condition 3 was referred to as positivity in Technical Point 2.3.

We will see that these conditions are often heroic, which explains why causal inferences from observational studies are viewed with suspicion. However, if the analogy between observational study and conditionally randomized experiment happens to be correct, then we can use the methods described in the previous chapter—IP weighting or standardization—to identify causal effects from observational studies. We therefore refer to these conditions as *identifiability* conditions or assumptions. For example, in the previous chapter, we computed a causal risk ratio equal to 1 using the data in Table 2.2, which arose from a conditionally randomized experiment. If the same data, now shown in Table 3.1, had arisen from an observational study and the three identifiability conditions above held true, we would also compute a causal risk ratio equal to 1.

Table 3.1

	L	A	Y
Rheia	0	0	0
Kronos	0	0	1
Demeter	0	0	0
Hades	0	0	0
Hestia	0	1	0
Poseidon	0	1	0
Hera	0	1	0
Zeus	0	1	1
Artemis	1	0	1
Apollo	1	0	1
Leto	1	0	0
Ares	1	1	1
Athena	1	1	1
Hephaestus	1	1	1
Aphrodite	1	1	1
Polyphemus	1	1	1
Persephone	1	1	1
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

Importantly, in ideal randomized experiments the identifiability conditions hold by design. That is, for a conditionally randomized experiment, we would only need the data in Table 3.1 to compute the causal risk ratio of 1. In contrast, to identify the causal risk ratio from an observational study, we would need to assume that the identifiability conditions held, which of course may not be true. Causal inference from observational data requires two elements: data and identifiability conditions. See Fine Point 3.1 for a more precise definition of identifiability.

When any of the identifiability conditions does not hold, the analogy between observational study and conditionally randomized experiment breaks down. In that situation, there are other possible approaches to causal inference from observational data, which require a different set of identifiability conditions. One of these approaches is hoping that a predictor of treatment, referred to as an *instrumental variable*, behaves as if it had been randomly assigned conditional on the measured covariates. We discuss instrumental variable methods in Chapter 16.

Rubin (1974, 1978) extended Neyman's theory for randomized experiments to observational studies. Rosenbaum and Rubin (1983) referred to the combination of exchangeability and positivity as *weak ignorability,* and to the combination of full exchangeability (see Technical Point 2.1) and positivity as *strong ignorability*.

Fine Point 3.1

Identifiability of causal effects. We say that an average causal effect is (nonparametrically) identifiable under a particular set of assumptions if these assumptions imply that the distribution of the observed data is compatible with a single value of the effect measure. Conversely, we say that an average causal effect is nonidentifiable under the assumptions when the distribution of the observed data is compatible with several values of the effect measure. For example, if the study in Table 3.1 had arisen from a conditionally randomized experiment in which the probability of receiving treatment depended on the value of L (and hence conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ holds by design) then we showed in the previous chapter that the causal effect is identifiable: the causal risk ratio equals 1, without requiring any further assumptions. However, if the data in Table 3.1 had arisen from an observational study, then the causal risk ratio equals 1 only if we supplement the data with the assumption of conditional exchangeability $Y^a \perp\!\!\!\perp A|L$. To identify the causal effect in observational studies, we need an assumption external to the data, an identifying assumption. In fact, if we decide not to supplement the data with the identifying assumption, then the data in Table 3.1 are consistent with a causal risk ratio

- lower than 1, if risk factors other than L are more frequent among the treated.
- greater than 1, if risk factors other than L are more frequent among the untreated.
- equal to 1, if all risk factors except L are equally distributed between the treated and the untreated or, equivalently, if $Y^a \perp\!\!\!\perp A|L$.

This chapter discusses the three identifiability conditions for nonparametric identification of average causal effects. In Chapter 16, we describe alternative identifiability conditions which suffice for nonparametric identification of average causal effects.

Not surprisingly, observational methods based on the analogy with a conditionally randomized experiment have been traditionally privileged in disciplines in which this analogy is often reasonable (e.g., epidemiology), whereas instrumental variable methods have been traditionally privileged in disciplines in which observational studies cannot often be conceptualized as conditionally randomized experiments given the measured covariates (e.g., economics). Until Chapter 16, we will focus on causal inference approaches that rely on the ability of the observational study to emulate a conditionally randomized experiment. We now describe in more detail each of the three identifiability conditions.

3.2 Exchangeability

An independent predictor of the outcome is a covariate associated with the outcome Y within levels of treatment. For dichotomous outcomes, independent predictors of the outcome are often referred to as *risk factors* for the outcome.

We have already said much about exchangeability $Y^a \perp\!\!\!\perp A$. In marginally (i.e., unconditionally) randomized experiments, the treated and the untreated are exchangeable because the treated, had they remained untreated, would have experienced the same average outcome as the untreated did, and vice versa. This is so because randomization ensures that the independent predictors of the outcome are equally distributed between the treated and the untreated groups.

For example, take the study summarized in Table 3.1. We said in the previous chapter that exchangeability clearly does not hold in this study because 69% treated versus 43% untreated individuals were in critical condition $L = 1$

at baseline. This imbalance in the distribution of an independent outcome predictor is not expected to occur in a marginally randomized experiment (actually, such imbalance might occur by chance but let us keep working under the illusion that our study is large enough to prevent chance findings).

On the other hand, an imbalance in the distribution of independent outcome predictors L between the treated and the untreated is expected by design in conditionally randomized experiments in which the probability of receiving treatment depends on L. The study in Table 3.1 is such a conditionally randomized experiment: the treated and the untreated are not exchangeable—because the treated had, on average, a worse prognosis at the start of the study—but the treated and the untreated are conditionally exchangeable within levels of the variable L. In the subset $L = 1$ (critical condition), the treated and the untreated are exchangeable because the treated, had they remained untreated, would have experienced the same average outcome as the untreated did, and vice versa. And similarly for the subset $L = 0$. An equivalent statement: conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ holds in conditionally randomized experiments because, within levels of L, all other outcome predictors are equally distributed between the treated and untreated groups.

Back to observational studies. When treatment is not randomly assigned by the investigators, the reasons for receiving treatment are likely to be associated with some outcome predictors. That is, like in a conditionally randomized experiment, the distribution of outcome predictors will generally vary between the treated and untreated groups in an observational study. For example, the data in Table 3.1 could have arisen from an observational study in which doctors tend to direct the scarce heart transplants to those who need them most, i.e., individuals in critical condition $L = 1$. In fact, if the only outcome predictor that is unequally distributed between the treated and the untreated is L, then one can refer to the study in Table 3.1 as either (i) an observational study in which the probability of treatment $A = 1$ is 0.75 among those with $L = 1$ and 0.50 among those with $L = 0$, or (ii) a (nonblinded) conditionally randomized experiment in which investigators randomly assigned treatment $A = 1$ with probability 0.75 to those with $L = 1$ and 0.50 to those with $L = 0$. Both characterizations of the study are logically equivalent. Under either characterization, conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ holds and standardization or IP weighting can be used to identify the causal effect.

In Chapter 7, we will refer to these type of outcome predictors as *confounders*.

Of course, the crucial question for the observational study is whether L is the only outcome predictor that is unequally distributed between the treated and the untreated. Sadly, the question must remain unanswered, so our investigators need to be willing to work under the *assumption* that conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ holds. Also, note that not all variables that are unequally distributed between treatment groups need to be included in L. For example, heart transplants are assigned to individuals with low probability of rejecting the transplant, i.e., a heart with certain human leukocyte antigen (HLA) genes will be assigned to an individual who happen to have compatible genes. Because HLA genes are not predictors of mortality, conditional on L and A, then treatment assignment is essentially random within levels of L and thus HLA needs not be considered in the analysis.

In the absence of randomization, there is no guarantee that conditional exchangeability holds. For example, suppose that, unknown to the investigators, doctors prefer to transplant hearts into nonsmokers. Consider two individuals with $L = 1$. One of them is a smoker ($U = 1$) and the other one is a nonsmoker ($U = 0$), the one with $U = 1$ has a lower probability of receiving treatment $A = 1$. When the distribution of smoking, an important outcome predictor,

Fine Point 3.2

Crossover randomized experiments. In Fine Point 2.1, we described crossover experiments in which an individual is observed during two or more periods—say $t = 0$ and $t = 1$—and the individual receives a different treatment value in each period. We showed that individual causal effects can be identified in crossover experiments when the following three strong conditions hold: i) no carryover effect of treatment: $Y_{it=1}^{a_0,a_1} = Y_{it=1}^{a_1}$, ii) the individual causal effect does not depend on time: $Y_{it}^{a_t=1} - Y_{it}^{a_t=0} = \alpha_i$ for $t = 0, 1$, and iii) the counterfactual outcome under no treatment does not depend on time: $Y_{it}^{a_t=0} = \beta_i$ for $t = 0, 1$. No randomization was required. We now turn our attention to crossover randomized experiments in which the order of treatment values that an individual receives is randomly assigned.

Randomized treatment assignment becomes important when, due to possible temporal effects, we do not assume iii) holds. For simplicity, assume that every individual is randomized to either ($A_{i1} = 1$, $A_{i0} = 0$) or ($A_{i1} = 0$, $A_{i0} = 1$) with probability 0.5. Let $Y_{i1}^{a_1=0} - Y_{i0}^{a_0=0} = r_i$. Then, under i) and ii) and consistency, if $A_{i0} = 0$ and $A_{i1} = 1$, then $Y_{i1} - Y_{i0} = \alpha_i + r_i$, and if $A_{i1} = 0$ and $A_{i0} = 1$, then $Y_{i0} - Y_{i1} = \alpha_i - r_i$. Because r_i is unknown we can no longer identify individual causal effects but, since A_{i1} and A_{i0} are randomized and therefore independent of r_i, the mean of $(Y_{i1} - Y_{i0})\, A_{i1} + (Y_{i0} - Y_{i1})\, A_{i0}$ estimates the average causal effect, i.e., $\mathrm{E}[\alpha_i]$. If we only assume i), then this mean estimates the average of the average treament effects at times 0 and 1, i.e., $(\mathrm{E}[\alpha_{i1}] + \mathrm{E}[\alpha_{i0}])/2$, where $\alpha_{it} = Y_{it}^{a_t=1} - Y_{it}^{a_t=0}$.

In conclusion, if assumption 1) of no carryover effect holds, then a crossover experiment can be used to estimate average causal effects. However, for the type of treatments and outcomes we study in this book, the assumption of no carryover effect is implausible.

We use U to denote unmeasured variables. Because unmeasured variables cannot be used for standardization or IP weighting, the causal effect cannot be identified when the measured variables L are insufficient to achieve conditional exchangeability.

To verify conditional exchangeability, one needs to confirm that $\Pr[Y^a = 1|A = a, L = l] = \Pr[Y^a = 1|A \neq a, L = l]$. But this is logically impossible because, for individuals who do not receive treatment a ($A \neq a$) the value of Y^a is unknown and so the right hand side cannot be empirically evaluated.

differs between the treated (with lower proportion of smokers $U = 1$) and the untreated (with higher proportion of smokers) in the stratum $L = 1$, conditional exchangeability given L does not hold. Importantly, collecting data on smoking would not prevent the possibility that other imbalanced outcome predictors, unknown to the investigators, remain unmeasured.

Thus exchangeability $Y^a \perp\!\!\!\perp A|L$ may not hold in observational studies. Specifically, conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ will not hold if there exist unmeasured independent predictors U of the outcome such that the probability of receiving treatment A depends on U within strata of L. Worse yet, even if conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ held, the investigators cannot empirically verify that is actually the case. How can they check that the distribution of smoking is equal in the treated and the untreated if they have not collected data on smoking? What about all the other unmeasured outcome predictors U that may also be differentially distributed between the treated and the untreated? When analyzing an observational study under conditional exchangeability, we must hope that our expert knowledge guides us correctly to collect enough data so that the assumption is at least approximately true.

Investigators can use their expert knowledge to enhance the plausibility of the conditional exchangeability assumption. They can measure many relevant variables L (e.g., determinants of the treatment that are also independent outcome predictors), rather than only one variable as in Table 3.1, and then assume that conditional exchangeability is approximately true within the strata defined by the combination of all those variables L. Unfortunately, no matter how many variables are included in L, there is no way to test that the assumption is correct, which makes causal inference from observational data a risky task. The validity of causal inferences requires that the investigators' expert knowledge is correct. This knowledge, encoded as the assumption of exchangeability conditional on the measured covariates, supplements the data in an attempt to identify the causal effect of interest.

3.3 Positivity

Some investigators plan to conduct an experiment to compute the average effect of heart transplant A on 5-year mortality Y. It goes without saying that the investigators will assign some individuals to receive treatment level $A = 1$ and others to receive treatment level $A = 0$. Consider the alternative: the investigators assign all individuals to either $A = 1$ or $A = 0$. That would be silly. With all the individuals receiving the same treatment level, computing the average causal effect would be impossible. Instead we must assign treatment so that, with near certainty, some individuals will be assigned to each of the treatment groups. In other words, we must ensure that there is a probability greater than zero—a positive probability—of being assigned to each of the treatment levels. This is the *positivity* condition.

The positivity condition is sometimes referred to as the *experimental treatment assumption*.

We did not emphasize positivity when describing experiments because positivity is taken for granted in those studies. In marginally randomized experiments, the probabilities $\Pr[A = 1]$ and $\Pr[A = 0]$ are both positive by design. In conditionally randomized experiments, the conditional probabilities $\Pr[A = 1|L = l]$ and $\Pr[A = 0|L = l]$ are also positive by design for all levels of the variable L that are eligible for the study. For example, if the data in Table 3.1 had arisen from a conditionally randomized experiment, the conditional probabilities of assignment to heart transplant would have been $\Pr[A = 1|L = 1] = 0.75$ for those in critical condition and $\Pr[A = 1|L = 0] = 0.50$ for the others. Positivity holds, conditional on L, because neither of these probabilities is 0 (nor 1, which would imply that the probability of no heart transplant $A = 0$ would be 0). Thus we say that there is positivity if $\Pr[A = a|L = l] > 0$ for all a involved in the causal contrast. Actually, this definition of positivity is incomplete because, if our study population were restricted to the group $L = 1$, then there would be no need to require positivity in the group $L = 0$. Positivity is only needed for the values l that are present in the population of interest.

Positivity: $\Pr[A = a|L = l] > 0$ for all values l with $\Pr[L = l] \neq 0$ in the population of interest.

In addition, positivity is only required for the variables L that are required for exchangeability. For example, in the conditionally randomized experiment of Table 3.1, we do not ask ourselves whether the probability of receiving treatment is greater than 0 in individuals with blue eyes because the variable "having blue eyes" is not necessary to achieve exchangeability between the treated and the untreated. (The variable "having blue eyes" is not an independent predictor of the outcome Y conditional on L and A, and was not even used to assign treatment.) That is, the standardized risk and the IP weighted risk are equal to the counterfactual risk after adjusting for L only; positivity does not apply to variables that, like "having blue eyes", do not need to be adjusted for.

In observational studies, neither positivity nor exchangeability are guaranteed. For example, positivity would not hold if doctors always transplant a heart to individuals in critical condition $L = 1$, i.e., if $\Pr[A = 0|L = 1] = 0$, as shown in Figure 3.1. A difference between the conditions of exchangeability and positivity is that positivity can sometimes be empirically verified (see Chapter 12). For example, if Table 3.1 corresponded to data from an observational study, we would conclude that positivity holds for L because there are people at all levels of treatment (i.e., $A = 0$ and $A = 1$) in every level of L (i.e., $L = 0$ and $L = 1$).Our discussion of standardization and IP weighting in the previous chapter was explicit about the exchangeability condition, but only implicitly assumed the positivity condition (explicitly in Technical Point 2.3). Our previous definitions of standardized risk and IP weighted risk are

actually only meaningful when positivity holds. To intuitively understand why the standardized and IP weighted risk are not well-defined when the positivity condition fails, consider Figure 3.1. If there were no untreated individuals ($A = 0$) with $L = 1$, the data would contain no information to simulate what would have happened had all treated individuals been untreated because there would be no untreated individuals with $L = 1$ that could be considered exchangeable with the treated individuals with $L = 1$. See Technical Point 3.1 for details.

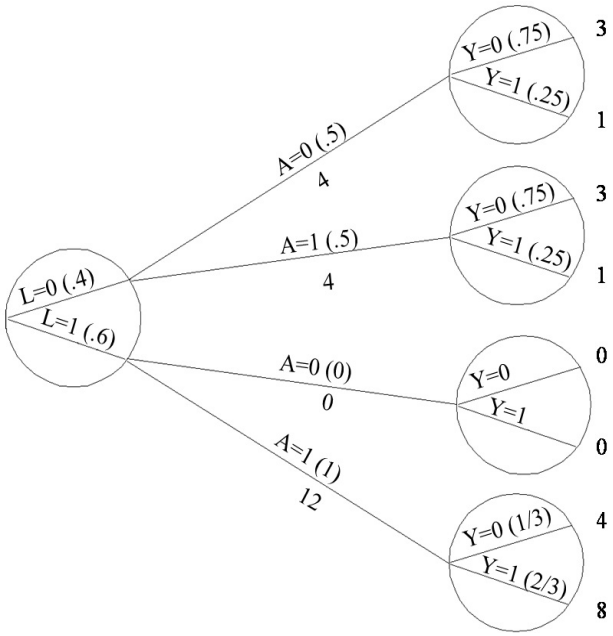

Figure 3.1

3.4 Consistency: First, define the counterfactual outcome

Robins and Greenland (2000) argued that well-defined counterfactuals, or mathematically equivalent concepts, are necessary for meaningful causal inference.

Consistency means that the observed outcome for every treated individual equals her outcome if she had received treatment, and that the observed outcome for every untreated individual equals her outcome if she had remained untreated, i.e., $Y^a = Y$ for every individual with $A = a$. This statement seems so obviously true that some readers may be wondering whether there are any situations in which consistency does not hold. After all, if I take aspirin $A = 1$ and I die ($Y = 1$), isn't it the case that my counterfactual outcome $Y^{a=1}$ under aspirin equals 1 by definition? The apparent simplicity of the consistency condition is deceptive. Let us unpack consistency by explicitly describing its two main components: (1) a precise definition of the counterfactual outcomes Y^a via a detailed specification of the superscript a, and (2) the linkage of the counterfactual outcomes to the observed outcomes. This section deals with the first component of consistency.

The methodology for causal inference described in this book is licensed by the existence of well-defined counterfactual outcomes Y^a. If Y^a is well-defined, then the causal effect $\Pr\left[Y^{a=1} = 1\right] - \Pr\left[Y^{a=0} = 1\right]$ is well-defined. A fundamental question is then, how do we know that the counterfactuals are well defined? The answer is straightforward when, as in randomized experiments with perfect adherence to the study protocol, the interventions of

Technical Point 3.1

Positivity for standardization and IP weighting. We have defined the standardized mean for treatment level a as $\sum_l \mathrm{E}[Y|A=a,L=l]\Pr[L=l]$. However, this expression can only be computed if the conditional quantity $\mathrm{E}[Y|A=a,L=l]$ is well defined, which will be the case when the conditional probability $\Pr[A=a|L=l]$ is greater than zero for all values l that occur in the population. That is, when positivity holds. (Note the statement $\Pr[A=a|L=l]>0$ for all l with $\Pr[L=l]\neq 0$ is effectively equivalent to $f[a|L]>0$ with probability 1.) Therefore, the standardized mean is defined as

$$\sum_l \mathrm{E}[Y|A=a,L=l]\Pr[L=l] \quad \text{if } \Pr[A=a|L=l]>0 \text{ for all } l \text{ with } \Pr[L=l]\neq 0,$$

and is undefined otherwise. The standardized mean can be computed only if, for each value of the covariate L in the population, there are some individuals that received the treatment level a.

The IP weighted mean $\mathrm{E}\left[\frac{I(A=a)Y}{f[A|L]}\right]$ is no longer equal to $\mathrm{E}\left[\frac{I(A=a)Y}{f[a|L]}\right]$ when positivity does not hold. Specifically, $\mathrm{E}\left[\frac{I(A=a)Y}{f[a|L]}\right]$ is undefined because the undefined ratio $\frac{0}{0}$ occurs in computing the expectation. On the other hand, the IP weighted mean $\mathrm{E}\left[\frac{I(A=a)Y}{f[A|L]}\right]$ is *always* well defined since its denominator $f[A|L]$ can never be zero. However, it is now a biased estimate of the counterfactual mean even under exchangeability when positivity fails to hold. In particular, $\mathrm{E}\left[\frac{I(A=a)Y}{f[A|L]}\right]$ is equal to $\Pr[L\in Q(a)]\sum_l \mathrm{E}[Y|A=a,L=l,L\in Q(a)]\Pr[L=l|L\in Q(a)]$ where $Q(a)=\{l;\Pr(A=a|L=l)>0\}$ is the set of values l for which $A=a$ may be observed with positive probability. Therefore, under exchangeability, $\mathrm{E}\left[\frac{I(A=a)Y}{f[A|L]}\right]$ equals $\mathrm{E}[Y^a|L\in Q(a)]\Pr[L\in Q(a)]$.

From the definition of $Q(a)$, $Q(0)$ cannot equal $Q(1)$ when A is binary and positivity does not hold. In this case the contrast $\mathrm{E}\left[\frac{I(A=1)Y}{f[A|L]}\right]-\mathrm{E}\left[\frac{I(A=0)Y}{f[A|L]}\right]$ has no causal interpretation, even under exchangeability, because it is a contrast between two different groups. Under positivity, $Q(1)=Q(0)$ and the contrast is the average causal effect if exchangeability holds.

Fine Point 1.2 introduced the concept of multiple versions of treatment.

interest are well defined themselves. Consider again a randomized experiment of heart transplant A and 5-year mortality Y. Before enrolling patients in the study, the investigators wrote a protocol in which the two interventions of interest—heart transplant $a=1$ and medical therapy $a=0$—were described in detail. For example, the investigators specified that individuals assigned to heart transplant were to receive certain pre-operative procedures, anesthesia, surgical technique, post-operative care, and immunosuppressive therapy in an attempt to ensure that each individual receives the same version of the treatment. Had the protocol not specified these details, it is possible that each doctor had conducted a different version of the treatment "heart transplant", perhaps using their preferred surgical technique or immunosuppressive therapy. We define $Y^{a=1}$ as the individual's outcome in this study if the instructions for intervention $a=1$ in the protocol of the experiment were followed, and analogously for $Y^{a=0}$.

But how do we know that the counterfactual outcomes are well defined in an observational study? Again, we will have well-defined counterfactuals if we have well-defined interventions. To do so, we need to specify what intervention is represented by a in Y^a. For example, for an observational study on the

Fine Point 3.3

On states and interventions. Variables like socioeconomic status or obesity define states, e.g., being poor, being obese. Causal questions about changes in states are of the form "what if these individuals become poor or obese?" These variables, however, are often not sufficiently well-defined for quantitative causal inference.

As an example, consider "the effect of becoming obese on myocardial infarction" (For simplicity, we use a traditional definition of obesity as body mass index ≥ 30.) The quoted text does not have a meaningful interpretation because, as discussed in the main text, the counterfactual outcome is ill-defined. One might think it would become well defined if we specified the start and end of the intervention (e.g., age 40 years through age 50 years) and the procedure by which body weight would be changed. However, if we specified all these details, we would not be describing the effect of obesity but the effect of whatever interventions we are specifying. Hernán and Taubman (2008) discuss the tribulations of two world leaders—a despotic king and a clueless president—when considering "the effect of obesity" in their countries.

In Technical Point 6.4, we provide mathematical conditions for well-defined counterfactual outcomes when the interventions are not well defined. However, those conditions are often extreme and of little practical relevance.

effect of heart transplant, we would specify the interventions of interest as for the randomized trial above, and for an observational study on the causal effect of exercise, we would specify duration, frequency, intensity, type (swimming, running...) and how the time devoted to exercise would otherwise be spent (rehearsing with your band, watching television...). The difficulty of specifying the treatment a increases for causal questions involving biological (e.g., blood pressure, LDL-cholesterol, body weight) or social (e.g., socioeconomic status) factors or states.

Take the effect of weight gain on fatal myocardial infarction. Suppose that, on Zeus's 35th birthday, we decided to make him gain 20 kilograms (kg) by age 40 by lowering his daily exercise. He had a fatal myocardial infarction at age 49. Now suppose that, in a parallel universe identical to ours, on Zeus's 35th birthday we decided to make him gain 20 kg by age 40 by increasing his caloric intake. He did not have a fatal myocardial infarction before age 50. That is, in both universes Zeus gains 20 kg, but only in one of them he had a fatal heart attack.

Because Zeus's counterfactual outcome under weight gain of 20 kg can be either death or no death, we conclude that specifying $a = 1$ as "weight gain of 20 kg" is too vague to define the counterfactual outcomes Y^a. The problem is that we can only change the value of weight gain by interventions (e.g., diet, exercise) that may have effects on the outcome through causal pathways that may not involve exclusively weight gain. See also Fine Point 3.3.

Whether causal effects are ill-defined depends on the outcome of interest. Consider the effect of obesity on job discrimination—as measured by the proportion of job applicants called for a personal interview after the employer reviews the applicant's resume and photograph. Here, the treatment is "obesity as perceived by the employer", rather than "biological obesity". Therefore, the biological mechanisms that led to obesity may be irrelevant.

In contrast, if we were interested in the causal effect of a weight loss pill A on mortality, this problem would not arise: "taking the pill" $a = 1$ is a well-defined intervention because it does not require considering any other interventions (e.g., on diet or exercise). Of course, the value of the counterfactual outcome Y^a can still depend on the individual's values of any previous interventions (e.g., on diet and exercise).

The more precisely we define the meaning of $a = 1$ and $a = 0$, the more precise our causal questions are. However, absolute precision in the definition of treatment is neither necessary nor possible. For example, for exercise, we do not need to specify the direction of running (clockwise or counterclockwise) around your neighborhood's park. Scientists agree that the direction of running is irrelevant because varying it would not lead to different outcomes. That is, we only need *sufficiently well-defined* interventions a for which no meaningful vagueness remains.

Which begs the question "How do we know that a treatment is sufficiently

Fine Point 3.4

Protocols open to interpretation. It is possible that $\Pr[Y^{a=1} = 1]$ differs between two randomized experiments with identical populations and protocols. To see this, consider the following scenario.

In both experiments, individuals assigned to $a = 1$ underwent a surgical operation according to the instructions in the protocol. However, the protocol did not specify how to match patients with surgeons. In the first experiment, individuals assigned to $a = 1$ were referred to and operated on by experienced surgeons if they were high risk patients, and by less experienced surgeons if they were low risk patients. Because of this, almost no patients died and $\Pr[Y^{a=1} = 1]$ was close to 0. In contrast, in the second experiment, individuals assigned to $a = 1$ were referred to a surgeon without regard to the patient's risk and the surgeon's experience. In this study $\Pr[Y^{a=1} = 1]$ is far from zero because many high-risk patients were operated on by inexperienced surgeons.

By definition, lack of exchangeability cannot explain the difference in $\Pr[Y^{a=1} = 1]$ because both experiments were randomized. Rather, the difference is explained by the different versions of treatment used in each trial. Because the protocol did not specify how to match patients with surgeons, the two trials ended up with different results.

well-defined?" Or, equivalently, how do we know that no meaningful vagueness remains? The answer is "We don't." Declaring a treatment sufficiently well-defined is a matter of agreement among experts based on the available substantive knowledge. Today we agree that the direction of running is irrelevant, but future research might prove us wrong if it is demonstrated that, say, leaning the body to the right, but not to the left, while running is harmful. At any point in history, experts who write the protocols of randomized experiments often attempt to eliminate as much vagueness as possible by employing the subject-matter knowledge at their disposal. However, some vagueness is inherent to all causal questions. The vagueness of causal questions can be reduced by a more detailed specification of treatment, but cannot be completely eliminated.

In pragmatic trials, the investigators may purposely choose not to specify all components of the intervention so that the treatment versions used in the trial reflect what happens in real world settings.

Even the protocols of randomized experiments may fail to specify some relevant components of the intervention. For example, the protocol of the above heart transplant study did not specify the surgeon's experience performing heart transplants. Thus, both experienced and inexperienced surgeons participated in the study. Because scant transplant experience is known to affect post-transplant mortality, the risk $\Pr[Y^{a=1} = 1]$ had all individuals received treatment according to the protocol will depend on the unknown distribution of experience of the participating surgeons. That is, the average causal effect in a new community with a different distribution of surgical experience will differ from the effect in the trial population, even if the new population follows the exact same protocol as in the trial.

In fact, the value of $\Pr[Y^{a=1} = 1]$, and therefore of the average causal effect, may differ between two experiments conducted in the same population and with the same protocol. This discrepancy would arise if the protocol allows for $a = 1$ to include several versions of treatment with different causal effects on the outcome of interest, and different versions of treatment are used in each experiment. Fine Point 3.4 describes an example of two randomized experiments with the same protocol but different causal effects. A different distribution of versions of treatment affects the transportability of causal effects (see Chapter 4). The same considerations apply to observational studies.

The discussion in this section illustrates an intrinsic feature of causal inference: the articulation of causal questions is contingent on domain expertise and informal judgment. What we view as a meaningful causal question at

Fine Point 3.5

Possible worlds. Philosophers of science have proposed counterfactual theories based on the concept of "possible worlds" (Stalnaker 1968, Lewis 1973). The counterfactual Y^a is defined to be the value of Y in the world in which the individual received the treatment that is closest to the actual world. In particular, these philosophers assume that $Y^a = Y$ if $A = a$ because the closest possible world to the actual world is itself. Hence, under their definition of counterfactuals, consistency always holds.

When $A \neq a$, the "closest possible world" and thus the counterfactual Y^a are always somewhat ill-defined and vague. Nonetheless, Lewis noted that his definition of counterfactuals is often useful. Robins and Greenland (2000) agreed but also argued that the concept of well-defined interventions should replace the concept of the closest possible world because, in observational studies, counterfactuals are vague and ill-defined to the degree that one fails to make precise the hypothetical interventions and causal contrasts under consideration.

The phrase "no causation without manipulation" (Holland 1986) captures the idea that meaningful causal inference requires sufficiently well-defined interventions. However, bear in mind that sufficiently well-defined interventions may not be humanly feasible, or practicable, interventions at a particular time in history. For example, the effect of genetic variants on disease was considered sufficiently well defined even before the existence of technology for genetic modification.

present may turn out to be viewed as too vague in the future after learning that unspecified components of the treatment affect the outcome and therefore the magnitude of the causal effect. Years from now, scientists will probably refine our obesity question in terms of cellular modifications which we barely understand at this time. Again, the term sufficiently well-defined treatment relies on expert consensus, which changes over time. Fine Point 3.5 links this discussion with previous proposals.

Refining the causal question, until it is agreed that no meaningful vagueness remains, is good practice for sound causal inference. For example, declaring our interest in "the effect of obesity" may be viewed as just a starting point for a discussion during which we will sharpen, and possibly modify, the causal question by refining the specification of the treatment until, hopefully, a consensus is reached with our colleagues. The more precisely we specify the treatment, the better defined the causal question is and the fewer opportunities for miscommunication between researchers and decision makers exist.

3.5 Consistency: Second, link counterfactuals to the observed data

For an expanded discussion of the issues described in Sections 3.4 and 3.5, see the text and references in Hernán (2016), and in Robins and Weissman (2016).

As a reminder, the consistency condition says that $Y^a = Y$ for individuals with $A = a$. In the previous section, we described the first component of consistency: sufficiently well-defined counterfactual outcomes Y^a such that no meaningful vagueness remains. In this section, we describe the second component of consistency in observational studies: ensuring that the equality $Y^a = Y$ holds, i.e., linking the counterfactual outcomes to the observed data.

Suppose our goal is quantifying the effect of heart transplant $a = 1$ vs. medical therapy $a = 0$ using observational data. We carefully specify the two treatment versions $a = 1$ and $a = 0$ of interest. Experts agree that $a = 1$ and $a = 0$ are sufficiently well-defined and, therefore, that no meaningful vagueness remains in the specification of the counterfactual outcomes $Y^{a=1}$ and $Y^{a=0}$. Specifically, we specified that heart transplant $a = 1$ includes certain preoperative procedures, anesthesia, surgical technique, post-operative care, and immunosuppressive therapy, as well as surgeons who had conducted at least 10 heart transplants in the last five years. Now suppose that, in our observational data, all surgeons have conducted only between 5 and 9 heart transplants in the last five years. Then, our carefully defined counterfactual outcome $Y^{a=1}$

cannot be linked to any of the observed outcomes Y because nobody in the study population received the treatment version $a = 1$.

That is, the validity of the consistency condition is threatened by ill-defined treatments like "weight gain" (previous section), but also by sufficiently well-defined treatments like "heart transplant" that are absent in the data (this section). This latter problem is the result of a lack of (unconditional) positivity for the treatment of interest.

To link the counterfactual outcomes $Y^{a=1}$ and the observed outcomes Y, we have to ensure that only individuals receiving treatment version $a = 1$ are considered as treated individuals ($A = 1$) in the analysis, and analogously for the untreated. The implication is that, if we want to quantify the causal effect $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$ using observational data, we need data in which some individuals received $a = 1$ and $a = 0$. Being able to describe a well-defined intervention a, as we did, is not helpful if the intervention does not occur in the observed data, i.e., if we cannot reasonably assume that the equality $Y^a = Y$ holds for at least some individuals.

An obvious approach to handling the mismatch between the treatment version of interest and the treatment versions in the observed data is to hypothesize that the effects of those versions of treatment are identical—that is, that there is *treatment variation irrelevance* (see Fine Point 1.2). In some cases, this hypothesis may be a good approximation. For example, it might be argued that no additional relevant experience is gained after performing 5 transplants and, thus, that we can use observational data in which all surgeons conducted at least 5 transplants to correctly identify the effect of "heart transplant" $a = 1$ under the protocol requirement that all surgeons had previously performed at least 10 heart transplants.

Assuming treatment variation irrelevance may also be reasonable for some biological factors. For example, if interested in the effect of pharmacologically-induced changes in blood pressure, investigators may be willing to assume that the particular medication used to achieve the change is not relevant.

In other cases, however, this assumption may not be reasonable. For example, it would be hard to justify that the use of observational data in which all surgeons had conducted only 1 transplant to correctly identify the effect of "heart transplant" $a = 1$ under the protocol requirement that all surgeons had previously performed at least 10 heart transplants.

Matching the intervention of interest $a = 1$ with the observed "treatment" value $A = 1$, and therefore equating the counterfactual outcome $Y^{a=1}$ with the observed outcome $Y^A = Y$, requires collecting detailed data. Not only is this information necessary to detect a mismatch between the treatment of interest and the data at hand, but also to have an informed discussion about whether the available versions of treatment can be used in lieu of the treatment version of interest. In our heart transplant study in this section, if information on surgeon experience was not collected, we would not be able to determine whether the counterfactual $Y^{a=1}$ can be linked to the observed Y.

Confusion often arises from the common practice of using the same letter to refer to the hypothetical intervention a and to the observed value A before enough information exists to match a and A.

Because data on treatment versions are often unavailable in observational studies, consistency is often compromised. Since achieving consistency is not easy in observational studies, a good practice is to make our reasoning as transparent as possible, so that others can directly challenge our arguments. The next section describes a procedure to achieve that transparency.

3.6 The target trial

The average causal effect is a contrast between mean counterfactual outcomes under different treatment values. Because the counterfactual outcomes need to be well defined, we can imagine a (hypothetical) randomized experiment to

quantify the causal effect of interest. We refer to that hypothetical experiment as the target experiment or the *target trial*. When conducting the target trial is not feasible, ethical, or timely, we resort to causal analyses of observational data. Generally, one can view these observational analyses as an attempt to emulate some target trial. If the emulation were successful, there would be no difference between the results from the observational study and from the target trial (had it been conducted).

The target trial—or its logical equivalents—has long been central to the causal inference framework. Dorn (1953), Wold (1954), Cochran (1972), Rubin (1974), Feinstein (1971), and Dawid (2000) used the concept. Robins (1986) generalized it for time-varying treatments.

In this chapter, we have explored three conditions—exchangeability, positivity, consistency—that help equate an observational study with a (conditionally randomized) target experiment. When these conditions hold, we can apply the methods described in the previous chapter—IP weighting or standardization—to compute causal effects from the observational data.

Fine Point 3.6 describes how to use observational data to compute the proportion of cases attributable to treatment.

Therefore "what randomized experiment are you trying to emulate?" can be a key question for causal inference from observational data. For each causal effect that we wish to estimate using observational data, we may (i) specify the target trial that we would like to, but cannot, conduct, and (ii) describe how the observational data can be used to emulate that target trial. Specifying the target trial, and therefore the causal effect of interest, requires specifying key components of the trial's protocol: eligibility criteria, interventions (or, in general, treatment strategies), assignment, outcomes, start and end of follow-up, and causal contrasts.

Hernán and Robins (2016) specified the key components of the target trial. The acronym PICO (Population, Intervention, Comparator, Outcome) is sometimes used to summarize some of those components (Richardson et al. 1995).

Therefore, a valid emulation of the target trial requires that the observational dataset includes sufficient information to identify eligible individuals, assign them to groups defined by the interventions they receive, and ascertain their outcomes during the follow-up. For example, to estimate the causal effect of heart transplant, we first specify the components of the protocol of the target trial, and then try to emulate each of them using the observational data. Such explicit emulation of a target trial improves causal inference from observational data by making the interventions, and therefore the causal question, well-defined (see Chapter 22 for an extended discussion of the target trial framework). Once the causal question is well-defined via a target trial, investigators can focus on whether and how conditional exchangeability across groups can be achieved.

When we are concerned that assuming conditional exchangeability may not be reasonable given the available data, we can consider alternative identifying assumptions (see Chapter 16) or perform sensitivity analyses.

All of the above assumes that the interventions of interest are sufficiently well-defined to translate them into a hypothetical experiment. But what can we do when, based on current scientific knowledge, the causal question cannot be translated into a target trial? As an example, consider "the causal effect of weight loss" on mortality in individuals who are obese and do not smoke at age 40. Because this causal question is somewhat vague, we replace it by one that specifies the actual intervention that would be implemented to bring about weight loss. For example, we could specify and emulate a target trial about inducing weight loss via, say, exercise or diet.

In contrast to the conceptualization of causal inference from observational data as a target trial emulation, some authors view "the causal effect of A on Y" as a well-defined quantity regardless of what A and Y stand for (as long as A temporally precedes Y). Often the argument goes like this:

For some examples of this point of view, see Pearl (2009), Schwartz et al (2016), and Glymour and Spiegelman (2016).

> Requiring well-defined counterfactual outcomes via a target trial imposes severe restrictions on the causal questions that can be asked. Suppose we study weight loss and heart disease using observational data. We may not be able to characterize precisely the causal effect of weight loss, but is that really so important if indeed some causal effect exists? There is value in learning that many

Fine Point 3.6

Attributable fraction. We have described effect measures like the causal risk ratio $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$ and the causal risk difference $\Pr[Y^{a=1}=1]-\Pr[Y^{a=0}=1]$, which compare the counterfactual risk under treatment $a=1$ with the counterfactual risk under treatment $a=0$. However, one could also be interested in measures that compare the observed risk with the counterfactual risk under either treatment $a=1$ or $a=0$. This latter contrast allows us to compute the proportion of cases that are attributable to treatment in an observational study, i.e., the proportion of cases that would not have occurred had treatment not occurred. For example, suppose that all 20 individuals in our population attended a dinner in which they were served either ambrosia $(A=1)$ or nectar $(A=0)$. The following day, 7 of the 10 individuals who received $A=1$, and 1 of the 10 individuals who received $A=0$, were sick. For simplicity, assume exchangeability of the treated and the untreated so that the causal risk ratio is $0.7/0.1=7$ and the causal risk difference is $0.7-0.1=0.6$. (In conditionally randomized experiments, one would compute these effect measures via standardization or IP weighting.) It was later discovered that the ambrosia had been contaminated by a flock of doves, which explains the increased risk summarized by both the causal risk ratio and the causal risk difference. We now address the question "What fraction of the cases was attributable to consuming ambrosia?"

In this study we observed 8 cases, i.e., the observed risk was $\Pr[Y=1]=8/20=0.4$. The risk that would have been observed if everybody had received $a=0$ is $\Pr[Y^{a=0}=1]=0.1$. The difference between these two risks is $0.4-0.1=0.3$. That is, there is an excess 30% of the individuals who did fall ill but would not have fallen ill if everybody in the population had received $a=0$ rather than their treatment A. Because $0.3/0.4=0.75$, we say that 75% of the cases are attributable to treatment $a=1$: compared with the 8 observed cases, only 2 cases would have occurred if everybody had received $a=0$. This *excess fraction* or *attributable fraction* is defined as

$$\frac{\Pr[Y=1]-\Pr[Y^{a=0}=1]}{\Pr[Y=1]}$$

See Fine Point 5.4 for a discussion of the excess fraction in the context of the sufficient-component-cause framework.

The excess fraction is generally different from the *etiologic fraction*, another version of the attributable fraction which is defined as the proportion of cases mechanically caused by exposure. For example, suppose the untreated $(A=0)$ would have had 7 cases if they have been treated, but these 7 cases would not have contained the 1 untreated case that actually occurred, i.e., treatment produces 7 cases but prevents 1 case. Also suppose that, if untreated, the treated would have had only 1 case but different from the 7 cases they actually had. Then the excess fraction would not be equal to the etiologic fraction. Here the excess fraction is a lower bound on the etiologic fraction. Because the etiologic fraction does not rely on the concept of excess cases, it can only be computed in randomized experiments under strong assumptions. See Greenland and Robins, 1988 and Robins and Greenland, 1989.

> deaths can be prevented if obese people, somehow, lost weight, even if the intervention required for achieving that transformation is unspecified.

This is an appealing argument, but it is problematic for two reasons.

First, unspecified interventions may be unreasonable or impractical. For example, suppose that some investigators are interested in learning about the health effects of weight loss, but they do not propose and explicitly emulate a target trial. Rather, they conduct a simplified analysis that compares the risk of death in obese versus non-obese individuals at age 40. That comparison corresponds implicitly to a target trial in which obese individuals are instantaneously transformed into individuals with a body mass index of 25 at baseline (through a massive liposuction?). Such target trial cannot be emulated because very few people, if anyone, undergo such drastic instantaneous change in the real world, and thus the counterfactual outcomes cannot generally be linked to the observed outcomes. Had this draconian intervention been made

Extreme or impossible interventions are more likely to go unrecognized when they are not explicitly specified.

explicit, all scientists, including those who conducted the data analyses, would have agreed that consistency does not hold. Explicit target trial emulation prevents investigators from making implicit consistency assumptions that do not cohere with their own beliefs.

In fact, anchoring observational data analyses to a target trial not only helps sharpen the causal question, but also makes the causal inference more relevant for decision makers. This is the case even if the available data are insufficient to emulate a contrast of well-defined interventions. For example, we may not have sufficient data to emulate an intervention that precisely specifies the method used to lose weight (e.g., diet, exercise, a pill), but we may have sufficient data to ensure that other components of the intervention remain realistic. If we had longitudinal data on body weight, we can specify a target trial in which some individuals are assigned to lose 5% of body mass index every year, starting at age 40 and for as long as their body mass index stays over 25. (Part III of this book revolves around interventions that, like this one, are sustained over time.) Though this intervention is not yet well-defined (because it does not specify *how* the weight loss is achieved), it at least avoids mandating an instantaneous weight loss, which corresponds to an intervention that is impossible to implement. That we may not be able to define well-defined counterfactual outcomes is no excuse to try to make them as less ill-defined as possible.

Danaei et al. (2016) tried to estimate the effect of weight loss using observational data. They left unspecified the method used to lose weight, but they carefully specified the timing of the weight loss over many years.

Second, unspecified interventions make it harder to achieve exchangeability and positivity. Because the set of covariates L that result in conditional exchangeability may vary across interventions, the usual uncertainty regarding conditional exchangeability in observational studies is greatly exacerbated if we forgo characterizing the interventions as well as possible. Also, if the interventions remain unspecified, it is hard to characterize the combinations of values of L that would make it impossible to receive the intervention in the observational data, which increases the risk of an inadvertent violation of positivity.

For an extended discussion about the differences between prediction and causal inference, which is a form of counterfactual prediction, see Hernán, Hsu, and Healy (2019).

So what can we do when a target trial cannot be reasonably specified and emulated or, more generally, when the counterfactual outcomes are not sufficiently well defined? In those settings, observational data may still be quite useful for non-causal *prediction*. That obese individuals have a higher mortality risk than nonobese individuals means that obesity is a predictor of—is associated with—mortality. This is an important piece of information to identify individuals at high risk of mortality. By saying that obesity predicts—is associated with—mortality, we remain causally agnostic: obesity might predict mortality in the sense that cigarette smoking predicts lung cancer or in the sense that carrying a lighter predicts lung cancer. Thus the association between obesity and mortality is an interesting hypothesis-generating exercise and a motivation for further research (why does obesity predict mortality anyway?), while acknowledging the magnitude of the association does not necessarily correspond to that of a causal effect.