Causal Inference

Miguel A. Hernán, James M. Robins

August 27, 2015

Contents

Part I

Causal inference without models

Chapter 1
A DEFINITION OF CAUSAL EFFECT

By reading this book you are expressing an interest in learning about causal inference. But, as a human being, you have already mastered the fundamental concepts of causal inference. You certainly know what a causal effect is; you clearly understand the difference between association and causation; and you have used this knowledge constantly throughout your life. In fact, had you not understood these causal concepts, you would have not survived long enough to read this chapter—or even to learn to read. As a toddler you would have jumped right into the swimming pool after observing that those who did so were later able to reach the jam jar. As a teenager, you would have skied down the most dangerous slopes after observing that those who did so were more likely to win the next ski race. As a parent, you would have refused to give antibiotics to your sick child after observing that those children who took their medicines were less likely to be playing in the park the next day.

Since you already understand the definition of causal effect and the difference between association and causation, do not expect to gain deep conceptual insights from this chapter. Rather, the purpose of this chapter is to introduce mathematical notation that formalizes the causal intuition that you already possess. Make sure that you can match your causal intuition with the mathematical notation introduced here. This notation is necessary to precisely define causal concepts, and we will use it throughout the book.

1.1 Individual causal effects

Zeus is a patient waiting for a heart transplant. On January 1, he receives a new heart. Five days later, he dies. Imagine that we can somehow know, perhaps by divine revelation, that had Zeus not received a heart transplant on January 1, he would have been alive five days later. Equipped with this information most would agree that the transplant caused Zeus's death. The heart transplant intervention had a causal effect on Zeus's five-day survival.

Another patient, Hera, also received a heart transplant on January 1. Five days later she was alive. Imagine we can somehow know that, had Hera not received the heart on January 1, she would still have been alive five days later. Hence the transplant did not have a causal effect on Hera's five-day survival.

These two vignettes illustrate how humans reason about causal effects: We compare (usually only mentally) the outcome when an action A is taken with the outcome when the action A is withheld. If the two outcomes differ, we say that the action A has a causal effect, causative or preventive, on the outcome. Otherwise, we say that the action A has no causal effect on the outcome. Epidemiologists, statisticians, economists, and other social scientists often refer to the action A as an intervention, an exposure, or a treatment.

To make our causal intuition amenable to mathematical and statistical analysis we shall introduce some notation. Consider a dichotomous treatment variable A (1: treated, 0: untreated) and a dichotomous outcome variable Y (1: death, 0: survival). In this book we shall refer to variables such as A and Y that may have different values for different individuals or subjects as *random variables*. Let $Y^{a=1}$ (read Y under treatment $a = 1$) be the outcome variable that would have been observed under the treatment value $a = 1$, and $Y^{a=0}$ (read Y under treatment $a = 0$) the outcome variable that would have been

> Capital letters represent random variables. Lower case letters and numbers denote particular values of a random variable

We abbreviate the expression "individual i has outcome $Y = 1$" by writing $Y_i = 1$, and analogously for other random variables.

Causal effect for individual i:
$Y_i^{a=1} \neq Y_i^{a=0}$

observed under the treatment value $a = 0$. $Y^{a=1}$ and $Y^{a=0}$ are also random variables. Zeus has $Y^{a=1} = 1$ and $Y^{a=0} = 0$ because he died when treated but would have survived if untreated, while Hera has $Y^{a=1} = 0$ and $Y^{a=0} = 0$ because she survived when treated and would also have survived if untreated.

We can now provide a formal definition of a *causal effect for an individual*: the treatment A has a causal effect on an individual's outcome Y if $Y^{a=1} \neq Y^{a=0}$ for the individual. Thus the treatment has a causal effect on Zeus's outcome because $Y^{a=1} = 1 \neq 0 = Y^{a=0}$, but not on Hera's outcome because $Y^{a=1} = 0 = Y^{a=0}$. The variables $Y^{a=1}$ and $Y^{a=0}$ are referred to as *potential outcomes* or as *counterfactual outcomes*. Some authors prefer the term "potential outcomes" to emphasize that, depending on the treatment that is received, either of these two outcomes can be potentially observed. Other authors prefer the term "counterfactual outcomes" to emphasize that these outcomes represent situations that may not actually occur (that is, counter to the fact situations).

For each subject, one of the counterfactual outcomes—the one that corresponds to the treatment value that the subject actually received—is actually factual. For example, because Zeus was actually treated $(A = 1)$, his counterfactual outcome under treatment $Y^{a=1} = 1$ is equal to his observed (actual) outcome $Y = 1$. That is, a subject with observed treatment A equal to a, has observed outcome Y equal to his counterfactual outcome Y^a. This equality can be succinctly expressed as $Y = Y^A$ where Y^A denotes the counterfactual Y^a evaluated at the value a corresponding to the subject's observed treatment A. The equality $Y = Y^A$ is referred to as *consistency*.

Consistency:
if $A_i = a$, then $Y_i^a = Y^{A_i} = Y_i$

Individual causal effects are defined as a contrast of the values of counterfactual outcomes, but only one of those outcomes is observed for each individual—the one corresponding to the treatment value actually experienced by the subject. All other counterfactual outcomes remain unobserved. The unhappy conclusion is that, in general, individual causal effects cannot be identified, i.e., computed from the observed data, because of missing data. (See Fine Point 2.1 for a possible exception.)

1.2 Average causal effects

We needed three pieces of information to define an individual causal effect: an outcome of interest, the actions $a = 1$ and $a = 0$ to be compared, and the individual whose counterfactual outcomes $Y^{a=0}$ and $Y^{a=1}$ are to be compared. However, because identifying individual causal effects is generally not possible, we now turn our attention to an aggregated causal effect: the average causal effect in a population of individuals. To define it, we need three pieces of information: an outcome of interest, the actions $a = 1$ and $a = 0$ to be compared, and a well defined population of individuals whose outcomes $Y^{a=0}$ and $Y^{a=1}$ are to be compared.

Take Zeus's extended family as our population of interest. Table 1.1 shows the counterfactual outcomes under both treatment $(a = 1)$ and no treatment $(a = 0)$ for all 20 members of our population. Let us first focus our attention on the last column: the outcome $Y^{a=1}$ that would have been observed for each individual if they had received the treatment (a heart transplant). Half of the members of the population (10 out of 20) would have died if they had received a heart transplant. That is, the proportion of individuals that would have developed the outcome had all population subjects received treatment

Fine Point 1.1

Interference between subjects. An implicit assumption in our definition of counterfactual outcome is that a subject's counterfactual outcome under treatment value a does not depend on other subjects' treatment values. For example, we implicitly assumed that Zeus would die if he received a heart transplant, regardless of whether Hera also received a heart transplant. That is, Hera's treatment value did not interfere with Zeus's outcome. On the other hand, suppose that Hera's getting a new heart upsets Zeus to the extent that he would not survive his own heart transplant, even though he would have survived had Hera not been transplanted. In this scenario, Hera's treatment interferes with Zeus's outcome. Interference between subjects is common in studies that deal with contagious agents or educational programs, in which an individual's outcome is influenced by their social interaction with other population members. In the presence of interference, the counterfactual Y_i^a for an individual i is not well defined because an individual's outcome depends also on other individuals' treatment values. As a consequence "the causal effect of heart transplant on Zeus's outcome" is not well defined when there is interference. Rather, one needs to refer to "the causal effect of heart transplant on Zeus's outcome when Hera does not get a new heart" or "the causal effect of heart transplant on Zeus's outcome when Hera does get a new heart." If other relatives and friends' treatment also interfere with Zeus's outcome, then one may need to refer to the causal effect of heart transplant on Zeus's outcome when "no relative or friend gets a new heart," "when only Hera gets a new heart," etc. because the causal effect of treatment on Zeus's outcome may differ for each particular allocation of hearts. The assumption of no interference was labeled "no interaction between units" by Cox (1958), and is included in the "stable-unit-treatment-value assumption (SUTVA)" described by Rubin (1980). Unless otherwise specified, we will assume no interference throughout this book.

Table 1.1

	$Y^{a=0}$	$Y^{a=1}$
Rheia	0	1
Kronos	1	0
Demeter	0	0
Hades	0	0
Hestia	0	0
Poseidon	1	0
Hera	0	0
Zeus	0	1
Artemis	1	1
Apollo	1	0
Leto	0	1
Ares	1	1
Athena	1	1
Hephaestus	0	1
Aphrodite	0	1
Cyclope	0	1
Persephone	1	1
Hermes	1	0
Hebe	1	0
Dionysus	1	0

Average causal effect in population:
$E[Y^{a=1}] \neq E[Y^{a=0}]$

$a = 1$ is $\Pr[Y^{a=1} = 1] = 10/20 = 0.5$. Similarly, from the other column of Table 1.1, we can conclude that half of the members of the population (10 out of 20) would have died if they had not received a heart transplant. That is, the proportion of subjects that would have developed the outcome had all population subjects received no treatment $a = 0$ is $\Pr[Y^{a=0} = 1] = 10/20 = 0.5$. Note that we have computed the counterfactual risk under treatment to be 0.5 by counting the number of deaths (10) and dividing them by the total number of individuals (20), which is the same as computing the average of the counterfactual outcome across all individuals in the population (if you do not see the equivalence between risk and average for a dichotomous outcome, please use the data in Table 1.1 to compute the average of $Y^{a=1}$).

We are now ready to provide a formal definition of the *average causal effect* in the population: an average causal effect of treatment A on outcome Y is present if $\Pr[Y^{a=1} = 1] \neq \Pr[Y^{a=0} = 1]$ in the population of interest. Under this definition, treatment A does not have an average causal effect on outcome Y in our population because both the risk of death under treatment $\Pr[Y^{a=1} = 1]$ and the risk of death under no treatment $\Pr[Y^{a=0} = 1]$ are 0.5. That is, it does not matter whether all or none of the individuals receive a heart transplant: half of them would die in either case. When, like here, the average causal effect in the population is null, we say that the *null hypothesis of no average causal effect* is true. Because the risk equals the average and because the letter E is usually employed to represent the population average or mean (also referred to as 'E'xpectation), we can rewrite the definition of a non-null average causal effect in the population as $E[Y^{a=1}] \neq E[Y^{a=0}]$ so that the definition applies to both dichotomous and nondichotomous outcomes.

The presence of an "average causal effect of heart transplant A" is defined by a contrast that involves the two actions "receiving a heart transplant ($a = 1$)" and "not receiving a heart transplant ($a = 0$)." When more than two actions are possible (i.e., the treatment is not dichotomous), the particular

Fine Point 1.2

Multiple versions of treatment. Another implicit assumption in our definition of a subject's counterfactual outcome under treatment value a is that there is only one version of treatment value $A = a$. For example, we said that Zeus would die if he received a heart transplant. This statement implicitly assumes that all heart transplants are performed by the same surgeon using the same procedure and equipment. That is, that there is only one version of the treatment "heart transplant." If there were multiple versions of treatment (e.g., surgeons with different skills), then it is possible that Zeus would survive if his transplant were performed by Asclepios, and would die if his transplant were performed by Hygieia. In the presence of multiple versions of treatment, the counterfactual Y_i^a for an individual i is not well defined because an individual's outcome depends on the version of treatment a. As a consequence "the causal effect of heart transplant on Zeus's outcome" is not well defined when there are multiple versions of treatment. Rather, one needs to refer to "the causal effect of heart transplant on Zeus's outcome when Asclepios performs the surgery" or "the causal effect of heart transplant on Zeus's outcome when Hygieia performs the surgery." If other components of treatment (e.g., procedure, place) are also relevant to the outcome, then one may need to refer to "the causal effect of heart transplant on Zeus's outcome when Asclepios performs the surgery using his rod at the temple of Kos" because the causal effect of treatment on Zeus's outcome may differ for each particular version of treatment. The assumption of no multiple versions of treatment is included in the "stable-unit-treatment-value assumption (SUTVA)" described by Rubin (1980). VanderWeele (2009) formalized the weaker assumption of "treatment variation irrelevance," i.e., the assumption that multiple versions of treatment $A = a$ may exist but they all result in the same outcome Y_i^a. Unless otherwise specified, we will assume treatment variation irrelevance throughout this book. See Chapter 3 for an extended discussion of this issue.

contrast of interest needs to be specified. For example, "the causal effect of aspirin" is meaningless unless we specify that the contrast of interest is, say, "taking, while alive, 150 mg of aspirin by mouth (or nasogastric tube if need be) daily for 5 years" versus "not taking aspirin." Note that this causal effect is well defined even if counterfactual outcomes under other interventions are not well defined or even do not exist (e.g., "taking, while alive, 500 mg of aspirin by absorption through the skin daily for 5 years").

Absence of an average causal effect does not imply absence of individual effects. In fact, Table 1.1 shows that treatment has an individual causal effect on the outcomes of 12 members (including Zeus) of the population because, for each of these 12 individuals, the value of their counterfactual outcomes $Y^{a=1}$ and $Y^{a=0}$ differ. Six of the twelve (including Zeus) were harmed by treatment $\left(Y^{a=1} - Y^{a=0} = 1\right)$; an equal number were helped $\left(Y^{a=1} - Y^{a=0} = -1\right)$. This equality is not an accident: the average causal effect $E[Y^{a=1}] - E[Y^{a=0}]$ is always equal to the average $E[Y^{a=1} - Y^{a=0}]$ of the individual causal effects $Y^{a=1} - Y^{a=0}$, as a difference of averages is equal to the average of the differences. When there is no causal effect for any individual in the population, i.e., $Y^{a=1} = Y^{a=0}$ for all subjects, we say that the *sharp causal null hypothesis* is true. The sharp causal null hypothesis implies the null hypothesis of no average effect.

As discussed in the next chapters, average causal effects *can* sometimes be identified from data, even if individual causal effects cannot. Hereafter we refer to 'average causal effects' simply as 'causal effects' and the null hypothesis of no average effect as the causal null hypothesis. We next describe different measures of the magnitude of a causal effect.

Technical Point 1.1

Causal effects in the population. Let $E[Y^a]$ be the mean counterfactual outcome had all subjects in the population received treatment level a. For discrete outcomes, the mean or expected value $E[Y^a]$ is defined as the weighted sum $\sum_y y\, p_{Y^a}(y)$ over all possible values y of the random variable Y^a, where $p_{Y^a}(\cdot)$ is the probability mass function of Y^a, i.e., $p_{Y^a}(y) = \Pr[Y^a = y]$. For dichotomous outcomes, $E[Y^a] = \Pr[Y^a = 1]$. For continuous outcomes, the expected value $E[Y^a]$ is defined as the integral $\int y f_{Y^a}(y)\, dy$ over all possible values y of the random variable Y^a, where $f_{Y^a}(\cdot)$ is the probability density function of Y^a. A common representation of the expected value that applies to both discrete and continuous outcomes is $E[Y^a] = \int y dF_{Y^a}(y)$, where $F_{Y^a}(\cdot)$ is the cumulative distribution function (CDF) of the random variable Y^a. We say that there is a non-null average causal effect in the population if $E[Y^a] \neq E[Y^{a'}]$ for any two values a and a'.

The average causal effect, defined by a contrast of means of counterfactual outcomes, is the most commonly used population causal effect. However, a population causal effect may also be defined as a contrast of, say, medians, variances, hazards, or CDFs of counterfactual outcomes. In general, a causal effect can be defined as a contrast of any functional of the distributions of counterfactual outcomes under different actions or treatment values. The causal null hypothesis refers to the particular contrast of functionals (mean, median, variance, hazard, CDF, ...) used to define the causal effect.

1.3 Measures of causal effect

We have seen that the treatment 'heart transplant' A does not have a causal effect on the outcome 'death' Y in our population of 20 family members of Zeus. The causal null hypothesis holds because the two counterfactual risks $\Pr[Y^{a=1} = 1]$ and $\Pr[Y^{a=0} = 1]$ are equal to 0.5. There are equivalent ways of representing the causal null. For example, we could say that the risk $\Pr[Y^{a=1} = 1]$ minus the risk $\Pr[Y^{a=0} = 1]$ is zero $(0.5 - 0.5 = 0)$ or that the risk $\Pr[Y^{a=1} = 1]$ divided by the risk $\Pr[Y^{a=0} = 1]$ is one $(0.5/0.5 = 1)$. That is, we can represent the causal null by

(i) $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1] = 0$

(ii) $\dfrac{\Pr[Y^{a=1} = 1]}{\Pr[Y^{a=0} = 1]} = 1$

(iii) $\dfrac{\Pr[Y^{a=1} = 1]/\Pr[Y^{a=1} = 0]}{\Pr[Y^{a=0} = 1]/\Pr[Y^{a=0} = 0]} = 1$

where the left-hand side of the equalities (i), (ii), and (iii) is the causal risk difference, risk ratio, and odds ratio, respectively.

Suppose now that another treatment A, cigarette smoking, has a causal effect on another outcome Y, lung cancer, in our population. The causal null hypothesis does not hold: $\Pr[Y^{a=1} = 1]$ and $\Pr[Y^{a=0} = 1]$ are not equal. In this setting, the causal risk difference, risk ratio, and odds ratio are not 0, 1, and 1, respectively. Rather, these causal parameters quantify the strength of the same causal effect on different scales. Because the causal risk difference, risk ratio, and odds ratio (and other summaries) measure the causal effect, we refer to them as *effect measures*.

Each effect measure may be used for different purposes. For example, imagine a large population in which 3 in a million individuals would develop the outcome if treated, and 1 in a million individuals would develop the outcome if untreated. The causal risk ratio is 3, and the causal risk difference is 0.000002. The causal risk ratio (multiplicative scale) is used to compute how many times

treatment, relative to no treatment, increases the disease risk. The causal risk difference (additive scale) is used to compute the absolute number of cases of the disease attributable to the treatment. The use of either the multiplicative or additive scale will depend on the goal of the inference.

1.4 Random variability

At this point you could complain that our procedure to compute effect measures is somewhat implausible. Not only did we ignore the well known fact that the immortal Zeus cannot die, but—more to the point—our population in Table 1.1 had only 20 individuals. The populations of interest are typically much larger.

In our tiny population, we collected information from all the subjects. In practice, investigators only collect information on a sample of the population of interest. Even if the counterfactual outcomes of all study subjects were known, working with samples prevents one from obtaining the exact proportion of subjects in the population who had the outcome under treatment value a, e.g., the probability of death under no treatment $\Pr[Y^{a=0} = 1]$ cannot be directly computed. One can only estimate this probability.

Consider the subjects in Table 1.1. We have previously viewed them as forming a twenty-subject population. Suppose we view them as a random sample from a much larger, near-infinite super-population (e.g., all immortals). We denote the proportion of subjects in the sample who would have died if unex-

1$^{\text{st}}$ source of random error: Sampling variability

An estimator $\hat{\theta}$ of θ is consistent if, with probability approaching 1, the difference $\hat{\theta}-\theta$ approaches zero as the sample size increases towards infinity.

Caution: the term 'consistency' when applied to estimators has a different meaning from that which it has when applied to counterfactual outcomes.

2$^{\text{nd}}$ source of random error: Nondeterministic counterfactuals

Table 1.2

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

posed as $\widehat{\Pr}[Y^{a=0} = 1] = 10/20 = 0.50$. The sample proportion $\widehat{\Pr}[Y^{a=0} = 1]$ does not have to be exactly equal to the proportion of subjects who would have died if the entire super-population had been unexposed, $\Pr[Y^{a=0} = 1]$. For example, suppose $\Pr[Y^{a=0} = 1] = 0.57$ in the population but, because of random error due to sampling variability, $\widehat{\Pr}[Y^{a=0} = 1] = 0.5$ in our particular sample. We use the sample proportion $\widehat{\Pr}[Y^{a} = 1]$ to estimate the super-population probability $\Pr[Y^{a} = 1]$ under treatment value a. The "hat" over Pr indicates that the sample proportion $\widehat{\Pr}[Y^{a} = 1]$ is an estimator of the corresponding population quantity $\Pr[Y^{a} = 1]$. We say that $\widehat{\Pr}[Y^{a} = 1]$ is a *consistent estimator* of $\Pr[Y^{a} = 1]$ because the larger the number of subjects in the sample, the smaller the difference between $\widehat{\Pr}[Y^{a} = 1]$ and $\Pr[Y^{a} = 1]$ is expected to be. This occurs because the error due to sampling variability is random and thus obeys the law of large numbers.

Because the super-population probabilities $\Pr[Y^{a} = 1]$ cannot be computed, only consistently estimated by the sample proportions $\widehat{\Pr}[Y^{a} = 1]$, one cannot conclude with certainty that there is, or there is not, a causal effect. Rather, a statistical procedure must be used to test the causal null hypothesis $\Pr[Y^{a=1} = 1] = \Pr[Y^{a=0} = 1]$; the procedure quantifies the chance that the difference $\widehat{\Pr}[Y^{a=1} = 1]$ and $\widehat{\Pr}[Y^{a=0} = 1]$ is wholly due to sampling variability.

So far we have only considered sampling variability as a source of random error. But there may be another source of random variability: perhaps the values of an individual's counterfactual outcomes are not fixed in advance. We have defined the counterfactual outcome Y^{a} as the subject's outcome had he received treatment value a. For example, in our first vignette, Zeus would have died if treated and would have survived if untreated. As defined, the values of the counterfactual outcomes are fixed or deterministic for each subject, e.g., $Y^{a=1} = 1$ and $Y^{a=0} = 0$ for Zeus. In other words, Zeus has a 100% chance of dying if treated and a 0% chance of dying if untreated. However, we could imagine another scenario in which Zeus has a 90% chance of dying if treated, and a 10% chance of dying if untreated. In this scenario, the counterfactual outcomes are stochastic or nondeterministic because Zeus's probabilities of dying under treatment (0.9) and under no treatment (0.1) are neither zero or one. The values of $Y^{a=1}$ and $Y^{a=0}$ shown in Table 1.1 would be possible realizations of "random flips of mortality coins" with these probabilities. Further, one would expect that these probabilities vary across subjects because not all subjects are equally susceptible to develop the outcome. Quantum mechanics, in contrast to classical mechanics, holds that outcomes are inherently nondeterministic. That is, if the quantum mechanical probability of Zeus dying is 90%, the theory holds that no matter how much data we collect about Zeus, the uncertainty about whether Zeus will actually develop the outcome if treated is irreducible and statistical methods are needed to quantify it.

Thus statistics is necessary in causal inference to quantify random error from sampling variability, nondeterministic counterfactuals, or both. However, for pedagogic reasons, we will continue to largely ignore statistical issues until Chapter 10. Specifically, we will assume that counterfactual outcomes are deterministic and that we have recorded data on every subject in a very large (perhaps hypothetical) super-population. This is equivalent to viewing our population of 20 subjects as a population of 20 billion subjects in which 1 billion subjects are identical to Zeus, 1 billion subjects are identical to Hera, and so on. Hence, until Chapter 10, we will carry out our computations with Olympian certainty.

Technical Point 1.2

Nondeterministic counterfactuals. For nondeterministic counterfactual outcomes, the mean outcome under treatment value a, $\mathrm{E}[Y^a]$, equals the weighted sum $\sum_y y\, p_{Y^a}(y)$ over all possible values y of the random variable Y^a, where the probability mass function $p_{Y^a}(\cdot) = \mathrm{E}\left[Q_{Y^a}(\cdot)\right]$, and $Q_{Y^a}(y)$ is a random probability of having outcome $Y = y$ under treatment level a. In the example described in the text, $Q_{Y^{a=1}}(1) = 0.9$ for Zeus. (For continuous outcomes, the weighted sum is replaced by an integral.)

More generally, a nondeterministic definition of counterfactual outcome does not attach some particular value of the random variable Y^a to each subject, but rather a statistical distribution $\Theta_{Y^a}(\cdot)$ of Y^a. The nondeterministic definition of causal effect is a generalization of the deterministic definition in which $\Theta_{Y^a}(\cdot)$ is a random CDF that may take values between 0 and 1. The average counterfactual outcome in the population $\mathrm{E}[Y^a]$ equals $\mathrm{E}\left\{\mathrm{E}\left[Y^a \mid \Theta_{Y^a}(\cdot)\right]\right\}$. Therefore, $\mathrm{E}[Y^a] = \mathrm{E}\left[\int y\, d\Theta_{Y^a}(y)\right] = \int y\, d\,\mathrm{E}[\Theta_{Y^a}(y)] = \int y\, dF_{Y^a}(y)$, because we define $F_{Y^a}(\cdot) = \mathrm{E}\left[\Theta_{Y^a_i}(\cdot)\right]$. Although the possibility of nondeterministic counterfactual outcomes implies no changes in our definitions of population causal effect and of effect measures, nondeterministic counterfactual outcomes introduce random variability. This additional variability has implications for the computation of confidence intervals for the effect measures (Robins 1988), as discussed in Chapter 10.

1.5 Causation versus association

Obviously, the data available from actual studies look different from those shown in Table 1.1. For example, we would not usually expect to learn Zeus's outcome if treated $Y^{a=1}$ *and* also Zeus's outcome if untreated $Y^{a=0}$. In the real world, we only get to observe one of those outcomes because Zeus is either treated or untreated. We referred to the observed outcome as Y. Thus, for each individual, we know the observed treatment level A and the outcome Y as in Table 1.2.

The data in Table 1.2 can be used to compute the proportion of subjects that developed the outcome Y among those subjects in the population that happened to receive treatment value a. For example, in Table 1.2, 7 subjects died ($Y = 1$) among the 13 individuals that were treated ($A = 1$). Thus the risk of death in the treated, $\Pr[Y = 1|A = 1]$, was 7/13. In general, we define the conditional probability $\Pr[Y = 1|A = a]$ as the proportion of subjects that developed the outcome Y among those subjects in the population of interest that happened to receive treatment value a.

When the proportion of subjects who develop the outcome in the treated $\Pr[Y = 1|A = 1]$ equals the proportion of subjects who develop the outcome in the untreated $\Pr[Y = 1|A = 0]$, we say that treatment A and outcome Y are independent, that A is not associated with Y, or that A does not predict Y. *Independence* is represented by $Y \coprod A$—or, equivalently, $A \coprod Y$— which is read as Y and A are independent. Some equivalent definitions of independence are

Dawid (1979) introduced the symbol ⫫ to denote independence

(i) $\Pr[Y = 1|A = 1] - \Pr[Y = 1|A = 0] = 0$

(ii) $\dfrac{\Pr[Y = 1|A = 1]}{\Pr[Y = 1|A = 0]} = 1$

(iii) $\dfrac{\Pr[Y = 1|A = 1]/\Pr[Y = 0|A = 1]}{\Pr[Y = 1|A = 0]/\Pr[Y = 0|A = 0]} = 1$

where the left-hand side of the inequalities (i), (ii), and (iii) is the associational risk difference, risk ratio, and odds ratio, respectively.

For a continuous outcome Y we define *mean independence* between treatment and outcome as: $E[Y|A = 1] = E[Y|A = 0]$. Independence and mean independence are the same concept for dichotomous outcomes.

We say that treatment A and outcome Y are dependent or associated when $\Pr[Y = 1|A = 1] \neq \Pr[Y = 1|A = 0]$. In our population, treatment and outcome are indeed associated because $\Pr[Y = 1|A = 1] = 7/13$ and $\Pr[Y = 1|A = 0] = 3/7$. The associational risk difference, risk ratio, and odds ratio (and other measures) quantify the strength of the association when it exists. They measure the association on different scales, and we refer to them as *association measures*. These measures are also affected by random variability. However, until Chapter 10, we will disregard statistical issues by assuming that the population in Table 1.2 is extremely large.

For dichotomous outcomes, the risk equals the average in the population, and we can therefore rewrite the definition of association in the population as $E[Y|A = 1] \neq E[Y|A = 0]$. For continuous outcomes Y, we can also define association as $E[Y|A = 1] \neq E[Y|A = 0]$. Under this definition, association is essentially the same as the statistical concept of *correlation* between A and a continuous Y.

In our population of 20 individuals, we found (i) no causal effect after comparing the risk of death if all 20 individuals had been treated with the risk of death if all 20 individuals had been untreated, and (ii) an association after comparing the risk of death in the 13 individuals who happened to be treated with the risk of death in the 7 individuals who happened to be untreated. Figure 1.1 depicts the causation-association difference. The population (represented by a diamond) is divided into a white area (the treated) and a smaller grey area (the untreated). The definition of causation implies a contrast between the whole white diamond (all subjects treated) and the whole grey diamond (all subjects untreated), whereas association implies a contrast between the white (the treated) and the grey (the untreated) areas of the original diamond.
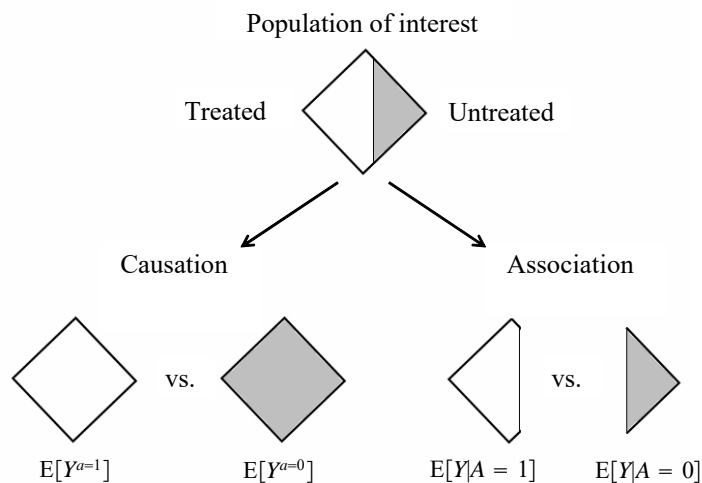
Figure 1.1

We can use the notation we have developed thus far to formalize the distinction between causation and association. The risk $\Pr[Y = 1|A = a]$ is a conditional probability: the risk of Y in the subset of the population that meet the condition 'having actually received treatment value a' (i.e., $A = a$). In contrast the risk $\Pr[Y^a = 1]$ is an unconditional—also known as marginal—probability, the risk of Y^a in the entire population. Therefore, *association* is defined by a different risk in two disjoint subsets of the population determined

by the subjects' actual treatment value ($A = 1$ or $A = 0$), whereas *causation* is defined by a different risk in the entire population under two different treatment values ($a = 1$ or $a = 0$). Throughout this book we often use the redundant expression 'causal effect' to avoid confusions with a common use of 'effect' meaning simply association.

The difference between association and causation is critical. Suppose the causal risk ratio of 5-year mortality is 0.5 for aspirin vs. no aspirin, and the corresponding associational risk ratio is 1.5. After a physician learns these results, she decides to withhold aspirin from her patients because those treated with aspirin have a greater risk of dying compared with the untreated. The doctor will be sued for malpractice.

These radically different definitions explain the well-known adage "association is not causation." In our population, there was association because the mortality risk in the treated ($7/13$) was greater than that in the untreated ($3/7$). However, there was no causation because the risk if everybody had been treated ($10/20$) was the same as the risk if everybody had been untreated. This discrepancy between causation and association would not be surprising if those who received heart transplants were, on average, sicker than those who did not receive a transplant. In Chapter 7 we refer to this discrepancy as *confounding*.

Causal inference requires data like the hypothetical data in Table 1.1, but all we can ever expect to have is real world data like those in Table 1.2. The question is then under which conditions real world data can be used for causal inference. The next chapter provides one answer: conduct a randomized experiment.

Chapter 2
RANDOMIZED EXPERIMENTS

Does your looking up at the sky make other pedestrians look up too? This question has the main components of any causal question: we want to know whether certain action (your looking up) affects certain outcome (other people's looking up) in certain population (say, residents of Madrid in 2011). Suppose we challenge you to design a scientific study to answer this question. "Not much of a challenge," you say after some thought, "I can stand on the sidewalk and flip a coin whenever someone approaches. If heads, I'll look up intently; if tails, I'll look straight ahead with an absentminded expression. I'll repeat the experiment a few thousand times. If the proportion of pedestrians who looked up within 10 seconds after I did is greater than the proportion of pedestrians who looked up when I didn't, I will conclude that my looking up has a causal effect on other people's looking up. By the way, I may hire an assistant to record what people do while I'm looking up." After conducting this study, you found that 55% of pedestrians looked up when you looked up but only 1% looked up when you looked straight ahead.

Your solution to our challenge was to conduct a randomized experiment. It was an experiment because the investigator (you) carried out the action of interest (looking up), and it was randomized because the decision to act on any study subject (pedestrian) was made by a random device (coin flipping). Not all experiments are randomized. For example, you could have looked up when a man approached and looked straight ahead when a woman did. Then the assignment of the action would have followed a deterministic rule (up for man, straight for woman) rather than a random mechanism. However, your findings would not have been nearly as convincing if you had conducted a non randomized experiment. If your action had been determined by the pedestrian's sex, critics could argue that the "looking up" behavior of men and women differs (women may not be as easily influenced by your actions) and thus your study compared essentially "noncomparable" groups of people. This chapter describes why randomization results in convincing causal inferences.

2.1 Randomization

In a real world study we will not know both of Zeus's potential outcomes $Y^{a=1}$ under treatment and $Y^{a=0}$ under no treatment. Rather, we can only know his observed outcome Y under the treatment value A that he happened to receive. Table 2.1 summarizes the available information for our population of 20 individuals. Only one of the two counterfactual outcomes is known for each individual: the one corresponding to the treatment level that he actually received. The data are missing for the other counterfactual outcomes. As we discussed in the previous chapter, this missing data creates a problem because it appears that we need the value of both counterfactual outcomes to compute effect measures. The data in Table 2.1 are only good to compute association measures.

Randomized experiments, like any other real world study, generate data with missing values of the counterfactual outcomes as shown in Table 2.1. However, randomization ensures that those missing values occurred by chance. As a result, effect measures can be computed —or, more rigorously, consistently estimated—in randomized experiments despite the missing data. Let us be more precise.

Suppose that the population represented by a diamond in Figure 1.1 was near-infinite, and that we flipped a coin for each subject in such population. We

Neyman (1923) applied counterfactual theory to the estimation of causal effects via randomized experiments

Table 2.1

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

Exchangeability:
$Y^a \coprod A$ for all a

assigned the subject to the white group if the coin turned tails, and to the grey group if it turned heads. Note this was not a fair coin because the probability of heads was less than 50%—fewer people ended up in the grey group than in the white group. Next we asked our research assistants to administer the treatment of interest ($A = 1$), to subjects in the white group and a placebo ($A = 0$) to those in the grey group. Five days later, at the end of the study, we computed the mortality risks in each group, $\Pr[Y = 1|A = 1] = 0.3$ and $\Pr[Y = 1|A = 0] = 0.6$. The associational risk ratio was $0.3/0.6 = 0.5$ and the associational risk difference was $0.3 - 0.6 = -0.3$. We will assume that this was an *ideal randomized experiment* in all other respects: no loss to follow-up, full adherence to the assigned treatment over the duration of the study, a single version of treatment, and double blind assignment (see Chapter 9). Ideal randomized experiments are unrealistic but useful to introduce some key concepts for causal inference. Later in this book we consider more realistic randomized experiments.

Now imagine what would have happened if the research assistants had misinterpreted our instructions and had treated the grey group rather than the white group. Say we learned of the misunderstanding after the study finished. How does this reversal of treatment status affect our conclusions? Not at all. We would still find that the risk in the treated (now the grey group) $\Pr[Y = 1|A = 1]$ is 0.3 and the risk in the untreated (now the white group) $\Pr[Y = 1|A = 0]$ is 0.6. The association measure would not change. Because subjects were randomly assigned to white and grey groups, the proportion of deaths among the exposed, $\Pr[Y = 1|A = 1]$ is expected to be the same whether subjects in the white group received the treatment and subjects in the grey group received placebo, or vice versa. When group membership is randomized, which particular group received the treatment is irrelevant for the value of $\Pr[Y = 1|A = 1]$. The same reasoning applies to $\Pr[Y = 1|A = 0]$, of course. Formally, we say that groups are exchangeable.

Exchangeability means that the risk of death in the white group would have been the same as the risk of death in the grey group had subjects in the white group received the treatment given to those in the grey group. That is, the risk under the potential treatment value a among the treated, $\Pr[Y^a = 1|A = 1]$, equals the risk under the potential treatment value a among the untreated, $\Pr[Y^a = 1|A = 0]$, for both $a = 0$ and $a = 1$. An obvious consequence of these (conditional) risks being equal in all subsets defined by treatment status in the population is that they must be equal to the (marginal) risk under treatment value a in the whole population: $\Pr[Y^a = 1|A = 1] = \Pr[Y^a = 1|A = 0] = \Pr[Y^a = 1]$. Because the counterfactual risk under treatment value a is the same in both groups $A = 1$ and $A = 0$, we say that the actual treatment A does not predict the counterfactual outcome Y^a. Equivalently, exchangeability means that the counterfactual outcome and the actual treatment are independent, or $Y^a \coprod A$, for all values a. Randomization is so highly valued because it is expected to produce exchangeability. When the treated and the untreated are exchangeable, we sometimes say that treatment is exogenous, and thus *exogeneity* is commonly used as a synonym for exchangeability.

The previous paragraph argues that, in the presence of exchangeability, the counterfactual risk under treatment in the white part of the population would equal the counterfactual risk under treatment in the entire population. But the risk under treatment in the white group is not counterfactual at all because the white group was actually treated! Therefore our ideal randomized experiment allows us to compute the counterfactual risk under treatment in the population $\Pr[Y^{a=1} = 1]$ because it is equal to the risk in the treated $\Pr[Y = 1|A = 1] =$

Technical Point 2.1

Full exchangeability and mean exchangeability. Randomization makes the Y^a jointly independent of A which implies, but is not implied by, exchangeability $Y^a \coprod A$ for each a. Formally, let $\mathcal{A} = \{a, a', a'', ...\}$ denote the set of all treatment values present in the population, and $Y^{\mathcal{A}} = \left\{Y^a, Y^{a'}, Y^{a''}, ...\right\}$ the set of all counterfactual outcomes. Randomization makes $Y^{\mathcal{A}} \coprod A$. We refer to this joint independence as *full exchangeability*. For a dichotomous treatment, $\mathcal{A} = \{0, 1\}$ and full exchangeability is $\left(Y^{a=1}, Y^{a=0}\right) \coprod A$.

For a dichotomous outcome and treatment, exchangeability $Y^a \coprod A$ can also be written as $\Pr\left[Y^a = 1 | A = 1\right] = \Pr\left[Y^a = 1 | A = 0\right]$ or, equivalently, as $\mathrm{E}[Y^a | A = 1] = \mathrm{E}[Y^a | A = 0]$ for all a. We refer to the last equality as *mean exchangeability*. For a continuous outcome, exchangeability $Y^a \coprod A$ implies mean exchangeability $\mathrm{E}[Y^a | A = a'] = \mathrm{E}[Y^a]$, but mean exchangeability does not imply exchangeability because distributional parameters other than the mean (e.g., variance) may not be independent of treatment.

Neither full exchangeability $Y^{\mathcal{A}} \coprod A$ nor exchangeability $Y^a \coprod A$ are required to prove that $\mathrm{E}[Y^a] = \mathrm{E}[Y | A = a]$. Mean exchangeability is sufficient. As sketched in the main text, the proof has two steps. First, $\mathrm{E}[Y | A = a] = \mathrm{E}[Y^a | A = a]$ by consistency. Second, $\mathrm{E}[Y^a | A = a] = \mathrm{E}[Y^a]$ by mean exchangeability. Because exchangeability and mean exchangeability are identical concepts for the dichotomous outcomes used in this chapter, we use the shorter term "exchangeability" throughout.

There are scenarios (e.g., the estimation of causal effects in randomized experiments with noncompliance) in which the fact that randomization implies joint independence rather than simply marginal or mean independence is of critical importance.

0.3. That is, the risk in the treated (the white part of the diamond) is the same as the risk if everybody had been treated (and thus the diamond had been entirely white). Of course, the same rationale applies to the untreated: the counterfactual risk under no treatment in the population $\Pr[Y^{a=0} = 1]$ equals the risk in the untreated $\Pr[Y = 1 | A = 0] = 0.6$. The causal risk ratio is 0.5 and the causal risk difference is -0.3. In ideal randomized experiments, association *is* causation.

Before proceeding, please make sure you understand the difference between $Y^a \coprod A$ and $Y \coprod A$. Exchangeability $Y^a \coprod A$ is defined as independence between the counterfactual outcome and the observed treatment. Again, this means that the treated and the untreated would have experienced the same risk of death if they had received the same treatment level (either $a = 0$ or $a = 1$). But independence between the counterfactual outcome and the observed treatment $Y^a \coprod A$ does not imply independence between the observed outcome and the observed treatment $Y \coprod A$. For example, in a randomized experiment in which exchangeability $Y^a \coprod A$ holds and the treatment has a causal effect on the outcome, then $Y \coprod A$ does not hold because the treatment is associated with the observed outcome.

Does exchangeability hold in our heart transplant study of Table 2.1? To answer this question we would need to check whether $Y^a \coprod A$ holds for $a = 0$ and for $a = 1$. Take $a = 0$ first. Suppose the counterfactual data in Table 1.1 are available to us. We can then compute the risk of death under no treatment $\Pr[Y^{a=0} = 1 | A = 1] = 7/13$ in the 13 treated subjects and the risk of death under no treatment $\Pr[Y^{a=0} = 1 | A = 0] = 3/7$ in the 7 untreated subjects. Since the risk of death under no treatment is greater in the treated than in the untreated subjects, i.e., $7/13 > 3/7$, we conclude that the treated have a worse prognosis than the untreated, that is, that the treated and the untreated are not exchangeable. Mathematically, we have proven that exchangeability $Y^a \coprod A$ does not hold for $a = 0$. (You can check that it does not hold for $a = 1$

Caution:
$Y^a \coprod A$ is different from $Y \coprod A$

Fine Point 2.1

Crossover randomized experiments. Individual (also known as subject-specific) causal effects can sometimes be identified via randomized experiments. For example, suppose we want to estimate the causal effect of lightning bolt use A on Zeus's blood pressure Y. We define the counterfactual outcomes $Y^{a=1}$ and $Y^{a=0}$ to be 1 if Zeus's blood pressure is temporarily elevated after calling or not calling a lightning strike, respectively. Suppose we convinced Zeus to use his lightning bolt only when suggested by us. Yesterday morning we flipped coin and obtained heads. We then asked Zeus to call a lightning strike ($a = 1$). His blood pressure was elevated after doing so. This morning we flipped a coin and obtained tails. We then asked Zeus to refrain from using his lightning bolt ($a = 0$). His blood pressure did not increase. We have conducted a *crossover randomized experiment* in which an individual's outcome is sequentially observed under two treatment values. One might argue that, because we have observed both of Zeus's counterfactual outcomes $Y^{a=1} = 1$ and $Y^{a=0} = 0$, using a lightning bolt has a causal effect on Zeus's blood pressure. We may repeat this procedure daily for some months to reduce random variability.

In crossover randomized experiments, an individual is observed during two or more periods. The individual receives a different treatment value in each period and the order of treatment values is randomly assigned. The main purported advantage of the crossover design is that, unlike in non crossover designs, for each treated subject there is a perfectly exchangeable untreated subject—him or herself. A direct contrast of a subject's outcomes under different treatment values allows the identification of individual effects under the following conditions: 1) treatment is of short duration and its effects do not carry-over to the next period, and 2) the outcome is a condition of abrupt onset that completely resolves by the next period. Therefore crossover randomized experiments cannot be used to study the effect of heart transplant, an irreversible action, on death, an irreversible outcome.

To eliminate random variability, one needs to randomly assign treatment at many different periods. If the individual causal effect changes with time, we obtain the average of the individual time-specific causal effects.

either.) Thus the answer to the question that opened this paragraph is 'No'.

But only the observed data in Table 2.1, not the counterfactual data in Table 1.1, are available in the real world. Since Table 2.1 is insufficient to compute counterfactual risks like the risk under no treatment in the treated $\Pr[Y^{a=0} = 1|A = 1]$, we are generally unable to determine whether exchangeability holds in our study. However, suppose for a moment, that we actually had access to Table 1.1 and determined that exchangeability does not hold in our heart transplant study. Can we then conclude that our study is not a randomized experiment? No, for two reasons. First, as you are probably already thinking, a twenty-subject study is too small to reach definite conclusions. Random fluctuations arising from sampling variability could explain almost anything. We will discuss random variability in Chapter 10. Until then, let us assume that each subject in our population represents 1 billion subjects that are identical to him or her. Second, it is still possible that a study is a randomized experiment even if exchangeability does not hold in infinite samples. However, unlike the type of randomized experiment described in this section, it would need to be a randomized experiment in which investigators use more than one coin to randomly assign treatment. The next section describes randomized experiments with more than one coin.

2.2 Conditional randomization

Table 2.2 shows the data from our heart transplant randomized study. Besides data on treatment A (1 if the subject received a transplant, 0 otherwise) and

outcome Y (1 if the subject died, 0 otherwise), Table 2.2 also contains data on the prognosis factor L (1 if the subject was in critical condition, 0 otherwise), which we measured before treatment was assigned. We now consider two mutually exclusive study designs and discuss whether the data in Table 2.2 could have arisen from either of them.

In design 1 we would have randomly selected 65% of the individuals in the population and transplanted a new heart to each of the selected individuals. That would explain why 13 out of 20 subjects were treated. In design 2 we would have classified all individuals as being in either critical ($L = 1$) or noncritical ($L = 0$) condition. Then we would have randomly selected 75% of the individuals in critical condition and 50% of those in noncritical condition, and transplanted a new heart to each of the selected individuals. That would explain why 9 out of 12 subjects in critical condition, and 4 out of 8 subjects in non critical condition, were treated.

Both designs are randomized experiments. Design 1 is precisely the type of randomized experiment described in Section 2.1. Under this design, we would use a single coin to assign treatment to all subjects (e.g., treated if tails, untreated if heads): a loaded coin with probability 0.65 of turning tails, thus resulting in 65% of the subjects receiving treatment. Under design 2 we would not use a single coin for all subjects. Rather, we would use a coin with a 0.75 chance of turning tails for subjects in critical condition, and another coin with a 0.50 chance of turning tails for subjects in non critical condition. We refer to design 2 experiments as *conditionally randomized experiments* because we use several randomization probabilities that depend (are conditional) on the values of the variable L. We refer to design 1 experiments as *marginally randomized experiments* because we use a single unconditional (marginal) randomization probability that is common to all subjects.

As discussed in the previous section, a marginally randomized experiment is expected to result in exchangeability of the treated and the untreated: $\Pr[Y^a = 1|A = 1] = \Pr[Y^a = 1|A = 0]$ or $Y^a \coprod A$. In contrast, a conditionally randomized experiment will not generally result in exchangeability of the treated and the untreated because, by design, each group may have a different proportion of subjects with bad prognosis.

Thus the data in Table 2.2 could not have arisen from a marginally randomized experiment because 69% treated versus 43% untreated individuals were in critical condition. This imbalance indicates that the risk of death in the treated, had they remained untreated, would have been higher than the risk of death in the untreated. In other words, treatment A predicts the counterfactual risk of death under no treatment, and exchangeability $Y^a \coprod A$ does not hold. Since our study was a randomized experiment, you can now safely conclude that the study was a randomized experiment with randomization conditional on L.

Our conditionally randomized experiment is simply the combination of two separate marginally randomized experiments: one conducted in the subset of individuals in critical condition ($L = 1$), the other in the subset of individuals in non critical condition ($L = 0$). Consider first the randomized experiment being conducted in the subset of individuals in critical condition. In this subset, the treated and the untreated are exchangeable. Formally, the counterfactual mortality risk under each treatment value a is the same among the treated and the untreated given that they all were in critical condition at the time of treatment assignment. That is, $\Pr[Y^a = 1|A = 1, L = 1] = \Pr[Y^a = 1|A = 0, L = 1]$ or Y^a and A are independent given $L = 1$, which is written as $Y^a \coprod A|L = 1$ for all a. Similarly, randomization also ensures that the treated

Table 2.2

	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
Cyclope	1	1	1
Persephone	1	1	1
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

and the untreated are exchangeable in the subset of individuals that were in noncritical condition, that is, $Y^a \coprod A|L=0$. When $Y^a \coprod A|L=l$ holds for all values l we simply write $Y^a \coprod A|L$. Thus, although conditional randomization does not guarantee unconditional (or marginal) exchangeability $Y^a \coprod A$, it guarantees *conditional exchangeability* $Y^a \coprod A|L$ within levels of the variable L. In summary, randomization produces either marginal exchangeability (design 1) or conditional exchangeability (design 2).

Conditional exchangeability:
$Y^a \coprod A|L$ for all a

We know how to compute effect measures under marginal exchangeability. In marginally randomized experiments the causal risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$ equals the associational risk ratio $\Pr[Y = 1|A = 1]/\Pr[Y = 1|A = 0]$ because exchangeability ensures that the counterfactual risk under treatment level a, $\Pr[Y^a = 1]$, equals the observed risk among those who received treatment level a, $\Pr[Y = 1|A = a]$. Thus, if the data in Table 2.2 had been collected during a marginally randomized experiment, the causal risk ratio would be readily calculated from the data on A and Y as $\dfrac{7/13}{3/7} = 1.26$.

In a marginally randomized experiment, the values of the counterfactual outcomes are missing completely at random (MCAR). In a conditionally randomized experiment, the values of the counterfactual outcomes are not MCAR, but they are missing at random (MAR) conditional on the covariate L. The terms MCAR, MAR, and NMAR (not missing at random) were introduced by Rubin (1976).

The question is how to compute the causal risk ratio in a conditionally randomized experiment. Remember that a conditionally randomized experiment is simply the combination of two (or more) separate marginally randomized experiments conducted in different subsets of the population, e.g., $L = 1$ and $L = 0$. Thus we have two options.

First, we can compute the average causal effect in each of these subsets of strata of the population. Because association is causation within each subset, the stratum-specific causal risk ratio $\Pr[Y^{a=1} = 1|L = 1]/\Pr[Y^{a=0} = 1|L = 1]$ among people in critical condition is equal to the stratum-specific associational risk ratio $\Pr[Y = 1|L = 1, A = 1]/\Pr[Y = 1|L = 1, A = 0]$ among people in critical condition. And analogously for $L = 0$. We refer to this method to compute stratum-specific causal effects as *stratification*. Note that the stratum-specific causal risk ratio in the subset $L = 1$ may differ from the causal risk ratio in $L = 0$. In that case, we say that the effect of treatment is modified by L, or that there is *effect modification* by L.

Stratification and effect modification are discussed in more detail in Chapter 4.

Second, we can compute the average causal effect $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$ in the entire population, as we have been doing so far. Whether our principal interest lies in the stratum-specific average causal effects versus the average causal effect in the entire population depends on practical and theoretical considerations discussed in detail in Chapter 4 and in Part III. As one example, you may be interested in the average causal effect in the entire population, rather than in the stratum-specific average causal effects, if you do not expect to have information on L for future subjects (e.g., the variable L is expensive to measure) and thus your decision to treat cannot depend on the value of L. Until Chapter 4, we will restrict our attention to the average causal effect in the entire population. The next two sections describe how to use data from conditionally randomized trials to compute the average causal effect in the entire population.

2.3 Standardization

Our heart transplant study is a conditionally randomized experiment: the investigators used a random procedure to assign hearts ($A = 1$) with probability 50% to the 8 individuals in noncritical condition ($L = 0$), and with probability 75% to the 12 individuals in critical condition ($L = 1$). First, let us focus on

the 8 individuals—remember, they are really the average representatives of 8 billion individuals—in noncritical condition. In this group, the risk of death among the treated is $\Pr[Y = 1|L = 0, A = 1] = \frac{1}{4}$, and the risk of death among the untreated is $\Pr[Y = 1|L = 0, A = 0] = \frac{1}{4}$. Because treatment was randomly assigned to subjects in the group $L = 0$, i.e., $Y^a \coprod A|L = 0$, the observed risks are equal to the counterfactual risks. That is, in the group $L = 0$, the risk in the treated equals the risk if everybody had been treated, $\Pr[Y = 1|L = 0, A = 1] = \Pr[Y^{a=1} = 1|L = 0]$, and the risk in the untreated equals the risk if everybody had been untreated, $\Pr[Y = 1|L = 0, A = 0] = \Pr[Y^{a=0} = 1|L = 0]$. Following an analogous reasoning, we can conclude that the observed risks equal the counterfactual risks in the group of 12 individuals in critical condition, i.e., $\Pr[Y = 1|L = 1, A = 1] = \Pr[Y^{a=1} = 1|L = 1] = \frac{2}{3}$, and $\Pr[Y = 1|L = 1, A = 0] = \Pr[Y^{a=0} = 1|L = 1] = \frac{2}{3}$.

Suppose now our goal is to compute the causal risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$. The numerator of the causal risk ratio is the risk if all 20 subjects in the population had been treated. From the previous paragraph, we know that the risk if all subjects had been treated is $\frac{1}{4}$ in the 8 subjects with $L = 0$ and $\frac{2}{3}$ in the 12 subjects with $L = 1$. Therefore the risk if all 20 subjects in the population had been treated will be a weighted average of $\frac{1}{4}$ and $\frac{2}{3}$ in which each group receives a weight proportional to its size. Since 40% of the subjects (8) are in group $L = 0$ and 60% of the subjects (12) and in group $L = 1$, the weighted average is $\frac{1}{4} \times 0.4 + \frac{2}{3} \times 0.6 = 0.5$. Thus the risk if everybody had been treated $\Pr[Y^{a=1} = 1]$ is equal to 0.5. By following the same reasoning we can calculate that the risk if nobody had been treated $\Pr[Y^{a=0} = 1]$ is also equal to 0.5. The causal risk ratio is then $0.5/0.5 = 1$.

More formally, the marginal counterfactual risk $\Pr[Y^a = 1]$ is the weighted average of the stratum-specific risks $\Pr[Y^a = 1|L = 0]$ and $\Pr[Y^a = 1|L = 1]$ with weights equal to the proportion of individuals in the population with $L = 0$ and $L = 1$, respectively. That is, $\Pr[Y^a = 1] = \Pr[Y^a = 1|L = 0]\Pr[L = 0] + \Pr[Y^a = 1|L = 1]\Pr[L = 1]$. Or, using a more compact notation, $\Pr[Y^a = 1] = \sum_l \Pr[Y^a = 1|L = l]\Pr[L = l]$, where $\sum_l$ means sum over all values l that occur in the population. By conditional exchangeability, we can replace the counterfactual risk $\Pr[Y^a = 1|L = l]$ by the observed risk $\Pr[Y = 1|L = l, A = a]$ in the expression above. That is, $\Pr[Y^a = 1] = \sum_l \Pr[Y = 1|L = l, A = a]\Pr[L = l]$. The left-hand side of this equality is an unobserved counterfactual risk whereas the right-hand side includes observed quantities only, which can be computed using data on L, A, and Y.

Standardized mean
$\sum_l \mathrm{E}[Y|L = l, A = a]$
$\quad \times \Pr[L = l]$

The method described above is known in epidemiology, demography, and other disciplines as *standardization*. For example, the numerator $\sum_l \Pr[Y = 1|L = l, A = 1]\Pr[L = l]$ of the causal risk ratio is the standardized risk in the treated using the population as the standard. In the presence of conditional exchangeability, this standardized risk can be interpreted as the (counterfactual) risk that would have been observed had all the individuals in the population been treated.

The standardized risks in the treated and the untreated are equal to the counterfactual risks under treatment and no treatment, respectively. Therefore, the causal risk ratio $\dfrac{\Pr[Y^{a=1} = 1]}{\Pr[Y^{a=0} = 1]}$ can be computed by standardization as $\dfrac{\sum_l \Pr[Y = 1|L = l, A = 1]\Pr[L = l]}{\sum_l \Pr[Y = 1|L = l, A = 0]\Pr[L = l]}$.

2.4 Inverse probability weighting

Figure 2.1 is an example of a finest fully randomized causally interpreted structured tree graph or FFRCISTG (Robins 1986, 1987). Did we win the prize for the worst acronym ever?

In the previous section we computed the causal risk ratio in a conditionally randomized experiment via standardization. In this section we compute this causal risk ratio via inverse probability weighting. The data in Table 2.2 can be displayed as a tree in which all 20 individuals start at the left and progress over time towards the right, as in Figure 2.1. The leftmost circle of the tree contains its first branching: 8 individuals were in non critical condition ($L = 0$) and 12 in critical condition ($L = 1$). The numbers in parentheses are the probabilities of being in noncritical, $\Pr[L = 0] = 8/20 = 0.4$, or critical, $\Pr[L = 1] = 12/20 = 0.6$, condition. Let us follow, for example, the branch $L = 0$. Of the 8 individuals in this branch, 4 were untreated ($A = 0$) and 4 were treated ($A = 1$). The conditional probability of being untreated is $\Pr[A = 0|L = 0] = 4/8 = 0.5$, as shown in parentheses. The conditional probability of being treated $\Pr[A = 1|L = 0]$ is 0.5 too. The upper right circle represents that, of the 4 individuals in the branch ($L = 0, A = 0$), 3 survived ($Y = 0$) and 1 died ($Y = 1$). That is, $\Pr[Y = 0|L = 0, A = 0] = 3/4$ and $\Pr[Y = 1|L = 0, A = 0] = 1/4$. The other branches of the tree are interpreted analogously. The circles contain the bifurcations defined by non treatment variables. We now use this tree to compute the causal risk ratio.

Figure 2.1

The denominator of the causal risk ratio, $\Pr[Y^{a=0} = 1]$, is the counterfactual risk of death had everybody in the population remained untreated. Let us calculate this risk. In Figure 2.1, 4 out of 8 individuals with $L = 0$ were untreated, and 1 of them died. How many deaths would have occurred had the 8 individuals with $L = 0$ remained untreated? Two deaths, because if 8 individuals rather than 4 individuals had remained untreated, then 2 deaths rather than 1 death would have been observed. If the number of individuals is multiplied times 2, then the number of deaths is also doubled. In Figure 2.1, 3 out of 12 individuals with $L = 1$ were untreated, and 2 of them died. How many deaths would have occurred had the 12 individuals with $L = 1$ remained

Fine Point 2.2

Risk periods. We have defined a risk as the proportion of subjects who develop the outcome of interest during a particular period. For example, the 5-day mortality risk in the treated $\Pr[Y = 1|A = 0]$ is the proportion of treated subjects who died during the first five days of follow-up. Throughout the book we often specify the period when the risk is first defined (e.g., 5 days) and, for conciseness, omit it later. That is, we may just say "the mortality risk" rather than "the five-day mortality risk."

The following example highlights the importance of specifying the risk period. Suppose a randomized experiment was conducted to quantify the causal effect of antibiotic therapy on mortality among elderly humans infected with the plague bacteria. An investigator analyzes the data and concludes that the causal risk ratio is 0.05, i.e., on average antibiotics decrease mortality by 95%. A second investigator also analyzes the data but concludes that the causal risk ratio is 1, i.e., antibiotics have a null average causal effect on mortality. Both investigators are correct. The first investigator computed the ratio of 1-year risks, whereas the second investigator computed the ratio of 100-year risks. The 100-year risk was of course 1 regardless of whether subjects received the treatment. When we say that a treatment has a causal effect on mortality, we mean that death is delayed, not prevented, by the treatment.

untreated? Eight deaths, or 2 deaths times 4, because 12 is 3×4. That is, if all $8 + 12 = 20$ individuals in the population had been untreated, then $2 + 8 = 10$ would have died. The denominator of the causal risk ratio, $\Pr[Y^{a=0} = 1]$, is $10/20 = 0.5$. The first tree in Figure 2.2 shows the population had everybody remained untreated. Of course, these calculations rely on the condition that treated individuals with $L = 0$, had they remained untreated, would have had the same probability of death as those who actually remained untreated. This condition is precisely exchangeability given $L = 0$.
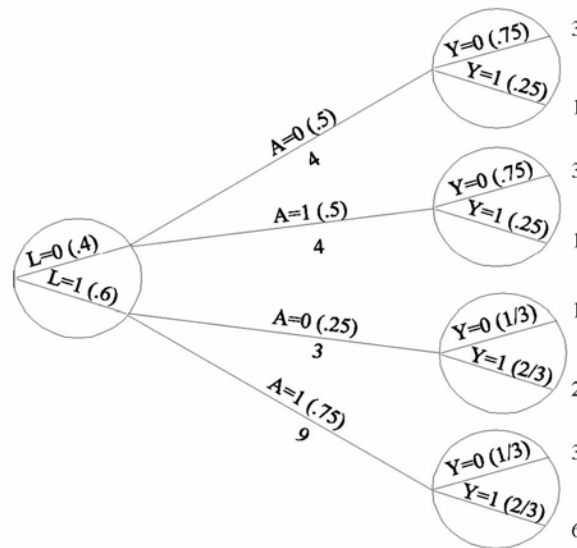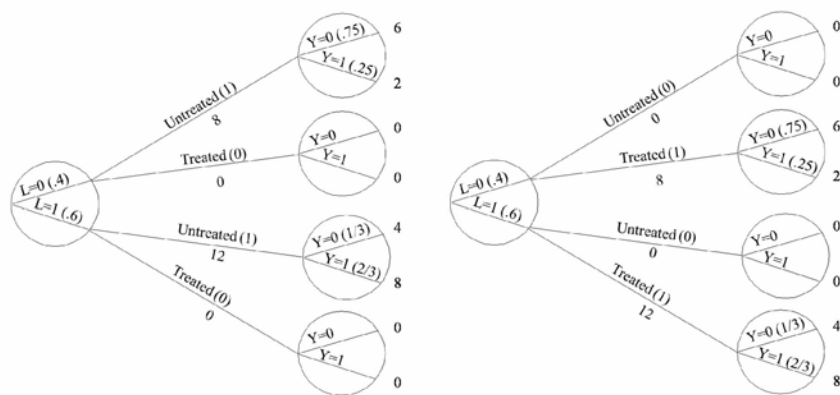
Figure 2.2

The numerator of the causal risk ratio $\Pr[Y^{a=1} = 1]$ is the counterfactual risk of death had everybody in the population been treated. Reasoning as in the previous paragraph, this risk is calculated to be also $10/20 = 0.5$, under exchangeability given $L = 1$. The second tree in Figure 2.2 shows the population had everybody been treated. Combining the results from this and the previous paragraph, the causal risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$ is equal to $0.5/0.5 = 1$. We are done.

Let us examine how this method works. The two trees in Figure 2.2 are essentially a simulation of what would have happened had all subjects in the

population been untreated and treated, respectively. These simulations are correct under conditional exchangeability. Both simulations can be pooled to create a hypothetical population in which every individual appears both as a treated and as an untreated individual. This hypothetical population, twice as large as the original population, is known as the *pseudo-population*. Figure 2.3 shows the entire pseudo-population. Under conditional exchangeability $Y^a \coprod A|L$ in the original population, the treated and the untreated are (unconditionally) exchangeable in the pseudo-population because the L is independent of A. In other words, the associational risk ratio in the pseudo-population is equal to the causal risk ratio in both the pseudo-population and the original population.

This method is known as *inverse probability (IP) weighting*. To see why, let us look at, say, the 4 untreated individuals with $L = 0$ in the population of Figure 2.1. These individuals are used to create 8 members of the pseudo-population of Figure 2.3. That is, each of them is assigned a weight of 2, which is equal to $1/0.5$. Figure 2.1 shows that 0.5 is the conditional probability of staying untreated given $L = 0$. Similarly, the 9 treated subjects with $L = 1$ in Figure 2.1 are used to create 12 members of the pseudo-population. That is, each of them is assigned a weight of $1.33 = 1/0.75$. Figure 2.1 shows that 0.75 is the conditional probability of being treated given $L = 1$. Informally, the pseudo-population is created by weighting each individual in the population by the inverse of the conditional probability of receiving the treatment level that she indeed received. These IP weights are shown in the last column of Figure 2.3.

IP weighted estimators were proposed by Horvitz and Thompson (1952) for surveys in which subjects are sampled with unequal probabilities

IP weight: $W^A = 1/f[A|L]$
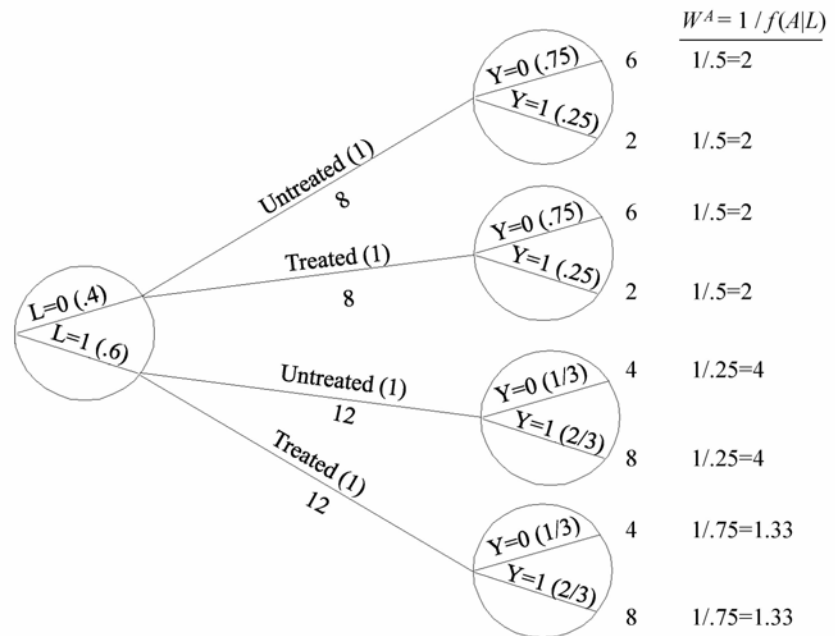

Figure 2.3

IP weighting yielded the same result as standardization—causal risk ratio equal to 1— in our example above. This is no coincidence: standardization and IP weighting are mathematically equivalent (see Technical Point 2.3). Each method uses a different set of the probabilities shown in Figure 2.1: IP weighting uses the conditional probability of treatment A given the covariate

Technical Point 2.2

Formal definition of IP weights. A subject's IP weight depends on her values of treatment A and covariate L. For example, a treated subject with $L = l$ receives the weight $1/\Pr[A = 1|L = l]$, whereas an untreated subject with $L = l'$ receives the weight $1/\Pr[A = 0|L = l']$. We can express these weights using a single expression for all subjects—regardless of their individual treatment and covariate values—by using the probability density function (PDF) of A rather than the probability of A. The conditional PDF of A given L evaluated at the values a and l is represented by $f_{A|L}[a|l]$, or simply as $f[a|l]$. For discrete variables A and L, $f[a|l]$ is the conditional probability $\Pr[A = a|L = l]$. In a conditionally randomized experiment, $f[a|l]$ is positive for all l such that $\Pr[L = l]$ is nonzero.

Since the denominator of the weight for each subject is the conditional density evaluated at the subject's own values of A and L, it can be expressed as the conditional density evaluated at the random arguments A and L (as opposed to the fixed arguments a and l), that is, as $f[A|L]$. This notation, which appeared in Figure 2.3, is used to define the IP weights $W^A = 1/f[A|L]$. It is needed to have a unified notation for the weights because $\Pr[A = A|L = L]$ is not considered proper notation.

L, standardization uses the probability of the covariate L and the conditional probability of outcome Y given A and L.

Because both standardization and IP weighting simulate what would have been observed if the variable (or variables in the vector) L had not been used to decide the probability of treatment, we often say that these methods *adjust for L*. (In a slight abuse of language we sometimes say that these methods *control for L*, but this "analytic control" is quite different from the "physical control" in a randomized experiment.) Standardization and IP weighting can be generalized to conditionally randomized studies with continuous outcomes (see Technical Point 2.3).

Why not finish this book here? We have a study design (an ideal randomized experiment) that, when combined with the appropriate analytic method (standardization or IP weighting), allows us to compute average causal effects. Unfortunately, randomized experiments are often unethical, impractical, or untimely. For example, it is questionable that an ethical committee would have approved our heart transplant study. Hearts are in short supply and society favors assigning them to subjects who are more likely to benefit from the transplant, rather than assigning them randomly among potential recipients. Also one could question the feasibility of the study even if ethical issues were ignored: double-blind assignment is impossible, individuals assigned to medical treatment may not resign themselves to forego a transplant, and there may not be compatible hearts for those assigned to transplant. Even if the study were feasible, it would still take several years to complete it, and decisions must be made in the interim. Frequently, conducting an observational study is the least bad option.

Technical Point 2.3

Equivalence of IP weighting and standardization. The standardized mean for treatment level a is defined as $\sum_l \mathrm{E}\left[Y|A=a,L=l\right]\Pr\left[L=l\right]$, and the IP weighted mean of Y for treatment level a is defined as $\mathrm{E}\left[\dfrac{I\left(A=a\right)Y}{f\left[A|L\right]}\right]$ i.e., the mean of Y, reweighted by the IP weight $W^A = 1/f\left[A|L\right]$, in subjects with treatment value $A=a$. The function $I\left(A=a\right)$ takes value 1 for subjects with $A=a$, and 0 for the others. The definitions of standardized and IP weighted means, as well as the proofs below, assume that $f\left[a|l\right]$ is positive for all l such that $\Pr\left[L=l\right]$ is nonzero. This positivity condition is guaranteed to hold in conditionally randomized experiments.

We now prove the equality of the IP weighted mean and the standardized mean. By definition of an expectation,
$$\mathrm{E}\left[\frac{I\left(A=a\right)Y}{f\left[A|L\right]}\right] = \sum_l \frac{1}{f\left[a|l\right]}\left\{\mathrm{E}\left[Y|A=a,L=l\right]f\left[a|l\right]\Pr\left[L=l\right]\right\}$$
$$= \sum_l \left\{\mathrm{E}\left[Y|A=a,L=l\right]\Pr\left[L=l\right]\right\} \text{ where in the final step we cancelled } f\left[a|l\right] \text{ from the numerator and denominator.}$$

The proof treats A and L as discrete but not necessarily dichotomous. For continuous L simply replace the sum over L with an integral.

The proof makes no reference to counterfactuals or to causality. However if we further assume conditional exchangeability then both the IP weighted and the standardized means are equal to the counterfactual mean $\mathrm{E}\left[Y^a\right]$. Here we provide two different proofs of this last statement. First, we prove equality of $\mathrm{E}\left[Y^a\right]$ and the standardized mean as in the text
$$\mathrm{E}\left[Y^a\right] = \sum_l \mathrm{E}\left[Y^a|L=l\right]\Pr\left[L=l\right] = \sum_l \mathrm{E}\left[Y^a|A=a,L=l\right]\Pr\left[L=l\right] = \sum_l \mathrm{E}\left[Y|A=a,L=l\right]\Pr\left[L=l\right]$$
where the second equality is by exchangeability and the third by consistency. Second, we prove equality of $\mathrm{E}\left[Y^a\right]$ and the IP weighted mean as follows:
$$\mathrm{E}\left[\frac{I\left(A=a\right)}{f\left[A|L\right]}Y\right] \text{ is equal to } \mathrm{E}\left[\frac{I\left(A=a\right)}{f\left[A|L\right]}Y^a\right] \text{ by consistency. Next:}$$
$$\mathrm{E}\left[\frac{I\left(A=a\right)}{f\left[A|L\right]}Y^a\right] = \mathrm{E}\left\{\mathrm{E}\left[\frac{I\left(A=a\right)}{f\left[a|L\right]}Y^a\middle|L\right]\right\} = \mathrm{E}\left\{\mathrm{E}\left[\frac{I\left(A=a\right)}{f\left[a|L\right]}\middle|L\right]\mathrm{E}\left[Y^a|L\right]\right\} \text{ (by conditional exchangeability)}$$
$$= \mathrm{E}\left\{\mathrm{E}\left[Y^a|L\right]\right\} \text{ (because } \mathrm{E}\left[\frac{I\left(A=a\right)}{f\left[a|L\right]}\middle|L\right] = 1 \text{)}$$
$$= \mathrm{E}\left[Y^a\right]$$

The extension to polytomous treatments (i.e., a can take more than two values) is straightforward. When treatment is continuous, which is unlikely in conditionally randomized experiments, effect estimates based on the IP weights $W^A = 1/f\left[A|L\right]$ have infinite variance and thus cannot be used. Chapter 12 describes generalized weights.

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 identify 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 passively observes and records the relevant data is 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 The randomized experiment paradigm

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. As discussed in Chapters 8 and 9, the previous sentence needs to be qualified in real, as opposed to ideal, randomized experiments with loss to follow-up and noncompliance with the assigned treatment.

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

Rubin (1974, 1978) extended Neyman's theory for randomized experiments to observational studies, and introduced the idea that one could view causal inference from observational studies as a missing data problem.

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? We analyze our data as if treatment had been randomly assigned conditional on the measured covariates—though we 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. An observational study can be conceptualized as a conditionally randomized experiment under the following three conditions:

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
Cyclope	1	1	1
Persephone	1	1	1
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

1. the values of treatment under comparison correspond to well-defined interventions

2. the conditional probability of receiving every value of treatment, though not decided by the investigators, depends only on the measured covariates

3. the conditional probability of receiving every value of treatment is greater than zero, i.e., positive

In this chapter we describe these three conditions in the context of observational studies. Condition 1 is necessary for the other two conditions to be defined. Condition 2 was referred to as exchangeability in 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.

When any of these conditions—and therefore the analogy between observational study and conditionally randomized experiment—does not hold, there is another possible approach to causal inference from observational data: hoping that a predictor of treatment, referred to as an *instrumental variable*, was randomly assigned conditional on the measured covariates. 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).

We discuss instrumental variable methods in Chapter REF. Until then, we will focus on causal inference approaches that rely on the ability of the observational study to emulate a conditionally randomized experiment. Therefore, for each causal question that we intend to answer using observational data, we will need to carefully describe (i) the randomized experiment that we would like to, but cannot, conduct, and (ii) how the observational study emulates that randomized experiment.

In ideal conditionally randomized experiments one can identify causal effects simply by applying IP weighting or standardization to the data. For example, in the previous chapter, we computed a causal risk ratio equal to 1 based exclusively on the data in Table 2.2, which arose from a conditionally randomized experiment. That is, in ideal randomized experiments, the data contain sufficient information to identify causal effects. In contrast, as we discuss in the following sections, the information contained in observational data

is insufficient to identify causal effects. Suppose the data in Table 3.1 arose from an observational study. (These are exactly the same data arising from the conditionally randomized study in Table 2.2.) To compute the causal risk ratio from this observational study, we need to supplement the information contained in the data with the information contained in the above conditions; only then the causal effect of treatment becomes identifiable. It follows that causal effects can be identified from observational studies by using IP weighting or standardization when the three above conditions—well-defined interventions, exchangeability, and positivity—hold. We therefore refer to them as *identifiability* conditions (see Fine Point 3.4). Causal inference from observational data requires two sources of information: data and identifiability assumptions. In Chapter REF we discuss identifiability assumptions other than the three discussed here.

The combination of exchangeability and positivity is also referred as weak ignorability, and the combination of full exchangeability (see Technical Point 2.1) and positivity as strong ignorability (Rosenbaum and Rubin 1983).

3.2 Exchangeability

We have already said much about exchangeability $Y^a \coprod A$. 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 cannot 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).

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.

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 $Y^a \coprod A|L$ holds in conditionally randomized experiments because, within levels of L, all other predictors of the outcome are equally distributed between the treated and the 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 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

Fine Point 3.1

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 $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$. Both the causal risk ratio and the causal risk difference are examples of effect measures that 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 contrast between observed and counterfactual risks 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 general, compute the effect measures via standardization if the identifiability conditions hold.) 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* 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.

Besides the excess fraction, other definitions of attributable fraction have been proposed. For example, the *etiologic fraction* is defined as the proportion of cases whose disease originated from a biologic (or other) process in which treatment had an effect. This is a mechanistic definition of attributable fraction that does not rely on the concept of excess cases and thus can only be computed in randomized experiments under strong assumptions. The etiologic fraction, also known as "probability of causation," has legal relevance because it is used to award compensation in lawsuits. There are yet other definitions of attributable fraction. Greenland and Robins (1988) and Robins and Greenland (1989) discuss the definition, interpretation, estimability, and estimation of the various attributable fractions.

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 (non blinded) 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 $Y^a \coprod A|L$ holds and standardization or IP weighting can be used to identify the causal effect.

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. For example, suppose the investigators of our observational study strongly believe that the treated and the untreated are exchangeable within levels of L. Their reasoning goes as follows: "Heart transplants are assigned to individuals with low probability of rejecting the transplant, that is, a heart with certain human leukocyte

antigen (HLA) genes will be assigned to a subject who happen to have compatible genes. Because HLA genes are not predictors of mortality, it turns out that treatment assignment is essentially random within levels of L." Thus our investigators are willing to work under the *assumption* that conditional exchangeability $Y^a \coprod A|L$ holds.

The key word is "assumption." No matter how convincing the investigators' story may be, 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. If two study subjects with $L = 1$ have similar HLA genes, but 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 predictor of the outcome, differs between the treated (lower proportion of smokers $U = 1$) and the untreated (higher proportion of smokers $U = 1$) with $L = 1$, 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.

We use U to denote unmeasured variables. Because unmeasured variables U 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.

Thus exchangeability $Y^a \coprod A|L$ cannot be generally expected to hold in observational studies. Specifically, conditional exchangeability $Y^a \coprod A|L$ 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 $Y^a \coprod A|L$ held, the investigators cannot empirically verify that is 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? Thus when we analyze an observational study under the assumption of conditional exchangeability, we must hope that the assumption is at least approximately true.

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.

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 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 subjects to either $A = 1$ or $A = 0$. That would be silly. With all the subjects receiving the same treatment level, computing the

average causal effect would be impossible. Instead we must assign treatment so that, with near certainty, some subjects 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, sometimes referred to as the *experimental treatment assumption*, that is required for causal inference.

<div style="float:left; width:30%;">

Positivity can be empirically violated in small samples (see Chapter 12) but, for now, we will only consider causal inference with very large sample sizes.

</div>

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. 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 individuals 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 $L = 0$ (our inference would be restricted to the group $L = 1$ anyway). Thus there is positivity if $\Pr[A = a|L = l] > 0$ for all l with $\Pr[L = l] \neq 0$ in the population of interest.

<div style="float:left; width:30%;">

Positivity: $\Pr[A = a|L = l] > 0$ for all l with $\Pr[L = l] \neq 0$

</div>

In our example of Table 3.1, we say that positivity holds 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 $L = 1$). When exchangeability is achieved conditional on some variables L, then it is sufficient to have positivity with respect to just those variables. For example, suppose the variable "having blue eyes" is not an independent predictor of the outcome given L and A. Suppose further that positivity does not hold for "having blue eyes" because in the group in critical condition $L = 1$, blue-eyed subjects were personally selected by the investigators and assigned to heart transplant $A = 1$ (and all others to medical treatment $A = 0$). Nonetheless the standardized risk (standardized with respect to L) and the IP weighted risk are still equal to the counterfactual risk.

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).

Our discussion of standardization and IP weighting in the previous chapter was explicit about the exchangeability condition, but only implicitly assumed the positivity condition (except 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 subjects ($A = 0$) with $L = 1$, the data would contain no information to simulate what would have happened had all treated subjects been untreated because there would be no untreated subjects with $L = 1$ that could be considered exchangeable with the treated subjects with $L = 1$. See Technical Point 3.1 for details.
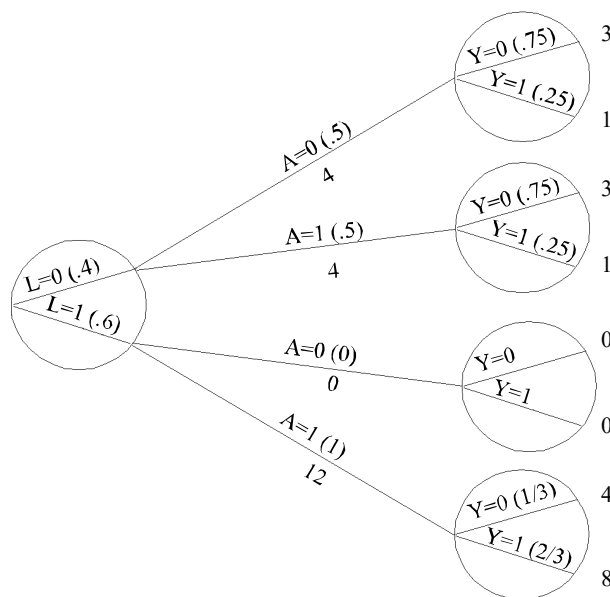
Figure 3.1

3.4 Well-defined interventions

Consider again a randomized experiment to compute the average effect of heart transplant A on 5-year mortality Y. Prior to 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. That is, the investigators specified that individuals assigned to $A = 1$ were to receive a particular type of pre-operative procedures, anesthesia, surgical technique, post-operative intensive care, and immunosuppressive treatment. Had the protocol not specified these details, it is possible that each doctor had conducted the heart transplant in a different way, perhaps using her preferred surgical technique or immunosuppressive therapy. That is, different versions of the treatment "heart transplant" might have been applied to each patient in the study (Fine Point 1.2 introduced the concept of multiple versions of treatment).

The presence of multiple versions of treatment is problematic when the causal effect varies across versions, i.e., when the versions of treatment are relevant for the outcome. Then the magnitude of the average causal effect depends on the proportion of individuals who received each version. For example, the average causal effect of "heart transplant" in a study in which most doctors used conventional immunosuppressive therapy may differ from that in a study in which most doctors used a novel immunosuppressive therapy. In this setting, the treatment "heart transplant" is not a unique treatment A but rather a collection R of different versions of treatment. We use $A(r)$, $A'(r)$, ... to refer to the versions of treatment $R = r$.

In the presence of multiple versions of treatment, the interventions of interest (e.g., heart transplant, medical therapy) are not well defined. And if the interventions are not well defined, the average causal effect is not well defined either. What do we mean by "the causal effect of heart transplant" if heart

Technical Point 3.1

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

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

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 subjects that received the treatment level a.

Similarly, the IP weighted mean for treatment level a, $\mathrm{E}\left[\dfrac{I\left(A=a\right)Y}{f\left[a|L\right]}\right]$, is only well defined under positivity. When positivity does not hold, the undefined ratio $\frac{0}{0}$ occurs in computing the expectation. Define the "apparent" IP weighted mean for treatment level a to be $\mathrm{E}\left[\dfrac{I\left(A=a\right)Y}{f\left[A|L\right]}\right]$. This mean is *always* well defined since its denominator $f\left[A|L\right]$ can never be zero. When positivity holds, the "apparent" and true IP weighted means for treatment level a are equal to one another (and to the standardized mean) and thus all quantities are well defined. When positivity fails to hold, the "apparent" IP weighted mean for treatment level a is equal to $\Pr\left[L \in Q(a)\right] \sum_l \mathrm{E}\left[Y|A=a, L=l, L \in Q(a)\right] \Pr\left[L=l|L \in Q(a)\right]$ where $Q(a) = \{l; \Pr\left(A=a|L=l\right) > 0\}$ is the set of values l for which $A=a$ may be observed with positive probability. Under exchangeability, the "apparent" IP weighted mean equals $\mathrm{E}\left[Y^a|L \in Q(a)\right] \Pr\left[L \in Q(a)\right]$.

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 $\mathrm{E}\left[\dfrac{I\left(A=1\right)Y}{f\left[A|L\right]}\right] - \mathrm{E}\left[\dfrac{I\left(A=0\right)Y}{f\left[A|L\right]}\right]$ 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.

Treatment-variation irrelevance (VanderWeele 2009) holds if, for any two versions $A(r)$ and $A'(r)$ of treatment $R = r$, $Y_i^{r,a(r)} = Y_i^{r,a'(r)}$ for all i and r. $Y_i^{r,a(r)}$ is subject i's counterfactual outcome under version $A(r) = a(r)$ of treatment $R = r$.

transplant R has multiple versions $A(r)$, $A'(r)$, ... each of them with different effects on the outcome? To avoid having to answer this question, we have so far assumed that the treatment of interest either does not have multiple versions, or has multiple versions with identical effects, i.e., the versions are not relevant for the outcome. For example, when discussing the causal effect of heart transplant ($A = 1$), we considered "heart transplant"—and also "medical treatment"—as an intervention that is well defined and that does not vary across individuals. That is, we assumed that all individuals receiving a heart transplant were receiving the same version of treatment $A = 1$, and similarly for $A = 0$.

This assumption of treatment-variation irrelevance may be reasonable in ideal randomized studies with protocols that clearly specify the interventions under study. On the other hand, in observational studies, the investigators have no control over the versions of the treatment that were applied to the individuals in the study. As a result, it is possible that multiple versions of treatment were used. When using observational data for causal inference, we need to carefully define the versions of interest and then collect sufficient data to characterize them. If interested in the causal effect of heart transplant, we need to decide which version(s) of heart transplant we are interested in (i.e.,

we need to define $A = 1$ unambiguously), and record enough information on the type of heart transplant so that only individuals receiving the version(s) of interest are considered as treated individuals ($A = 1$) in the analysis. If interested in the causal effect of exercise, we need to specify the version(s) of exercise we are interested in (e.g., duration, intensity, frequency, type of physical activity) and then include in the group $A = 1$ only individuals whose observed data are consistent with the version(s) of interest.

Therefore, in theory, the problem of multiple versions of treatment can be solved by restriction. If interested in the effect of exercise $A = 1$ defined as running exactly 30 minutes per day at moderate intensity on a flat terrain, we would only include individuals who run exactly 30 minutes per day at moderate intensity on a flat terrain in the group $A = 1$. But restriction may be impractical: if we define the intervention of interest in much detail, perhaps no individual's data are consistent with that version of treatment. The more precise we get the higher the risk of nonpositivity in some subsets of the study population. In practice, we need a compromise. For example, we may consider as $A = 1$ running, or playing soccer, between 25 and 35 minutes per day. That is, we effectively assume that several versions of treatment (e.g., running 25 minutes, running 26 minutes, running 27 minutes...) are not relevant for the outcome, and pool them together.

There is another reason why restriction to the version of interest may be problematic: we often have no data on the version of treatment. In fact, we may not be able to even enumerate the versions. Suppose we conduct an observational study to estimate the average causal effect of "obesity" R on the risk of mortality Y. All individuals aged 40 in the country are followed until their 50th birthday. At age 40, some subjects happen to be obese (body mass index$\geqslant 30$ or $R = 1$) and others happen to be nonobese (body mass index< 30 or $R = 0$). It turns out that obese subjects ($R = 1$) have a greater 10-year risk of death ($Y = 1$) than nonobese subjects ($R = 0$), i.e., the associational risk ratio $\Pr[Y = 1|R = 1]/\Pr[Y = 1|R = 0]$ is greater than 1. This finding establishes that obesity is associated with mortality or, equivalently, that obesity is a predictor of mortality. This finding does not establish that obesity has a causal effect on the 10-year mortality risk. To do so, we would need to compare the risks if all subjects had been obese $\Pr[Y^{r=1} = 1]$ and nonobese $\Pr[Y^{r=0} = 1]$ at age 40. But what exactly is meant by "the risks if all subjects had been obese and nonobese"? The answer is not straightforward because it is unclear what the treatment R means, which implies that the counterfactual outcomes Y^r are ill-defined.

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

To see this, take Kronos, an obese individual ($R = 1$) who died ($Y = 1$). There are many different ways in which an obese individual could have been nonobese. That is, there are multiple versions $A(r = 0)$ of the treatment $R = 0$. Here are some of them: more exercise, less food intake, more cigarette smoking, genetic modification, bariatric surgery, any combination of the above. The counterfactual outcome $Y^{r=0}$ if Kronos had been nonobese rather than obese is not well defined because its value depends on the particular version of $A(r = 0)$ of treatment $R = 0$ that we consider. A nonobese Kronos might have died if he had been nonobese through a lifetime of exercise (a bicycle accident), cigarette smoking (lung cancer), or bariatric surgery (adverse reaction to anesthesia), and might have survived if he had been nonobese through a better diet (fewer calories from devouring his children) or more favorable genes (less fat tissue).

The definition of the counterfactual outcome $Y^{r=1}$ is also problematic. Suppose Kronos was obese because his genes predisposed him to large amounts of fat tissue in both his waist and his coronary arteries. He had a fatal myocardial

infarction at age 49 despite not smoking, exercising moderately, and keeping a healthy diet. However, if he had been obese not because of his genes but because of lack of exercise and too many calories in the diet, then he would not have died by age 50. The outcome of an obese Kronos's might have been 0 if he had been obese through mechanisms $A(r = 1)$ other than the ones that actually made him obese, even though he was actually obese $R = 1$ and his observed outcome Y was 1.

Obesity may lead to less vague causal questions in other settings. Consider the effect of obesity on job discrimination as defined by the proportion of job applicants called for a personal interview after the employer reviews the applicant's resume and photograph. Because the treatment here is really "obesity as perceived by the prospective employer," the mechanisms that led to obesity are irrelevant.

The question "Does obesity have a causal effect on mortality?" is quite vague because the answer depends on *how* one intervenes on obesity. This problem arises for treatments R with multiple versions $A(r)$ when the versions are relevant for the outcome of interest. For example, if the data in Table 3.1 came from observational study, we would need to assume that the many versions of treatment $A = 0$ (e.g., the heart donor happened to survive the chariot crash that would have led to his death, Zeus killed the surgeon before the surgery appointment) are not relevant for the outcome, i.e., that Zeus's counterfactual outcome is the same under any of the versions of treatment and thus can be unambiguously represented by $Y^{a=0}$.

Because treatment-variation irrelevance cannot be taken for granted in observational studies, the interpretation of the causal effect is not always straightforward. At the very least, investigators need to characterize the versions of treatment that operate in the population. Such characterization is simple in experiments (i.e., whatever intervention investigators use to assign treatment), and relatively unambiguous in some observational studies (e.g., those studying the effects of medical treatments). For example, in an observational study of aspirin and mortality, one can imagine how to hypothetically manipulate an individual treatment's level by simply withholding or administering aspirin at the start of the study. On the other hand, the characterization of the versions of "treatments" that are complex biological (e.g., body weight, LDL-cholesterol, CD4 cell count, or C-reactive protein) or social (e.g., socioeconomic status) processes is often vague. This inherent vagueness has led some authors to propose that only the causal effects of treatments that can be hypothetically manipulated should ever be considered. See Fine Point 3.2 for additional discussion on the vagueness of hypothetical interventions.

Holland (1986): "no causation without manipulation"

This is an example of a random intervention or regime. Hernán and Taubman (2008) discussed this implicit intervention in observational studies of obesity and health outcomes; they refer to settings with ill-defined interventions as settings in which the consistency condition does not hold. Hernán and Vander-Weele (2011), and this book, propose a sharper distinction of the conditions of well-defined interventions and consistency.

There is one trick to address the vagueness of causal effects when the versions of treatment are unknown. Consider the following hypothetical intervention: 'assign everybody to being nonobese by changing the determinants of body weight to reflect the distribution of those determinants in those who already have nonobese weight in the study population.' This hypothetical intervention would randomly assign a version of treatment to each individual in the study population so that the resulting distribution of versions of treatment exactly matches the distribution of versions of treatment in the study population. We can propose an analogous intervention 'assign everybody to being obese.' This trick is implicitly used in the analysis of many observational studies that compare the risks $\Pr[Y = 1 | R = 1]$ and $\Pr[Y = 1 | R = 0]$ (often conditional on other variables) to endow the contrast with a causal interpretation. The problem with this trick is, of course, that the proposed random interventions do not match any realistic interventions we are interested in. Learning that intervening on 'the determinants of body weight to reflect the distribution of those determinants in those with nonobese weight' decreases mortality by, say, 30% does not imply that any real world intervention on obesity (e.g., by modifying caloric intake or exercise levels) will decrease mortality by 30% too. In fact, if intervening on 'determinants of body weight in the population' requires intervening on genetic factors, then a 30% reduction in mortality may

Fine Point 3.2

Refining causal questions. The well-defined versus ill-defined dichotomy for causal questions is a pedagogic simplification because no intervention is perfectly specified. The question "Does obesity have a causal effect on mortality?" is more ill-defined than the question "Does low-fat diet have a causal effect on mortality?" or than the question "Does exercise have a causal effect on mortality?" However, the latter question is more ill-defined than "Does 1 additional hour of daily strenuous exercise have a causal effect on mortality?" And even this question is not perfectly defined because the effect of the intervention will depend on how that hour would otherwise be spent. Reducing time spent laughing with your friends, playing with your children, or rehearsing with your band may have a different effect on mortality than reducing time eating, watching television, or studying.

No matter how much refining of the causal question, all causal effects from observational data are inherently vague. But there is a question of degree of vagueness. The vagueness inherent in increased "exercise" is less serious than that in "obesity" and can be further reduced by a more detailed specification of the intervention on exercise. That some interventions sound technically unfeasible or plainly crazy simply indicates that the formulation of causal questions is not straightforward. An explicit (counterfactual) approach to causal inference highlights the imprecision of ambiguous causal questions, and the need for a common understanding of the interventions involved (Robins and Greenland, 2000).

be unattainable by interventions that can actually be implemented in the real world.

3.5 Well-defined interventions are a pre-requisite for causal inference

What's so wrong with estimating the causal effects of ill-defined interventions? We may not precisely know which particular causal effect is being estimated in an observational study, but is that really so important if indeed some causal effect exists? A strong association between obesity and mortality may imply that there exists some intervention on body weight that reduces mortality. As described in the previous section, complex interventions that vary across subjects are implicit in the analysis of many observational studies—for example, those that attempt to estimate "the" causal effect of obesity. These implicit interventions are too difficult to implement and may not be meaningful for public health purposes. Yet one could argue that there is some value in learning that many deaths could have been prevented if all obese people had been forced to be of normal weight, even if the intervention required for achieving that transformation is unspecified and likely unrealistic. This is an appealing, but risky, argument. We now discuss why accepting ill-defined interventions prevents a proper consideration of exchangeability and positivity in observational studies.

Investigators use their subject-matter knowledge to measure the covariates L that will be adjusted for in the analysis. However, even in the absence of multiple versions of treatment, investigators cannot be certain that their efforts to measure covariates have resulted in approximate conditional exchangeability. This uncertainty, which is a fundamental shortcoming of causal inference from observational data, is greatly exacerbated when the interventions of interest are not well defined because of unknown versions of treatment. If we renounce to characterize the intervention corresponding to the causal effect of obesity R, how can we identify and measure the covariates L that make obese and nonobese subjects conditionally exchangeable, i.e., covariates L that are determinants of the versions $A(r)$ of treatment (obesity) and also risk factors for

Fine Point 3.3

Possible worlds. Some philosophers of science define causal effects using the concept of "possible worlds." The actual world is the way things actually are. A possible world is a way things might be. Imagine a possible world a where everybody receives treatment value a, and a possible world a' where everybody receives treatment value a'. The mean of the outcome is $E[Y^a]$ in the first possible world and $E[Y^{a'}]$ in the second one. These philosophers say that there is an average causal effect if $E[Y^a] \neq E[Y^{a'}]$ and the worlds a and a' are the two worlds closest to the actual world where all subjects receive treatment value a and a', respectively.

   We introduced an individual's counterfactual outcome Y^a as her outcome under a well specified intervention that assigned treatment value a to her. These philosophers prefer to think of the counterfactual Y^a as the outcome in the possible world that is closest to our world and where the subject was treated with a. Both definitions are equivalent when the only difference between the closest possible world and the actual world is that the intervention of interest took place. The possible worlds formulation of counterfactuals replaces the sometimes difficult problem of specifying the intervention of interest by the equally difficult problem of describing the closest possible world that is minimally different from the actual world. Stalnaker (1968) and Lewis (1973) proposed counterfactual theories based on possible worlds.

the outcome (mortality)? When trying to estimate the effect of an unspecified intervention, the concept of conditional exchangeability remains undefined.

   The acceptance of unspecified interventions also affects positivity. Suppose we decide to compute the effect of obesity on mortality by adjusting for some measured covariates L that include some genetic factors. It is possible that some genetic traits are so strongly associated to body weight that no subject possessing them will be obese; that is, positivity does not hold. If enough biologic knowledge is available, one could preserve positivity by restricting the analysis to the strata of L in which the population contains both obese and nonobese subjects. The price to pay for this strategy is potential lack of generalizability of the estimated effect (see Chapter 4), as these strata may no longer be representative of the original population.

   Positivity violations point to another potential problem: unspecified interventions may be unreasonable. The apparently straightforward comparison of obese and nonobese subjects in observational studies masks the true complexity of the interventions 'make everybody in the population nonobese' and 'make everybody in the population obese.' Had these interventions been made explicit, investigators would have realized that the interventions were too extreme to be relevant for public health because drastic changes in body weight (say, from body mass index of 30 to 25) in a short period are unachievable. Further, these drastic changes are unlikely to be observed in the study data, and thus any estimate of the effect of that intervention will rely heavily on modeling assumptions (see Part II). A more reasonable, even if still ill-characterized, intervention may be to reduce body mass index by 5% over a two-year period. In summary, violations of positivity are more likely to occur when estimating the effect of extreme interventions, and extreme interventions are more likely to go unrecognized when they are not explicitly specified.

   The problems generated by unspecified interventions cannot be dealt with by applying sophisticated statistical methods. All analytic methods for causal inference from observational data described in this book yield effect estimates that are only as well defined as the interventions that are being compared. Although the exchangeability condition can be replaced by other conditions (see Chapter 16) and the positivity condition can be waived if one is willing to

Technical Point 3.2

Consistency and multiple versions of treatment. The consistency condition is necessary, together with conditional exchangeability and positivity, to prove that the standardized mean and the IP weighted mean equal the counterfactual mean (see Technical Point 2.3). In Chapter 1, we defined consistency as follows: For all individuals i in the study, if $A_i = a$ then $Y_i^a = Y_i$. That is, consistency simply means that the outcome for every treated individual equals his outcome if he had received treatment, and the outcome for every untreated individual equals his outcome if he had remained untreated. This statement seems obviously true when treatment A has only one version, i.e., when it is a simple treatment. Let us now consider a treatment R with multiple versions, i.e., a compound treatment.

For a compound treatment R, the consistency assumption effectively reduces to the consistency assumption for a simple treatment when the version of treatment is not relevant for outcome Y. When the version of treatment is actually relevant for outcome Y, we may still articulate a consistency assumption as follows. For individuals with $R_i = r$ we let $A_i(r)$ denote the version of treatment $R_i = r$ actually received by individual i; for individuals with $R_i \neq r$ we define $A_i(r) = 0$ so that $A_i(r) \in \{0\} \cup \mathcal{A}(r)$. The consistency condition then requires for all i,

$$Y_i = Y_i^{r,a(r)} \text{ when } R_i = r \text{ and } A_i(r) = a(r).$$

That is, the outcome for every individual who received a particular version of treatment $R = r$ equals his outcome if he had received that particular version of treatment. This statement is true by definition of version of treatment if we in fact define the counterfactual $Y_i^{r,a(r)}$ for individual i with $R_i = r$ and $A_i(r) = a(r)$ as individual i's outcome that he actually had under actual treatment r and actual version $a(r)$. Hernán and VanderWeele (2011) discussed consistency for compound treatments with multiple versions.

Thus "consistency" and "multiple versions of treatment" are closely related concepts, and they have been often discussed together. Interestingly, the consistency condition is a red herring in these discussions because consistency can always be articulated in such a way that the condition is guaranteed to hold.

make untestable modeling assumptions to extrapolate to conditions that are not observed in the population, the requirement of well defined interventions is so fundamental that it cannot be waived aside without simultaneously negating the possibility of describing the causal effect that is being estimated.

3.6 Causation or prediction?

What ideal randomized experiment are you trying to emulate? This is a key question for causal inference from observational data. We may not be able to emulate a randomized experiment using observational data because of lack of conditional exchangeability or positivity for well-defined interventions, or because of ill-defined interventions

Consider again an observational study to estimate "the causal effect of obesity on mortality." Because there are many procedures to reduce body weight, one could try to emulate many different randomized experiments. Some of those experiments (e.g., chopping off an arm, starvation, smoking) can be ruled out because they clearly do not correspond to any interesting intervention from a public health standpoint. Interestingly, the experiment implicitly emulated by many observational studies—which simply compare the conditional risk of death in obese versus nonobese individuals—also lacks any public health interest because the corresponding unspecified intervention is a complex and probably unfeasible random regime.

Also, we have argued that unspecified interventions make it impossible to define whether exchangeability is achieved conditional on whatever variables L are measured. It is even possible that the data analyst adjusts for some variables L that are actually versions $A(r)$ of treatment R. These versions of treatment get then effectively excluded from the implicit intervention. For example, the better we are able to adjust for known, measurable factors that determine both body weight and mortality, e.g., diet, exercise, cigarette smoking, the more we are isolating an implied intervention that changes the remaining determinants of body weight, e.g., genes, asymptomatic illness. If the goal was informing public health policy, it would then seem that we have strayed from the most interesting questions. If we also try to adjust for genes and physiology, we may be delving so far into biological (or social) processes that we may encounter positivity violations.

Is everything lost when the observational data cannot be used to emulate an interesting randomized experiment? Not really. Observational data may still be quite useful by focusing on *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 subjects at high risk of mortality. Note, however, that by simply saying that obesity predicts—is associated with—mortality, we remain agnostic about the causal effects of obesity on mortality: obesity might predict mortality 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?), but not necessarily an appropriate justification to recommend a weight loss intervention targeted to the entire population.

By retreating into prediction from observational data, we avoid tackling questions that cannot be logically asked in randomized experiments. On the other hand, when causal inference is the ultimate goal, prediction may be unsatisfying.

Fine Point 3.4

Identifiability of causal effects. We say that an average causal effect is (non parametrically) identifiable when 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 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 $Y^a \coprod A|L$ 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 $Y^a \coprod A|L$. 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 $Y^a \coprod A|L$.

In the absence of selection bias (see Chapter 8), the assumption of conditional exchangeability given L is often known as the assumption of no unmeasured confounding given L (see Chapter 7). We now relate the concepts of identifiability and confounding in a setting in which the two other identifying assumptions—positivity and well-defined interventions—hold.

In a marginally randomized experiment, exchangeability $Y^a \coprod A$ ensures that effect measures can be computed when complete data on treatment A and outcome Y are available. For example, the causal risk ratio equals the associational risk ratio. There is no confounding or, equivalently, the causal effect is identifiable given data on A and Y.

In an ideal conditionally randomized experiment, conditional exchangeability $Y^a \coprod A|L$ ensures that effect measures can be computed when complete data on treatment A, outcome Y, and variable L are available. For example, the causal risk ratio equals the ratio of standardized risks. There is no unmeasured confounding given the measured variable L or, equivalently, the causal effect is identifiable given data on L, A and Y.

In an observational study, there is no guarantee that the treated and the untreated are conditionally exchangeable given L only. Thus the effect measures may not be computed even if complete data on L, A, and Y are available because of unmeasured confounding (i.e., other variables besides L must be measured and conditioned on to achieve exchangeability). Equivalently, the causal effect is not identifiable given the measured data.

This chapter discussed the assumptions required for nonparametric identification of average causal effects, that is, for identification that does not require any modeling assumptions when the size of the study population is quasi-infinite. Part II will discuss the use of models to estimate average causal effects.

Chapter 4
EFFECT MODIFICATION

So far we have focused on the average causal effect in an entire population of interest. However, many causal questions are about subsets of the population. Consider again the causal question "does one's looking up at the sky make other pedestrians look up too?" You might be interested in computing the average causal effect of treatment—your looking up to the sky— in city dwellers and visitors separately, rather than the average effect in the entire population of pedestrians.

The decision whether to compute average effects in the entire population or in a subset depends on the inferential goals. In some cases, you may not care about the variations of the effect across different groups of subjects. For example, suppose you are a policy maker considering the possibility of implementing a nationwide water fluoridation program. Because this public health intervention will reach all households in the population, your primary interest is in the average causal effect in the entire population, rather than in particular subsets. You will be interested in characterizing how the causal effect varies across subsets of the population when the intervention can be targeted to different subsets, or when the findings of the study need to be applied to other populations.

This chapter emphasizes that there is not such a thing as *the* causal effect of treatment. Rather, the causal effect depends on the characteristics of the particular population under study.

4.1 Definition of effect modification

Table 4.1

	M	Y^0	Y^1
Rheia	1	0	1
Demeter	1	0	0
Hestia	1	0	0
Hera	1	0	0
Artemis	1	1	1
Leto	1	0	1
Athena	1	1	1
Aphrodite	1	0	1
Persephone	1	1	1
Hebe	1	1	0
Kronos	0	1	0
Hades	0	0	0
Poseidon	0	1	0
Zeus	0	0	1
Apollo	0	1	0
Ares	0	1	1
Hephaestus	0	0	1
Cyclope	0	0	1
Hermes	0	1	0
Dionysus	0	1	0

We started this book by computing the average causal effect of heart transplant A on death Y in a population of 20 members of Zeus's extended family. We used the data in Table 1.1, whose columns show the individual values of the (generally unobserved) counterfactual outcomes $Y^{a=0}$ and $Y^{a=1}$. After examining the data in Table 1.1, we concluded that the average causal effect was null. Half of the members of the population would have died if everybody had received a heart transplant, $\Pr[Y^{a=1} = 1] = 10/20 = 0.5$, and half of the members of the population would have died if nobody had received a heart transplant, $\Pr[Y^{a=0} = 1] = 10/20 = 0.5$. The causal risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$ was $0.5/0.5 = 1$ and the causal risk difference $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$ was $0.5 - 0.5 = 0$.

We now consider two new causal questions: What is the average causal effect of A on Y in women? And in men? To answer these questions we will use Table 4.1, which contains the same information as Table 1.1 plus an additional column with an indicator M for sex. For convenience, we have rearranged the table so that women ($M = 1$) occupy the first 10 rows, and men ($M = 0$) the last 10 rows.

Let us first compute the average causal effect in women. To do so, we need to restrict the analysis to the first 10 rows of the table with $M = 1$. In this subset of the population, the risk of death under treatment is $\Pr[Y^{a=1} = 1|M = 1] = 6/10 = 0.6$ and the risk of death under no treatment is $\Pr[Y^{a=0} = 1|M = 1] = 4/10 = 0.4$. The causal risk ratio is $0.6/0.4 = 1.5$ and the causal risk difference is $0.6 - 0.4 = 0.2$. That is, on average, heart transplant A *increases* the risk of death Y in women.

Let us next compute the average causal effect in men. To do so, we need to restrict the analysis to the last 10 rows of the table with $M = 0$. In this subset of the population, the risk of death under treatment is $\Pr[Y^{a=1} = 1|M = 0] = 4/10 = 0.4$ and the risk of death under no treatment is $\Pr[Y^{a=0} = 1|M = 0] = 6/10 = 0.6$. The causal risk ratio is $0.4/0.6 = 2/3$ and the causal risk difference is $0.4 - 0.6 = -0.2$. That is, on average, heart transplant A *decreases* the risk of death Y in men.

Our example shows that a null average causal effect in the population does not imply a null average causal effect in a particular subset of the population. In Table 4.1, the *null hypothesis of no average causal effect* is true for the entire population, but not for men or women when taken separately. It just happens that the average causal effects in men and in women are of equal magnitude but in opposite direction. Because the proportion of each sex is 50%, both effects cancel out exactly when considering the entire population. Although exact cancellation of effects is probably rare, heterogeneity of the individual causal effects of treatment is often expected because of variations in individual susceptibilities to treatment. An exception occurs when the *sharp null hypothesis of no causal effect* is true. Then no heterogeneity of effects exists because the effect is null for every individual and thus the average causal effect in any subset of the population is also null.

We are now ready to provide a definition of effect modifier. We say that M is a modifier of the effect of A on Y when the average causal effect of A on Y varies across levels of M. Since the average causal effect can be measured using different effect measures (e.g., risk difference, risk ratio), the presence of effect modification depends on the effect measure being used. For example, sex M is an effect modifier of the effect of heart transplant A on mortality Y on the *additive* scale because the causal risk difference varies across levels of M. Sex M is also an effect modifier of the effect of heart transplant A on mortality Y on the multiplicative scale because the causal risk ratio varies across levels of M. Note that we only consider variables M that are not affected by treatment A as effect modifiers. Variables affected by treatment may be *mediators* of the effect of treatment, as described in Chapter REF.

In Table 4.1 the causal risk ratio is greater than 1 in women ($M = 1$) and less than 1 in men ($M = 0$). Similarly, the causal risk difference is greater than 0 in women ($M = 1$) and less than 0 in men ($M = 0$). That is, we say that there is *qualitative effect modification* because the average causal effects in the subsets $M = 1$ and $M = 0$ are in the opposite direction. In the presence of qualitative effect modification, additive effect modification implies multiplicative effect modification, and vice versa. In the absence of qualitative effect modification, however, one can find effect modification on one scale (e.g., multiplicative) but not on the other (e.g., additive). To illustrate this point, suppose that, in a second study, we computed the quantities shown to the left of this line. In this second study, there is no additive effect modification by M because the causal risk difference among individuals with $M = 1$ equals that among individuals with $M = 0$, that is, $0.9 - 0.8 = 0.1 = 0.2 - 0.1$. However, in this second study there is multiplicative effect modification by M because the causal risk ratio among individuals with $M = 1$ differs from that among individuals with $M = 0$, that is, $0.9/0.8 = 1.1 \neq 0.2/0.1 = 2$. Since one cannot generally state that there is, or there is not, effect modification without referring to the effect measure being used (e.g., risk difference, risk ratio), some authors use the term *effect-measure modification*, rather than effect modification, to emphasize the dependence of the concept on the choice of effect measure.

See Section 6.5 for a structural classification of effect modifiers.

Additive effect modification:
$E[Y^{a=1} - Y^{a=0}|M = 1] \neq E[Y^{a=1} - Y^{a=0}|M = 0]$

Multiplicative effect modification:
$\frac{E[Y^{a=1}|M=1]}{E[Y^{a=0}|M=1]} \neq \frac{E[Y^{a=1}|M=0]}{E[Y^{a=0}|M=0]}$

Note that we do not consider effect modification on the odds ratio scale because the odds ratio is rarely, if ever, the parameter of interest for causal inference.

Multiplicative, but not additive, effect modification by M:
$\Pr[Y^{a=0} = 1|M = 1] = 0.8$
$\Pr[Y^{a=1} = 1|M = 1] = 0.9$
$\Pr[Y^{a=0} = 1|M = 0] = 0.1$
$\Pr[Y^{a=1} = 1|M = 0] = 0.2$

4.2 Stratification to identify effect modification

Stratification: the causal effect of A on Y is computed in each stratum of M. For dichotomous M, the stratified causal risk differences are:
$\Pr[Y^{a=1} = 1|M = 1] -$
$\Pr[Y^{a=0} = 1|M = 1]$
and
$\Pr[Y^{a=1} = 1|M = 0] -$
$\Pr[Y^{a=0} = 1|M = 0]$

Table 4.2
Stratum $M = 0$

	L	A	Y
Cybele	0	0	0
Saturn	0	0	1
Ceres	0	0	0
Pluto	0	0	0
Vesta	0	1	0
Neptune	0	1	0
Juno	0	1	1
Jupiter	0	1	1
Diana	1	0	0
Phoebus	1	0	1
Latona	1	0	0
Mars	1	1	1
Minerva	1	1	1
Vulcan	1	1	1
Venus	1	1	1
Seneca	1	1	1
Proserpina	1	1	1
Mercury	1	1	0
Juventas	1	1	0
Bacchus	1	1	0

A stratified analysis is the natural way to identify effect modification. To determine whether M modifies the causal effect of A on Y, one computes the causal effect of A on Y in each level (stratum) of the variable M. In the previous section, we used the data in Table 4.1 to compute the causal effect of transplant A on death Y in each of the two strata of sex M. Because the causal effect differed between the two strata (on both the additive and the multiplicative scale), we concluded that there was (additive and multiplicative) effect modification by M of the causal effect of A on Y.

But the data in Table 4.1 are not the typical data one encounters in real life. Instead of the two columns with each individual's counterfactual outcomes $Y^{a=1}$ and $Y^{a=0}$, one will find two columns with each individual's treatment level A and observed outcome Y. How does the unavailability of the counterfactual outcomes affect the use of stratification to detect effect modification? The answer depends on the study design.

Consider first an ideal marginally randomized experiment. In Chapter 2 we demonstrated that, leaving aside random variability, the average causal effect of treatment can be computed using the observed data. For example, the causal risk difference $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$ is equal to the observed associational risk difference $\Pr[Y = 1|A = 1] - \Pr[Y = 1|A = 0]$. The same reasoning can be extended to each stratum of the variable M because, if treatment assignment was random and unconditional, exchangeability is expected in every subset of the population. Thus the causal risk difference in women, $\Pr[Y^{a=1} = 1|M = 1] - \Pr[Y^{a=0} = 1|M = 1]$, is equal to the associational risk difference in women, $\Pr[Y = 1|A = 1, M = 1] - \Pr[Y = 1|A = 0, M = 1]$. And similarly for men. Thus, to identify effect modification by M in an ideal experiment with unconditional randomization, one just needs to conduct a stratified analysis by computing the association measure in each level of the variable M.

Consider now an ideal randomized experiment with conditional randomization. In a population of 40 people, transplant A has been randomly assigned with probability 0.75 to those in severe condition ($L = 1$), and with probability 0.50 to the others ($L = 0$). The 40 individuals can be classified into two nationalities: 20 are Greek ($M = 1$) and 20 are Roman ($M = 0$). The data on L, A, and death Y for the 20 Greeks are shown in Table 2.2 (same as Table 3.1). The data for the 20 Romans are shown in Table 4.2. The population risk under treatment, $\Pr[Y^{a=1} = 1]$, is 0.55, and the population risk under no treatment, $\Pr[Y^{a=0} = 1]$, is 0.40. (Both risks are readily calculated by using either standardization or IP weighting. We leave the details to the reader.) The average causal effect of transplant A on death Y is therefore $0.55 - 0.40 = 0.15$ on the risk difference scale, and $0.55/0.40 = 1.375$ on the risk ratio scale. In this population, heart transplant increases the mortality risk. As discussed in the previous chapter, the calculation of the causal effect would have been the same if the data had arisen from an observational study in which we believe that conditional exchangeability $Y^a \perp\!\!\!\perp A|L$ holds.

We now discuss how to conduct a stratified analysis to investigate whether nationality M modifies the effect of A on Y. The goal is to compute the causal effect of A on Y in the Greeks, $\Pr[Y^{a=1} = 1|M = 1] - \Pr[Y^{a=0} = 1|M = 1]$, and in the Romans, $\Pr[Y^{a=1} = 1|M = 0] - \Pr[Y^{a=0} = 1|M = 0]$. If these two causal risk differences differ, we will say that there is additive effect modification by M. And similarly for the causal risk ratios if interested in multiplicative effect modification.

The procedure to compute the conditional risks $\Pr[Y^{a=1} = 1|M = m]$ and

Fine Point 4.1

Effect in the treated. This chapter is concerned with average causal effects in subsets of the population. One particular subset is the treated ($A = 1$). The *average causal effect in the treated* is not null if $\Pr[Y^{a=1} = 1|A = 1] \neq \Pr[Y^{a=0} = 1|A = 1]$ or, by consistency, if

$$\Pr[Y = 1|A = 1] \neq \Pr[Y^{a=0} = 1|A = 1].$$

That is, there is a causal effect in the treated if the observed risk among the treated subjects does not equal the counterfactual risk had the treated subjects been untreated. The causal risk difference in the treated is $\Pr[Y = 1|A = 1] - \Pr[Y^{a=0} = 1|A = 1]$. The causal risk ratio in the treated, also known as the standardized morbidity ratio (SMR), is $\Pr[Y = 1|A = 1]/\Pr[Y^{a=0} = 1|A = 1]$. The causal risk difference and risk ratio in the untreated are analogously defined by replacing $A = 1$ by $A = 0$. Figure 4.1 shows the groups that are compared when computing the effect in the treated and the effect in the untreated.

The average effect in the treated will differ from the average effect in the population if the distribution of individual causal effects varies between the treated and the untreated. That is, when computing the effect in the treated, treatment group $A = 1$ is used as a marker for the factors that are truly responsible for the modification of the effect between the treated and the untreated groups. However, even though one could say that there is effect modification by the pretreatment variable M even if M is only a surrogate (e.g., nationality) for the causal effect modifiers, one would not say that there is modification of the effect A by treatment A because it sounds nonsensical. See Section 6.6 for a graphical representation of true and surrogate effect modifiers. The bulk of this book is focused on the causal effect in the population because the causal effect in the treated, or in the untreated, cannot be directly generalized to time-varying treatments (see Part III).

$\Pr[Y^{a=0} = 1|M = m]$ in each stratum m has two stages: 1) stratification by M, and 2) standardization by L (or, equivalently, IP weighting). We computed the standardized risks in the Greek stratum ($M = 1$) in Chapter 2: the causal risk difference was 0 and the causal risk ratio was 1. Using the same procedure in the Roman stratum ($M = 0$), we can compute the risks $\Pr[Y^{a=1} = 1|M = 0] = 0.6$ and $\Pr[Y^{a=0} = 1|M = 0] = 0.3$. (Again we leave the details to the reader.) Therefore, the causal risk difference is 0.3 and the causal risk ratio is 2 in the stratum $M = 0$. Because these effect measures differ from those in the stratum $M = 1$, we say that there is both additive and multiplicative effect modification by nationality M of the effect of transplant A on death Y. This effect modification is not qualitative because the effect is harmful or null in both strata $M = 0$ and $M = 1$.

We have shown that, in our population, nationality M modifies the effect of heart transplant A on the risk of death Y. However, we have made no claims about the mechanisms involved in such effect modification. In fact, it is possible that nationality is simply a marker for the factor that is truly responsible for the effect modification. For example, suppose that the quality of heart surgery is better in Greece than in Rome. One would then find effect modification by nationality even though, technically, passport-defined nationality does not modify the effect. For example, improving the quality of heart surgery in Rome, or moving Romans to Greece, would eliminate the modification of the effect by nationality. We refer to nationality as a *surrogate effect modifier*, and to quality of care as a *causal effect modifier*. See Section 6.6 for a graphical representation of surrogate and causal effect modifiers.

Similarly, we may say that there is effect modification by time (or by age) if the effect measure in a certain group of people at one time differs from the effect measure in the same group at a later time. In this case the strata defined by the effect modifier M do not correspond to different groups of people but to

the same people at different times. Thus our use of the term effect modification by M does not necessarily imply that M plays a causal role. To avoid potential confusions, some authors prefer to use the more neutral term "heterogeneity of causal effects across strata of M" rather than "effect modification by M." The next chapter introduces "interaction," a concept related to effect modification, that does attribute a causal role to the variables involved.
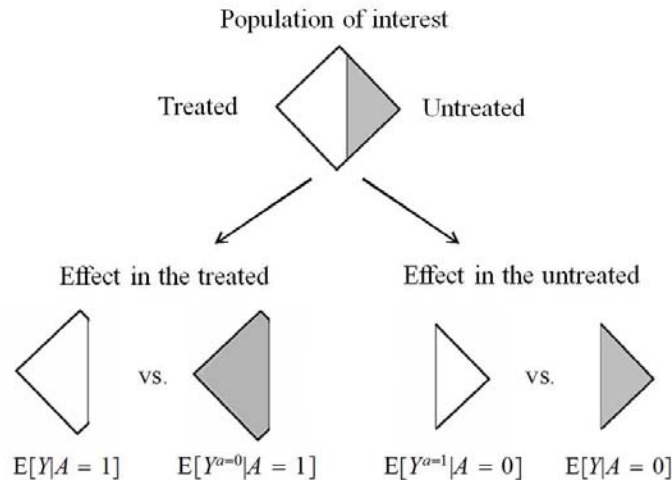

Figure 4.1

4.3 Reasons to care about effect modification

There are several related reasons why investigators are interested in identifying effect modification, and why it is important to collect data on pre-treatment descriptors M even in randomized experiments.

First, if a factor M modifies the effect of treatment A on the outcome Y then the average causal effect will differ between populations with different prevalence of M. For example, the average causal effect in the population of Table 4.1 is harmful in women and beneficial in men. Because there are 50% of subjects of each sex and the sex-specific harmful and beneficial effects are equal but of opposite sign, the average causal effect in the entire population is null. However, had we conducted our study in a population with a greater proportion of women (e.g., graduating college students), the average causal effect in the entire population would have been harmful. Other examples: the effects of exposure to asbestos differ between smokers and nonsmokers, the effects of antiretroviral therapy differ between relatively healthy and severely ill HIV-infected individuals, the effects of universal health care differ between low-income and high-income families. That is, the average causal effect in a population depends on the distribution of individual causal effects in the population. There is generally no such a thing as "*the* average causal effect of treatment A on outcome Y (period)", but "the average causal effect of treatment A on outcome Y in a population with a particular mix of causal effect modifiers."

Technical Point 4.1

Computing the effect in the treated. We computed the average causal effect in the population under conditional exchangeability $Y^a \coprod A|L$ for both $a = 0$ and $a = 1$. Computing the average causal effect in the treated only requires *partial exchangeability* $Y^{a=0} \coprod A|L$. In other words, it is irrelevant whether the risk in the untreated, had they been treated, equals the risk in those who were actually treated. The *average causal effect in the untreated* is computed under the partial exchangeability condition $Y^{a=1} \coprod A|L$.

We now describe how to compute the counterfactual risk $\Pr[Y^a = 1|A = a']$ via standardization, and a more general approach to compute the counterfactual mean $\mathrm{E}[Y^a|A = a']$ via IP weighting, under the above assumptions of partial exchangeability:

- Standardization: $\Pr[Y^a = 1|A = a']$ is equal to $\sum_l \Pr[Y = 1|A = a, L = l]\Pr[L = l|A = a']$. See Miettinen (1973) for a discussion of standardized risk ratios.

- IP weighting: $\mathrm{E}[Y^a|A = a']$ is equal to the IP weighted mean $\dfrac{\mathrm{E}\left[\dfrac{I(A = a)Y}{f(A|L)}\Pr[A = a'|L]\right]}{\mathrm{E}\left[\dfrac{I(A = a)}{f(A|L)}\Pr[A = a'|L]\right]}$ with weights

  $\dfrac{\Pr[A = a'|L]}{f(A|L)}$. For dichotomous A, this equality was derived by Sato and Matsuyama (2003). See Hernán and Robins (2006) for further details.

Some refer to lack of transportability as lack of external validity.

In our example, the effect of heart transplant A on risk of death Y differs between men and women, and between Romans and Greeks. Thus our knowledge about the average causal effect in this population may not be transportable to other populations with a different distribution of the effect modifiers sex and nationality. We then say that the average causal effect is not transportable or generalizable to other populations.

The extrapolation of causal effects computed in one population to a second population, which is also referred to as the *transportability* of causal inferences across populations, can be improved by restricting our attention to the average causal effects in the strata defined by the effect modifiers rather than to the average effect in the population. Unfortunately, there is no guarantee that this conditional effect measures accurately quantify the conditional effects in the second population. There could be other unmeasured, or unknown, causal effect modifiers whose conditional distribution varies between the two populations. See also Fine Point 4.2. Transportability of causal effects is an assumption that relies heavily on subject-matter knowledge. For example, most experts would agree that the health effects (on either the additive or multiplicative scale) of increasing a household's annual income by \$100 in Niger cannot be generalized to the Netherlands, but most experts would agree that the health effects of use of cholesterol-lowering drugs in Europeans can be generalized to Canadians.

A setting in which generalizability may not be an issue: Smith and Pell (2003) could not identify any major modifiers of the effect of parachute use on death after "gravitational challenge" (e.g., jumping from an airplane at high altitude). They concluded that conducting randomized trials of parachute use restricted to a particular group of people would not compromise the transportability of the findings to other groups.

Second, evaluating the presence of effect modification is helpful to identify the groups of subjects that would benefit most from an intervention. In our example of Table 4.1, the average causal effect of treatment A on outcome Y was null. However, treatment A had a beneficial effect in men ($M = 0$), and a harmful effect in women ($M = 1$). If a physician knew that there is qualitative effect modification by sex then, in the absence of additional information, she

would treat the next patient only if he happens to be a man. The situation is slightly more complicated when, as in our second example, there is multiplicative, but not additive, effect modification. Here treatment reduces the risk of the outcome by 10% in subjects with $M = 0$ and also by 10% in subjects with $M = 1$, i.e., there is no additive effect modification by M because the causal risk difference is 0.1 in all levels of M. Thus, an intervention to treat all patients would be equally effective in reducing risk in both strata of M, despite the fact that there is multiplicative effect modification. Additive, but not multiplicative, effect modification is the appropriate scale to identify the groups that will benefit most from intervention. To see this, note that an effect modifier on either the additive or the multiplicative scale is guaranteed to exist when the sharp causal null hypothesis does not hold (i.e., when the treatment has a non-null effect on some subjects' outcomes). However, in the absence of additive effect modification, it is usually not very helpful to learn that there is multiplicative effect modification. In our second example, the presence of multiplicative effect modification follows from the mathematical fact that, because the risk under no treatment in the stratum $M = 1$ equals 0.8, the maximum possible causal risk ratio in the $M = 1$ stratum is $1/0.8 = 1.25$. Thus the causal risk ratio in the stratum $M = 1$ is guaranteed to differ from the causal risk ratio of 2 in the $M = 0$ stratum. In these situations, the presence of multiplicative effect modification is simply the consequence of different risk under no treatment $\Pr[Y^{a=0} = 1|M = m]$ across levels of M. In these cases, it is more informative to report the risk differences (and, even better, the absolute risks) than the risk ratios.

Several authors (e.g., Blot and Day, 1979; Rothman et al., 1980; Saracci, 1980) have referred to additive effect modification as the one of interest for public health purposes.

Finally, the identification of effect modification may help understand the biological, social, or other mechanisms leading to the outcome. For example, a greater risk of HIV infection in uncircumcised compared with circumcised men may provide new clues to understand the disease. The identification of effect modification may be a first step towards characterizing the interactions between two treatments. In fact, the terms "effect modification" and "interaction" are sometimes used as synonymous in the scientific literature. The next chapter describes "interaction" as a causal concept that is related to, but different from, effect modification.

4.4 Stratification as a form of adjustment

Until this chapter, our only goal was to compute the average causal effect in the entire population. In the absence of marginal randomization, achieving this goal requires adjustment for the variables L that ensure conditional exchangeability of the treated and the untreated. For example, in Chapter 2 we determined that the average causal effect of heart transplant A on mortality Y was null, that is, the causal risk ratio $\Pr\left[Y^{a=1} = 1\right] / \Pr\left[Y^{a=0} = 1\right] = 1$. We used the data in Table 2.2 to adjust for the prognostic factor L via both standardization and IP weighting.

The present chapter adds another potential goal to the analysis: to identify effect modification by variables M. To achieve this goal, we need to stratify by M before adjusting for L. For example, in this chapter we stratified by nationality M before adjusting for the prognostic factor L to determine that the average causal effect of heart transplant A on mortality Y differed between Greeks and Romans. In summary, standardization (or IP weighting) is used to adjust for L and stratification is used to identify effect modification by M.

Fine Point 4.2

Transportability.   Causal effects estimated in one population (the study population) are often intended to make decisions in another population (the target population). Suppose we have correctly estimated the average causal effect of treatment R in our study population under exchangeability, positivity, and well-defined interventions. Will the effect be the same in the target population? That is, can we "transport" the effect from one population to the other? The answer to this question depends on the characteristics of both populations. Specifically, transportability of effects from one population to another may be justified if the following characteristics are similar between the two populations:

- Effect modification: The causal effect of treatment R may differ across individuals with different susceptibility to the outcome. For example, if women are more susceptible to the effects of treatment than men, we say that sex is an effect modifier. The distribution of effect modifiers in a population will generally affect the magnitude of the causal effect of treatment in that population. Discussions about *generalizability* of causal effects are often focused on effect modification.

- Interference: In many settings treating one individual may indirectly affect the treatment level of other individuals in the population. For example, a socially and physically active individual may convince his friends to get treated, and thus an intervention on that individual may be more effective than an intervention on a socially isolated individual. The distribution of contact patterns among individuals may affect the magnitude of the causal effect of treatment R in a population. See Halloran and Struchiner (1995), Sobel (2006), Rosenbaum (2007), and Hudgens and Halloran (2009) for a more detailed discussion of the role of interference in the definition of causal effects.

- Versions of treatment: The causal effect of treatment R depends on the distribution of versions of treatment in the population. If this distribution differs between the study population and the target population, then the magnitude of the causal effect of treatment R will differ too.

Note that the transportability of causal inferences across populations may be improved by restricting our attention to the average causal effects in the strata defined by the effect modifiers (rather than to the average effect), or by using the stratum-specific effects in the study population to reconstruct the average causal effect in the target population. For example, the four stratum-specific effect measures (Roman women, Greek women, Roman men, and Greek men) in our population can be combined in a weighted average to reconstruct the average causal effect in another population with a different mix of sex and nationality. The weight assigned to each stratum-specific measure is the proportion of subjects in that stratum in the second population. However, there is no guarantee that this reconstructed effect will coincide with the true effect in the target population because of possible unmeasured effect modifiers, and between-populations differences in interference patterns and distribution of versions of treatment.

Under conditional exchangeability given L, the risk ratio in the subset $L = l$ measures the average causal effect in the subset $L = l$ because, if $Y^a \coprod A|L$, then
$\Pr[Y = 1|A = a, L = 0] = \Pr[Y^a = 1|L = 0]$

But stratification is not always used to identify effect modification by M. In practice stratification is often used as an alternative to standardization (and IP weighting) to adjust for L. In fact, the use of stratification as a method to adjust for L is so widespread that many investigators consider the terms "stratification" and "adjustment" as synonymous. For example, suppose you ask an epidemiologist to adjust for the prognostic factor L to compute the effect of heart transplant A on mortality Y. Chances are that she will immediately split Table 2.2 into two subtables—one restricted to subjects with $L = 0$, the other to subjects with $L = 1$—and would provide the effect measure (say, the risk ratio) in each of them. That is, she would calculate the risk ratios $\Pr[Y = 1|A = 1, L = l] / \Pr[Y = 1|A = 0, L = l] = 1$ for both $l = 0$ and $l = 1$.

These two stratum-specific associational risk ratios can be endowed with a causal interpretation under conditional exchangeability given L: they measure the average causal effect in the subsets of the population defined by $L = 0$

and $L = 1$, respectively. They are *conditional effect measures*. In contrast the risk ratio of 1 that we computed in Chapter 2 was a marginal (unconditional) effect measure. In this particular example, all three risk ratios—the two conditional ones and the marginal one—happen to be equal because there is no effect modification by L. Stratification necessarily results in multiple stratum-specific effect measures (one per stratum defined by the variables L). Each of them quantifies the average causal effect in a nonoverlapping subset of the population but, in general, none of them quantifies the average causal effect in the entire population. Therefore, we did not consider stratification when describing methods to compute the average causal effect of treatment in the population in Chapter 2. Rather, we focused on standardization and IP weighting.

In addition, unlike standardization and IP weighting, adjustment via stratification requires computing the effect measures in subsets of the population defined by a combination of *all* variables L that are required for conditional exchangeability. For example, when using stratification to estimate the effect of heart transplant in the population of Tables 2.2 and 4.2, one must compute the effect in Romans with $L = 1$, in Greeks with $L = 1$, in Romans with $L = 0$, and in Greeks with $L = 0$; but one cannot compute the effect in Romans by simply computing the association in the stratum $M = 0$ because nationality M, by itself, is insufficient to guarantee conditional exchangeability. That is,

Robins (1986, 1987) described the conditions under which stratum-specific effect measures for time-varying treatments will not have a causal interpretation even if in the presence of exchangeability, positivity, and well-defined interventions.

the use of stratification forces one to evaluate effect modification by all variables L required to achieve conditional exchangeability, regardless of whether one is interested in such effect modification. In contrast, stratification by M followed by IP weighting or standardization to adjust for L allows one to deal with exchangeability and effect modification separately, as described above. Other problems associated with the use of stratification are noncollapsibility of certain effect measures (see Fine Point 4.3) and inappropriate adjustment when, in the case for time-varying treatments, it is necessary to adjust for time-varying variables L that are affected by prior treatment (see Part III).

Stratification requires positivity in addition to exchangeability: the causal effect cannot be computed in subsets $L = l$ in which there are only treated, or untreated, individuals.

Sometimes investigators compute the causal effect in only some of the strata defined by the variables L. That is, no stratum-specific effect measure is computed for some strata. This form of stratification is known as *restriction*. Stratification is simply the application of restriction to several comprehensive and mutually exclusive subsets of the population. An important use of restriction is the preservation of positivity (see Chapter 3).

4.5 Matching as another form of adjustment

Our discussion on matching applies to cohort studies only. Under study designs not yet discussed (i.e., case-control studies), matching is used for purposes other than adjustment, and thus needs to be followed by some form of stratification to estimate conditional (stratum-specific) effect measures.

Matching is another adjustment method. The goal of matching is to construct a subset of the population in which the variables L have the same distribution in both the treated and the untreated. As an example, take our heart transplant example in Table 2.2 in which the variable L is sufficient to achieve conditional exchangeability. For each untreated individual in non critical condition ($A = 0, L = 0$) randomly select a treated individual in non critical condition ($A = 1, L = 0$), and for each untreated individual in critical condition ($A = 0, L = 1$) randomly select a treated individual in critical condition ($A = 1, L = 1$). We refer to each untreated individual and her corresponding treated individual as a matched pair, and to the variable L as the matching factor. Suppose we formed the following 7 matched pairs: Rheia-Hestia, Kronos-Poseidon, Demeter-Hera, Hades-Zeus for $L = 0$, and Artemis-Ares, Apollo-Aphrodite, Leto-Hermes for

$L = 1$. All the untreated, but only a sample of treated, in the population were selected. In this subset of the population comprised of matched pairs, the proportion of individuals in critical condition ($L = 1$) is the same, by design, in the treated and in the untreated ($3/7$).

To construct our matched population we replaced the treated in the population by a subset of the treated in which the matching factor L had the same distribution as that in the untreated. Under the assumption of conditional exchangeability given L, the result of this procedure is (unconditional) exchangeability of the treated and the untreated in the matched population. Because the treated and the untreated are exchangeable in the matched population, their average outcomes can be directly compared: the risk in the treated is $3/7$, the risk in the untreated is $3/7$, and hence the causal risk ratio is 1. Note that matching ensures *positivity* in the matched population because strata with only treated, or untreated, individuals are excluded from the analysis.

Often one chooses the group with fewer subjects (the untreated in our example) and uses the other group (the treated in our example) to find their matches. The chosen group defines the subpopulation on which the causal effect is being computed. In the previous paragraph we computed the *effect in the untreated*. In settings with fewer treated than untreated individuals across all strata of L, we generally compute the *effect in the treated*. Also, matching needs not be one-to-one (matching pairs), but it can be one-to-many (matching sets).

In many applications, L is a vector of several variables. Then, for each untreated individual in a given stratum defined by a combination of values of all the variables in L, we would have randomly selected one (or several) treated individual(s) from the same stratum.

As the number of matching factors increases, so does the probability that no exact matches exist for an individual. There is a vast literature, beyond the scope of this book, on how to find approximate matches in those settings.

Matching can be used to create a matched population with any chosen distribution of L, not just the distribution in the treated or the untreated. The distribution of interest can be achieved by individual matching, as described above, or by *frequency matching*. An example of the latter is a study in which one randomly selects treated subjects in such a way that 70% of them have $L = 1$, and then repeats the same procedure for the untreated.

Because the matched population is a subset of the original study population, the distribution of causal effect modifiers in the matched study population will generally differ from that in the original, unmatched study population, as discussed in the next section.

4.6 Effect modification and adjustment methods

Part II describes how standardization, IP weighting, and stratification can be used in combination with parametric or semiparametric models. For example, standard regression models are a form of stratification in which the association between treatment and outcome is estimated within levels of all the other covariates in the model.

Standardization, IP weighting, stratification/restriction, and matching are different approaches to estimate average causal effects, but they estimate different types of causal effects. These four approaches can be divided into two groups according to the type of effect they estimate: standardization and IP weighting can be used to compute either marginal or conditional effects, stratification/restriction and matching can only be used to compute conditional effects in certain subsets of the population. All four approaches require exchangeability, positivity, and well-defined interventions, but the subsets of the population in which these conditions need to hold depend on the causal effect of interest. For example, to compute the conditional effect among individuals with $L = l$, any of the above methods requires exchangeability in that subset only; to estimate the marginal effect in the entire population, IP weighting and

Technical Point 4.2

Pooling of stratum-specific effect measures. So far we have focused on the conceptual, non statistical, aspects of causal inference by assuming that we work with the entire population rather than with a sample from it. Thus we talk about computing causal effects rather than about (consistently) estimating them. In the real world, however, we can rarely compute causal effects in the population. We need to estimate them from samples, and thus obtaining reasonably narrow confidence intervals around our estimated effect measures is an important practical concern.

When dealing with stratum-specific effect measures, one commonly used strategy to reduce the variability of the estimates is to combine all stratum-specific effect measures into one pooled stratum-specific effect measure. The idea is that, *if the effect measure is the same in all strata* (i.e., if there is no effect-measure modification), then the pooled effect measure will be a more precise estimate of the common effect measure. Several methods (e.g., Woolf, Mantel-Haenszel, maximum likelihood) yield a pooled estimate, sometimes by computing a weighted average of the stratum-specific effect measures with weights chosen to reduce the variability of the pooled estimate. Greenland and Rothman (2008) review some commonly used methods for stratified analysis. Regression models can also be used to compute pooled effect measures. To do so, the model needs to include all possible product ("interaction") terms between all covariates L, but no product terms between treatment A and covariates L. That is, the model must be saturated (see Chapter 11) with respect to L.

The main goal of pooling is to obtain a narrower confidence interval around the common stratum-specific effect measure, but the pooled effect measure is still a conditional effect measure. In our heart transplant example, the pooled stratum-specific risk ratio (Mantel-Haenszel method) was 0.88 for the outcome Z. This result is only meaningful if the stratum-specific risk ratios 2 and 0.5 are indeed estimates of the same stratum-specific causal effect. For example, suppose that the causal risk ratio is 0.9 in both strata but, because of the small sample size, we obtained estimates of 0.5 and 2.0. In that case, pooling would be appropriate and the Mantel-Haenszel risk ratio would be closer to the truth than either of the stratum-specific risk ratios. Otherwise, if the causal stratum-specific risk ratios are truly 0.5 and 2.0, then pooling makes little sense and the Mantel-Haenszel risk ratio could not be easily interpreted. In practice, it is not always obvious to determine whether the heterogeneity of the effect measure across strata is due to sampling variability or to effect-measure modification. The finer the stratification, the greater the uncertainty introduced by random variability.

Table 4.3

	L	A	Z
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	1
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	0
Cyclope	1	1	0
Persephone	1	1	0
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

standardization require exchangeability in all levels of L.

In the absence of effect modification, the effect measures computed via these four approaches will be equal. For example, we concluded that the average causal effect of heart transplant A on mortality Y was null both in the entire population of Table 2.2 (standardization and IP weighting), in the subsets of the population in critical condition $L = 1$ and non critical condition $L = 0$ (stratification), and in the untreated (matching). All methods resulted in a causal risk ratio equal to 1. However, the effect measures computed via these four approaches will not generally be equal. To illustrate how the effects may vary, let us compute the effect of heart transplant A on high blood pressure Z (1: yes, 0 otherwise) using the data in Table 4.3. We assume that exchangeability $Z^a \coprod A|L$ and positivity hold. We use the risk ratio scale for no particular reason.

Standardization and IP weighting yield the average causal effect in the entire population $\Pr[Z^{a=1} = 1]/\Pr[Z^{a=0} = 1] = 0.8$ (these and the following calculations are left to the reader). Stratification yields the conditional causal risk ratios $\Pr[Z^{a=1} = 1|L = 0]/\Pr[Z^{a=0} = 1|L = 0] = 2.0$ in the stratum $L = 0$, and $\Pr[Z^{a=1} = 1|L = 1]/\Pr[Z^{a=0} = 1|L = 1] = 0.5$ in the stratum $L = 1$. Matching, using the matched pairs selected in the previous section, yields the causal risk ratio in the untreated $\Pr[Z^{a=1} = 1|A = 0]/\Pr[Z = 1|A = 0] = 1.0$.

We have computed four causal risk ratios and have obtained four different numbers: $0.8, 2.0, 0.5,$ and 1.0. All of them are correct. Leaving aside random

variability (see Technical Point 4.2), the explanation of the differences is qualitative effect modification: Treatment doubles the risk among individuals in noncritical condition ($L = 0$, causal risk ratio 2.0) and halves the risk among individuals in critical condition ($L = 1$, causal risk ratio 0.5). The average causal effect in the population (causal risk ratio 0.8) is beneficial because there are more individuals in critical condition than in noncritical condition. The causal effect in the untreated is null (causal risk ratio 1.0), which reflects the larger proportion of individuals in noncritical condition in the untreated compared with the entire population. This example highlights the primary importance of specifying the population, or the subset of a population, to which the effect measure corresponds.

The previous chapter argued that a well defined causal effect is a prerequisite for meaningful causal inference. This chapter argues that a well characterized target population is another such prerequisite. Both prerequisites are automatically present in experiments that compare two or more interventions in a population that meets certain a priori eligibility criteria. However, these prerequisites cannot be taken for granted in observational studies. Rather, investigators conducting observational studies need to explicitly define the causal effect of interest and the subset of the population in which the effect is being computed. Otherwise, misunderstandings might easily arise when effect measures obtained via different methods are different. In our example above, one investigator who used IP weighting (and computed the effect in the entire population) and another one who used matching (and computed the effect in the untreated) need not engage in a debate about the superiority of one analytic approach over the other. Their discrepant effect measures result from the different causal question asked by each investigator rather than from their choice of analytic approach. In fact, the second investigator could have used IP weighting to compute the effect in the untreated or in the treated (see Technical Point 4.1).

A final note. Stratification can be used to compute average causal effects in subsets of the population, but not individual (subject-specific) effects. We cannot generally compare the mortality outcome had Zeus been treated with the mortality outcome had he been untreated. Estimating subject-specific effects would require subject-specific exchangeability, e.g., for a treated subject we need a perfectly exchangeable untreated subject. Because the assumption of individual exchangeability is generally untenable, adjustment methods require only exchangeability between groups (i.e., the treated and the untreated). As a result, only average causal effects in groups—populations or subsets of populations—can be computed in general.

Table 4.4

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

Fine Point 4.3

Collapsibility and the odds ratio. In the absence of effect modification by M, the causal risk ratio in the entire population, $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$ is equal to the conditional causal risk ratios $\Pr[Y^{a=1}=1|M=m]/\Pr[Y^{a=0}=1|M=m]$ in every stratum m of M. More generally, the causal risk ratio is a weighted average of the stratum-specific risk ratios. For example, if the causal risk ratios in the strata $M=1$ and $M=0$ were equal to 2 and 3, respectively, then the causal risk ratio in the population would be greater than 2 and less than 3. That the value of the causal risk ratio (and the causal risk difference) in the population is always constrained by the range of values of the stratum-specific risk ratios is not only obvious but also a desirable characteristic of any effect measure.

Now consider a hypothetical effect measure (other than the risk ratio or the risk difference) such that the population effect measure were not a weighted average of the stratum-specific measures. That is, the population effect measure would not necessarily lie inside of the range of values of the stratum-specific effect measures. Such effect measure would be an odd one. The odds ratio (pun intended) is such an effect measure, as we now discuss.

Suppose the data in Table 4.4 were collected to compute the causal effect of altitude A on depression Y in a population of 20 individuals. The treatment A is 1 if the subject moved to a high altitude residence (on the top of Mount Olympus), 0 otherwise; the outcome Y is 1 if the subject developed depression, 0 otherwise; and M is 1 if the subject was female, 0 if male. The decision to move was random, i.e., those more prone to develop depression were as likely to move as the others; effectively $Y^a \amalg A$. Therefore the risk ratio $\Pr[Y=1|A=1]/\Pr[Y=1|A=0]=2.3$ is the causal risk ratio in the population, and the odds ratio $\dfrac{\Pr[Y=1|A=1]/\Pr[Y=0|A=1]}{\Pr[Y=1|A=0]/\Pr[Y=0|A=0]}=5.4$ is the causal odds ratio $\dfrac{\Pr[Y^{a=1}=1]/\Pr[Y^{a=1}=0]}{\Pr[Y^{a=0}=1]/\Pr[Y^{a=0}=0]}$ in the population. The risk ratio and the odds ratio measure the same causal effect on different scales.

Let us now compute the sex-specific causal effects on the risk ratio and odds ratio scales. The (conditional) causal risk ratio $\Pr[Y=1|M=m,A=1]/\Pr[Y=1|M=m,A=0]$ is 2 for men ($M=0$) and 3 for women ($M=1$). The (conditional) causal odds ratio $\dfrac{\Pr[Y=1|M=m,A=1]/\Pr[Y=0|M=m,A=1]}{\Pr[Y=1|M=m,A=0]/\Pr[Y=0|M=m,A=0]}$ is 6 for men ($M=0$) and 6 for women ($M=1$). The causal risk ratio in the population, 2.3, is in between the sex-specific causal risk ratios 2 and 3. In contrast, the causal odds ratio in the population, 5.4, is smaller (i.e., closer to the null value) than both sex-specific odds ratios, 6. The causal effect, when measured on the odds ratio scale, is bigger in each half of the population than in the entire population. In general, the population causal odds ratio can be closer to the null value than any of the non-null stratum-specific causal odds ratios when M is associated with Y (Miettinen and Cook, 1981).

We say that an effect measure is collapsible when the population effect measure can be expressed as a weighted average of the stratum-specific measures. In follow-up studies the risk ratio and the risk difference are collapsible effect measures, but the odds ratio—or the rarely used odds difference—is not (Greenland 1987). The noncollapsibility of the odds ratio, which is a special case of Jensen's inequality (Samuels 1981), may lead to counterintuitive findings like those described above. The odds ratio is collapsible under the sharp null hypothesis—both the conditional and unconditional effect measures are then equal to the null value—and it is approximately collapsible—and approximately equal to the risk ratio—when the outcome is rare (say, $<10\%$) in every stratum of a follow-up study.

One important consequence of the noncollapsibility of the odds ratio is the logical impossibility of equating "lack of exchangeability" and "change in the conditional odds ratio compared with the unconditional odds ratio." In our example, the change in odds ratio was about 10% $(1-6/5.4)$ even though the treated and the untreated were exchangeable. Greenland, Robins, and Pearl (1999) reviewed the relation between noncollapsibility and lack of exchangeability.

Chapter 5
INTERACTION

Consider again a randomized experiment to answer the causal question "does one's looking up at the sky make other pedestrians look up too?" We have so far restricted our interest to the causal effect of a single treatment (looking up) in either the entire population or a subset of it. However, many causal questions are actually about the effects of two or more simultaneous treatments. For example, suppose that, besides randomly assigning your looking up, we also randomly assign whether you stand in the street dressed or naked. We can now ask questions like: what is the causal effect of your looking up if you are dressed? And if you are naked? If these two causal effects differ we say that the two treatments under consideration (looking up and being dressed) interact in bringing about the outcome.

When joint interventions on two or more treatments are feasible, the identification of interaction allows one to implement the most effective interventions. Thus understanding the concept of interaction is key for causal inference. This chapter provides a formal definition of interaction between two treatments, both within our already familiar counterfactual framework and within the sufficient-component-cause framework.

5.1 Interaction requires a joint intervention

Suppose that in our heart transplant example, individuals were assigned to receiving either a multivitamin complex ($E = 1$) or no vitamins ($E = 0$) before being assigned to either heart transplant ($A = 1$) or no heart transplant ($A = 0$). We can now classify all individuals into 4 treatment groups: vitamins-transplant ($E = 1$, $A = 1$), vitamins-no transplant ($E = 1$, $A = 0$), no vitamins-transplant ($E = 0$, $A = 1$), and no vitamins-no transplant ($E = 0$, $A = 0$). For each individual, we can now imagine 4 potential or counterfactual outcomes, one under each of these 4 treatment combinations: $Y^{a=1,e=1}$, $Y^{a=1,e=0}$, $Y^{a=0,e=1}$, and $Y^{a=0,e=0}$. In general, an individual's counterfactual outcome $Y^{a,e}$ is the outcome that would have been observed if we had intervened to set the individual's values of A and E to a and e, respectively. We refer to interventions on two or more treatments as *joint interventions*.

We are now ready to provide a definition of interaction within the counterfactual framework. There is interaction between two treatments A and E if the causal effect of A on Y after a joint intervention that set E to 1 differs from the causal effect of A on Y after a joint intervention that set E to 0. For example, there would be an interaction between transplant A and vitamins E if the causal effect of transplant on survival had everybody taken vitamins were different from the causal effect of transplant on survival had nobody taken vitamins.

When the causal effect is measured on the risk difference scale, we say that there is *interaction between A and E on the additive scale* if

$$\Pr\left[Y^{a=1,e=1} = 1\right] - \Pr\left[Y^{a=0,e=1} = 1\right] \neq \Pr\left[Y^{a=1,e=0} = 1\right] - \Pr\left[Y^{a=0,e=0} = 1\right].$$

For example, suppose the causal risk difference for transplant A when everybody receives vitamins, $\Pr\left[Y^{a=1,e=1} = 1\right] - \Pr\left[Y^{a=0,e=1} = 1\right]$, were 0.1, and

that the causal risk difference for transplant A when nobody receives vitamins, $\Pr\left[Y^{a=1,e=0}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]$, were 0.2. We say that there is interaction between A and E on the additive scale because the risk difference $\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=0,e=1}=1\right]$ is less than the risk difference $\Pr\left[Y^{a=1,e=0}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]$. It can be easily shown that this inequality implies that the causal risk difference for vitamins E when everybody receives a transplant, $\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=1,e=0}=1\right]$, is also less than the causal risk difference for vitamins E when nobody receives a transplant A, $\Pr\left[Y^{a=0,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]$. That is, we can equivalently define interaction between A and E on the additive scale as

$$\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=1,e=0}=1\right]\neq\Pr\left[Y^{a=0,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right].$$

Let us now review the difference between interaction and effect modification. As described in the previous chapter, a variable M is a modifier of the effect of A on Y when the average causal effect of A on Y varies across levels of M. Note the concept of effect modification refers to the causal effect of A, not to the causal effect of M. For example, sex was an effect modifier for the effect of heart transplant in Table 4.1, but we never discussed the effect of sex on death. Thus, when we say that M modifies the effect of A we are not considering M and A as variables of equal status, because only A is considered to be a variable on which we could hypothetically intervene. That is, the definition of effect modification involves the counterfactual outcomes Y^{a}, not the counterfactual outcomes $Y^{a,m}$. In contrast, the definition of interaction between A and E gives equal status to both treatments A and E, as reflected by the two equivalent definitions of interaction shown above. The concept of interaction refers to the joint causal effect of two treatments A and E, and thus involves the counterfactual outcomes $Y^{a,e}$ under a joint intervention.

5.2 Identifying interaction

In previous chapters we have described the conditions that are required to identify the average causal effect of a treatment A on an outcome Y, either in the entire population or in a subset of it. The three key identifying conditions were exchangeability, positivity, and well-defined interventions. Because interaction is concerned with the joint effect of two (or more) treatments A and E, identifying interaction requires exchangeability, positivity, and well-defined interventions for both treatments.

Suppose that vitamins E were randomly, and unconditionally, assigned by the investigators. Then positivity and well-defined interventions hold, and the treated $E=1$ and the untreated $E=0$ are expected to be exchangeable. That is, the risk that would have been observed if all subjects had been assigned to transplant $A=1$ and vitamins $E=1$ equals the risk that would have been observed if all subjects who received $E=1$ had been assigned to transplant $A=1$. Formally, the marginal risk $\Pr\left[Y^{a=1,e=1}=1\right]$ is equal to the conditional risk $\Pr\left[Y^{a=1}=1|E=1\right]$. As a result, we can rewrite the definition of interaction between A and E on the additive scale as

$$\Pr\left[Y^{a=1}=1|E=1\right]-\Pr\left[Y^{a=0}=1|E=1\right]$$
$$\neq\Pr\left[Y^{a=1}=1|E=0\right]-\Pr\left[Y^{a=0}=1|E=0\right],$$

Technical Point 5.1

Interaction on the additive and multiplicative scales. The equality of causal risk differences $\Pr\left[Y^{a=1,e=1}=1\right]-$
$\Pr\left[Y^{a=0,e=1}=1\right]=\Pr\left[Y^{a=1,e=0}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]$ can be rewritten as

$$\Pr\left[Y^{a=1,e=1}=1\right]=\left\{\Pr\left[Y^{a=1,e=0}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]\right\}+\Pr\left[Y^{a=0,e=1}=1\right].$$

By subtracting $\Pr\left[Y^{a=0,e=0}=1\right]$ from both sides of the equation, we get $\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]=$

$$\left\{\Pr\left[Y^{a=1,e=0}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]\right\}+\left\{\Pr\left[Y^{a=0,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]\right\}.$$

This equality is a compact way to show that treatments A and E have equal status in the definition of interaction.

When the above equality holds, we say that there is no *interaction between A and E on the additive scale*, and we say that the causal risk difference $\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]$ is additive because it can be written as the sum of the causal risk differences that measure the effect of A in the absence of E and the effect of E in the absence of A. Conversely, there is interaction between A and E on the additive scale if $\Pr\left[Y^{a=1,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]\neq$

$$\left\{\Pr\left[Y^{a=1,e=0}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]\right\}+\left\{\Pr\left[Y^{a=0,e=1}=1\right]-\Pr\left[Y^{a=0,e=0}=1\right]\right\}.$$

The interaction is *superadditive* if the 'not equal to' ($\neq$) symbol can be replaced by a 'greater than' ($>$) symbol. The interaction is *subadditive* if the 'not equal to' ($\neq$) symbol can be replaced by a 'less than' ($<$) symbol.

Analogously, one can define interaction on the multiplicative scale when the effect measure is the causal risk ratio, rather than the causal risk difference. We say that there is *interaction between A and E on the multiplicative scale* if

$$\frac{\Pr\left[Y^{a=1,e=1}=1\right]}{\Pr\left[Y^{a=0,e=0}=1\right]}\neq\frac{\Pr\left[Y^{a=1,e=0}=1\right]}{\Pr\left[Y^{a=0,e=0}=1\right]}\times\frac{\Pr\left[Y^{a=0,e=1}=1\right]}{\Pr\left[Y^{a=0,e=0}=1\right]}.$$

The interaction is *supermultiplicative* if the 'not equal to' ($\neq$) symbol can be replaced by a 'greater than' ($>$) symbol. The interaction is *submultiplicative* if the 'not equal to' ($\neq$) symbol can be replaced by a 'less than' ($<$) symbol.

which is exactly the definition of modification of the effect of A by E on the additive scale. In other words, when treatment E is randomly assigned, then the concepts of interaction and effect modification coincide. The methods described in Chapter 4 to identify modification of the effect of A by M can now be applied to identify interaction of A and E by simply replacing the effect modifier M by the treatment E.

Now suppose treatment E was not assigned by investigators. To assess the presence of interaction between A and E, one still needs to compute the four marginal risks $\Pr\left[Y^{a,e}=1\right]$. In the absence of marginal randomization, these risks can be computed for both treatments A and E, under the usual identifying assumptions, by standardization or IP weighting conditional on the measured covariates. An equivalent way of conceptualizing this problem follows: rather than viewing A and E as two distinct treatments with two possible levels (1 or 0) each, one can view AE as a combined treatment with four possible levels (11, 01, 10, 00). Under this conceptualization the identification of interaction between two treatments is not different from the identification of the causal effect of one treatment that we have discussed in previous chapters. The same methods, under the same identifiability conditions, can be used. The only difference is that now there is a longer list of values that the treatment of interest can take, and therefore a greater number of counterfactual outcomes.

Sometimes one may be willing to assume (conditional) exchangeability for

treatment A but not for treatment E, e.g., when estimating the causal effect of A in subgroups defined by E in a randomized experiment. In that case, one cannot generally assess the presence of interaction between A and E, but can still assess the presence of effect modification by E. This is so because one does not need any identifying assumptions involving E to compute the effect of A in each of the strata defined by E. In the previous chapter we used the notation M (rather than E) for variables for which we are not willing to make assumptions about exchangeability, positivity, or well-defined interventions. For example, we concluded that the effect of transplant A was modified by nationality M, but we never required any identifying assumptions for the effect of M because we were not interested in using our data to compute the causal effect of M on Y. Yet we use our subject-matter knowledge to argue that nationality M does not have a causal effect on any individual's Y. That M does not act on the outcome implies that it does not interact with A—no action, no interaction. But M is a modifier of the effect of A on Y because M is correlated with (e.g., it is a proxy for) an unidentified variable that actually has an effect on Y and interacts with A. Thus there can be modification of the effect of A by another variable without interaction between A and that variable.

> Interaction between A and E without modification of the effect of A by E is also logically possible, though probably rare, because it requires dual effects of A and exact cancellations (VanderWeele 2009).

In the above paragraphs we have argued that a sufficient condition for identifying interaction between two treatments A and E is that exchangeability, positivity, and well-defined interventions are all satisfied for the joint treatment (A, E) with the four possible values $(0, 0)$, $(0, 1)$, $(1, 0)$, and $1, 1)$. Then standardization or IP weighting can be used to estimate the joint effects of the two treatments and thus to evaluate interaction between them. In Part III, we show that this condition is not necessary when the two treatments occur at different times. For the remainder of Part I (except this chapter) and most of Part II, we will focus on the causal effect of a single treatment A.

Up to here, we used deterministic counterfactuals for simplicity even though nothing hinged on that. In contrast, the following sections of this chapter review several concepts related to interaction that do actually require that counterfactuals are assumed to be deterministic, and that treatments and outcomes are dichotomous. This oversimplification, though not necessary, makes the study of these concepts manageable and helps clarify some issues that are often misunderstood. As a downside, the oversimplification impedes the practical application of these concepts to many realistic settings.

5.3 Counterfactual response types and interaction

Individuals can be classified in terms of their counterfactual responses. For example, in Table 4.1 (same as Table 1.1), there are four types of people: the "doomed" who will develop the outcome regardless of what treatment they receive (Artemis, Athena, Persephone, Ares), the "immune" who will not develop the outcome regardless of what treatment they receive (Demeter, Hestia, Hera, Hades), the "preventative" who will develop the outcome only if untreated (Hebe, Kronos, Poseidon, Apollo, Hermes, Dyonisus), and the "causative" who will develop the outcome only if treated (Rheia, Leto, Aphrodite, Zeus, Hephaestus, Cyclope). Each combination of counterfactual responses is often referred to as a response pattern or a *response type*. Table 5.1 display the four possible response types.

When considering two dichotomous treatments A and E, there are 16 possible response types because each individual has four counterfactual outcomes,

Table 5.1

Type	$Y^{a=0}$	$Y^{a=1}$
Doomed	1	1
Preventative	1	0
Causative	0	1
Immune	0	0

one under each of the four possible joint interventions on treatments A and E: $(1,1)$, $(0,1)$, $(1,0)$, and $(0,0)$. Table 5.2 shows the 16 response types for two treatments. This section explores the relation between response types and the presence of interaction in the case of two dichotomous treatments A and E and a dichotomous outcome Y.

The first type in Table 5.2 has the counterfactual outcome $Y^{a=1,e=1}$ equal to 1, which means that an individual of this type would die if treated with both transplant and vitamins. The other three counterfactual outcomes are also equal to 1, i.e., $Y^{a=1,e=1} = Y^{a=0,e=1} = Y^{a=1,e=0} = Y^{a=0,e=0} = 1$, which means that an individual of this type would also die if treated with (no transplant, vitamins), (transplant, no vitamins), or (no transplant, no vitamins). In other words, neither treatment A nor treatment E has any effect on the outcome of such individual. He would die no matter what joint treatment he is assigned to. Now consider type 16. All the counterfactual outcomes are 0, i.e., $Y^{a=1,e=1} = Y^{a=0,e=1} = Y^{a=1,e=0} = Y^{a=0,e=0} = 0$. Again, neither treatment A nor treatment E has any effect on the outcome of an individual of this type. She would survive no matter what joint treatment she is assigned to. If all individuals in the population were of types 1 and 16, we would say that neither A nor E has any causal effect on Y; the sharp causal null hypothesis would be true for the joint treatment (A, E). As a consequence, the causal effect of A is independent of E, and vice versa.

Let us now focus our attention on types 4, 6, 11, and 13. Individuals of type 4 would only die if treated with vitamins, whether they do or do not receive a transplant, i.e., $Y^{a=1,e=1} = Y^{a=0,e=1} = 1$ and $Y^{a=1,e=0} = Y^{a=0,e=0} = 0$. Individuals of type 13 would only die if not treated with vitamins, whether they do or do not receive a transplant, i.e., $Y^{a=1,e=1} = Y^{a=0,e=1} = 0$ and $Y^{a=1,e=0} = Y^{a=0,e=0} = 1$. Individuals of type 6 would only die if treated with transplant, whether they do or do not receive vitamins, i.e., $Y^{a=1,e=1} = Y^{a=1,e=0} = 1$ and $Y^{a=0,e=1} = Y^{a=0,e=0} = 0$. Individuals of type 11 would only die if not treated with transplant, whether they do or do not receive vitamins, i.e., $Y^{a=1,e=1} = Y^{a=1,e=0} = 0$ and $Y^{a=0,e=1} = Y^{a=0,e=0} = 1$. If all individuals in the population were of types 4, 6, 11, and 13, we would again say that the causal effect of A is independent of E, and vice versa.

Of the 16 possible response types in Table 5.2, we have identified 6 types (numbers 1, 4, 6, 11, 13, 16) with a common characteristic: for a subject with one of those response types, the causal effect of treatment A on the outcome Y is the same regardless of the value of treatment E, and the causal effect of treatment E on the outcome Y is the same regardless of the value of treatment A. In a population in which every subject has one of these 6 response types, the causal effect of treatment A in the presence of treatment E, as measured by the causal risk difference $\Pr\left[Y^{a=1,e=1} = 1\right] - \Pr\left[Y^{a=0,e=1} = 1\right]$, would equal the causal effect of treatment A in the absence of treatment E, as measured by the causal risk difference $\Pr\left[Y^{a=1,e=0} = 1\right] - \Pr\left[Y^{a=0,e=0} = 1\right]$. That is, if all individuals in the population have response types 1, 4, 6, 11, 13 and 16 then there will be no interaction between A and E on the additive scale.

The presence of additive interaction between A and E implies that, for some individuals in the population, the value of their two counterfactual outcomes under $A = a$ cannot be determined without knowledge of the value of E, and vice versa. That is, there must be individuals in at least one of the following three classes:

1. those who would develop the outcome under only one of the four treatment combinations (types 8, 12, 14, and 15 in Table 5.2)

Table 5.2

	$Y^{a,e}$ for each a, e value			
Type	1,1	0,1	1,0	0,0
1	1	1	1	1
2	1	1	1	0
3	1	1	0	1
4	1	1	0	0
5	1	0	1	1
6	1	0	1	0
7	1	0	0	1
8	1	0	0	0
9	0	1	1	1
10	0	1	1	0
11	0	1	0	1
12	0	1	0	0
13	0	0	1	1
14	0	0	1	0
15	0	0	0	1
16	0	0	0	0

Miettinen (1982) described the 16 possible response types under two binary treatments and outcome.

Greenland and Poole (1988) noted that Miettinen's response types were not invariant to recoding of A and E (i.e., switching the labels "0" and "1"). They partitioned the 16 response types of Table 5.2 into these three equivalence classes that are invariant to recoding.

Technical Point 5.2

Monotonicity of causal effects. Consider a setting with a dichotomous treatment A and outcome Y. The value of the counterfactual outcome $Y^{a=0}$ is greater than that of $Y^{a=1}$ only among subjects of the "preventative" type. For the other 3 types, $Y^{a=1} \geq Y^{a=0}$ or, equivalently, an individual's counterfactual outcomes are monotonically increasing (i.e., nondecreasing) in a. Thus, when the treatment cannot prevent any subject's outcome (i.e., in the absence of "preventative" subjects), all individuals' counterfactual response types are monotonically increasing in a. We then simply say that the causal effect of A on Y is monotonic.

The concept of monotonicity can be generalized to two treatments A and E. The causal effects of A and E on Y are monotonic if every individual's counterfactual outcomes $Y^{a,e}$ are monotonically increasing in both a and e. That is, if there are no subjects with response types $\left(Y^{a=1,e=1} = 0, Y^{a=0,e=1} = 1\right)$, $\left(Y^{a=1,e=1} = 0, Y^{a=1,e=0} = 1\right)$, $\left(Y^{a=1,e=0} = 0, Y^{a=0,e=0} = 1\right)$, and $\left(Y^{a=0,e=1} = 0, Y^{a=0,e=0} = 1\right)$.

2. those who would develop the outcome under two treatment combinations, with the particularity that the effect of each treatment is exactly the opposite under each level of the other treatment (types 7 and 10)

3. those who would develop the outcome under three of the four treatment combinations (types 2, 3, 5, and 9)

On the other hand, the absence of additive interaction between A and E implies that either no individual in the population belongs to one of the three classes described above, or that there is a perfect cancellation of equal deviations from additivity of opposite sign. Such cancellation would occur, for example, if there were an equal proportion of individuals of types 7 and 10, or of types 8 and 12.

For more on cancellations that result in additivity even when interaction types are present, see Greenland, Lash, and Rothman (2008).

The meaning of the term "interaction" is clarified by the classification of individuals according to their counterfactual response types (see also Fine Point 5.1). We now introduce a tool to conceptualize the causal mechanisms involved in the interaction between two treatments.

5.4 Sufficient causes

The meaning of interaction is clarified by the classification of individuals according to their counterfactual response types. We now introduce a tool to represent the causal mechanisms involved in the interaction between two treatments. Consider again our heart transplant example with a single treatment A. As reviewed in the previous section, some individuals die when they are treated, others when they are not treated, others die no matter what, and others do not die no matter what. This variety of response types indicates that treatment A is not the only variable that determines whether or not the outcome Y occurs.

Take those individuals who were actually treated. Only some of them died, which implies that treatment alone is insufficient to always bring about the outcome. As an oversimplified example, suppose that heart transplant $A = 1$ only results in death in subjects allergic to anesthesia. We refer to the smallest set of background factors that, together with $A = 1$, are sufficient to inevitably produce the outcome as U_1. The simultaneous presence of treatment ($A = 1$)

Fine Point 5.1

More on counterfactual types and interaction. The classification of subjects by counterfactual response types makes it easier to consider specific forms of interaction. For example, we may be interested in learning whether some individuals will develop the outcome when receiving both treatments $E = 1$ and $A = 1$, but not when receiving only one of the two. That is, whether individuals with counterfactual responses $Y^{a=1,e=1} = 1$ and $Y^{a=0,e=1} = Y^{a=1,e=0} = 0$ (types 7 and 8) exist in the population. VanderWeele and Robins (2007a, 2008) developed a theory of sufficient cause interaction for 2 and 3 treatments, and derived the identifying conditions for synergism that are described here. The following inequality is a sufficient condition for these individuals to exist:

$$\Pr\left[Y^{a=1,e=1} = 1\right] - \left(\Pr\left[Y^{a=0,e=1} = 1\right] + \Pr\left[Y^{a=1,e=0} = 1\right]\right) > 0$$

or, equivalently, $\Pr\left[Y^{a=1,e=1} = 1\right] - \Pr\left[Y^{a=0,e=1} = 1\right] > \Pr\left[Y^{a=1,e=0} = 1\right]$

That is, in an experiment in which treatments A and E are randomly assigned, one can compute the three counterfactual risks in the above inequality, and empirically check that individuals of types 7 and 8 exist.

Because the above inequality is a sufficient but not a necessary condition, it may not hold even if types 7 and 8 exist. In fact this sufficient condition is so strong that it may miss most cases in which these types exist. A weaker sufficient condition for synergism can be used if one knows, or is willing to assume, that receiving treatments A and E cannot prevent any individual from developing the outcome, i.e., if the effects are monotonic (see Technical Point 5.2). In this case, the inequality

$$\Pr\left[Y^{a=1,e=1} = 1\right] - \Pr\left[Y^{a=0,e=1} = 1\right] > \Pr\left[Y^{a=1,e=0} = 1\right] - \Pr\left[Y^{a=0,e=0} = 1\right]$$

is a sufficient condition for the existence of types 7 and 8. In other words, when the effects of A and E are monotonic, the presence of superadditive interaction implies the presence of type 8 (monotonicity rules out type 7). This sufficient condition for synergism under monotonic effects was originally reported by Greenland and Rothman in a previous edition of their book. It is now reported in Greenland, Lash, and Rothman (2008).

In genetic research it is sometimes interesting to determine whether there are individuals of type 8, a form of interaction referred to as *compositional epistasis*. VanderWeele (2010) reviews empirical tests for compositional epistasis.

and allergy to anesthesia ($U_1 = 1$) is a minimal *sufficient cause* of the outcome Y.

Now take those individuals who were not treated. Again only some of them died, which implies that lack of treatment alone is insufficient to bring about the outcome. As an oversimplified example, suppose that no heart transplant $A = 0$ only results in death if subjects have an ejection fraction less than 20%. We refer to the smallest set of background factors that, together with $A = 0$, are sufficient to produce the outcome as U_2. The simultaneous absence of treatment ($A = 0$) and presence of low ejection fraction ($U_2 = 1$) is another sufficient cause of the outcome Y.

Finally, suppose there are some individuals who do not have neither U_1 nor U_2 and that would have developed the outcome whether they had been treated or untreated. The existence of these "doomed" individuals implies that there are some other background factors that are themselves sufficient to bring about the outcome. As an oversimplified example, suppose that all subjects with pancreatic cancer at the start of the study will die. We refer to the smallest set of background factors that are sufficient to produce the outcome regardless of treatment status as U_0. The presence of pancreatic cancer ($U_0 = 1$) is another sufficient cause of the outcome Y.

By definition of background factors, the dichotomous variables U cannot be intervened on, and cannot be affected by treatment A.

We described 3 sufficient causes for the outcome: treatment $A = 1$ in the presence of U_1, no treatment $A = 0$ in the presence of U_2, and presence

of U_0 regardless of treatment status. Each sufficient cause has one or more *components*, e.g., $A = 1$ and $U_1 = 1$ in the first sufficient cause. Figure 5.1 represents each sufficient cause by a circle and its components as sections of the circle. The term *sufficient-component causes* is often used to refer to the sufficient causes and their components.
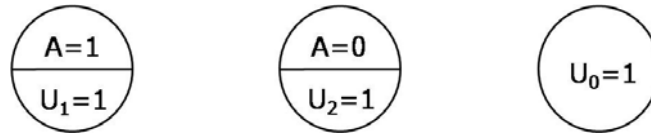
The graphical representation of sufficient-component causes helps visualize a key consequence of effect modification: as discussed in Chapter 4, the magnitude of the causal effect of treatment A depends on the distribution of effect modifiers. Imagine two hypothetical scenarios. In the first one, the population includes only 1% of individuals with $U_1 = 1$ (i.e., allergy to anesthesia). In the second one, the population includes 10% of individuals with $U_1 = 1$. The distribution of U_2 and U_0 is identical between these two populations. Now, separately in each population, we conduct a randomized experiment of heart transplant A in which half of the population is assigned to treatment $A = 1$. The average causal effect of heart transplant A on death will be greater in the second population because there are more subjects susceptible to develop the outcome if treated. One of the 3 sufficient causes, $A = 1$ plus $U_1 = 1$, is 10 times more common in the second population than in the first one, whereas the other two sufficient causes are equally frequent in both populations.

The graphical representation of sufficient-component causes also helps visualize an alternative concept of interaction, which is described in the next section. First we need to describe the sufficient causes for two treatments A and E. Consider our vitamins and heart transplant example. We have already described 3 sufficient causes of death: presence/absence of A (or E) is irrelevant, presence of transplant A regardless of vitamins E, and absence of transplant A regardless of vitamins E. In the case of two treatments we need to add 2 more ways to die: presence of vitamins E regardless of transplant A, and absence of vitamins regardless of transplant A. We also need to add four more sufficient causes to accommodate those who would die only under certain combination of values of the treatments A and E. Thus, depending on which background factors are present, there are 9 possible ways to die:

Greenland and Poole (1988) first enumerated these 9 sufficient causes.

1. by treatment A (treatment E is irrelevant)

2. by the absence of treatment A (treatment E is irrelevant)

3. by treatment E (treatment A is irrelevant)

4. by the absence of treatment E (treatment A is irrelevant)

5. by both treatments A and E

6. by treatment A and the absence of E

7. by treatment E and the absence of A

8. by the absence of both A and E

9. by other mechanisms (both treatments A and E are irrelevant)

In other words, there are 9 possible sufficient causes with treatment components $A = 1$ only, $A = 0$ only, $E = 1$ only, $E = 0$ only, $A = 1$ and $E = 1$, $A = 1$ and $E = 0$, $A = 0$ and $E = 1$, $A = 0$ and $E = 0$, and neither A or E matter. Each of these sufficient causes includes a set of background factors from $U_1,..., U_8$ and U_0. Figure 5.2 represents the 9 sufficient-component causes for two treatments A and E.
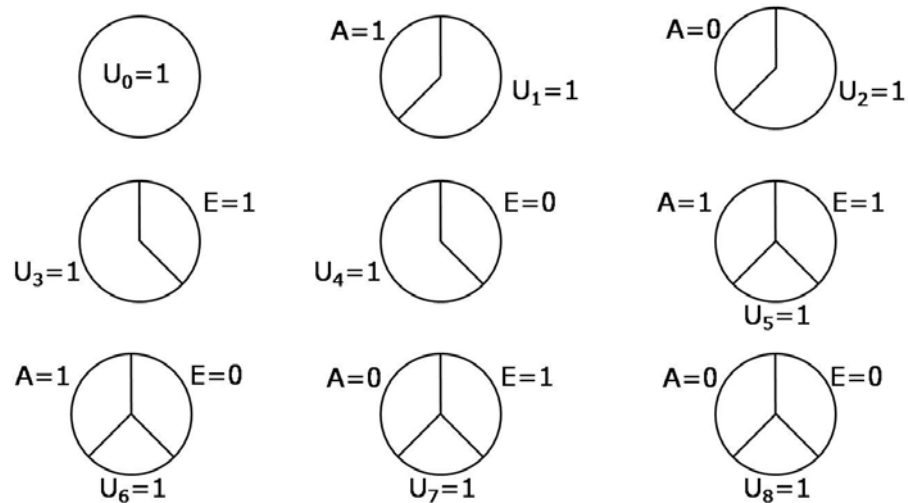
Figure 5.2

This graphical representation of sufficient-component causes is often referred to as "the causal pies."

Not all 9 sufficient-component causes for a dichotomous outcome and two treatments exist in all settings. For example, if receiving vitamins $E = 1$ does not kill any individual, regardless of her treatment A, then the 3 sufficient causes with the component $E = 1$ will not be present. The existence of those 3 sufficient causes would mean that some individuals (e.g., those with $U_3 = 1$) would be killed by receiving vitamins ($E = 1$), that is, their death would be prevented by not giving vitamins ($E = 0$) to them. Also note that some of the background factors U may be unnecessary. For example, if lack of vitamins and transplant were sufficient to bring about the outcome by themselves in some people, then the background factor U_8 in the last sufficient-component cause could be omitted.

5.5 Sufficient cause interaction

The colloquial use of the term "interaction between treatments A and E" evokes the existence of some causal mechanism by which the two treatments work together (i.e., "interact") to produce certain outcome. Interestingly, the definition of interaction within the counterfactual framework does not require any knowledge about those mechanisms nor even that the treatments work together (see Fine Point 5.3). In our example of vitamins E and heart transplant A, we said that there is an interaction between the treatments A and E if the causal effect of A when everybody receives E is different from the causal effect of A when nobody receives E. That is, interaction is defined by the contrast of counterfactual quantities, and can therefore be identified by conducting an ideal randomized experiment in which the conditions of exchangeability, positivity, and well-defined interventions hold for both treatments A and E. There

Fine Point 5.2

From counterfactuals to sufficient-component causes, and vice versa. There is a correspondence between the counterfactual response types and the sufficient component causes. In the case of a dichotomous treatment and outcome, suppose an individual has none of the background factors U_0, U_1, U_2. She will have an "immune" response type because she lacks the components necessary to complete all of the sufficient causes, whether she is treated or not. The table below displays the mapping between response types and sufficient-component causes in the case of one treatment A.

Type	$Y^{a=0}$	$Y^{a=1}$	Component causes
Doomed	1	1	$U_0 = 1$ or $\{U_1 = 1$ and $U_2 = 1\}$
Preventative	1	0	$U_0 = 0$ and $U_1 = 0$ and $U_2 = 1$
Causative	0	1	$U_0 = 0$ and $U_1 = 1$ and $U_2 = 0$
Immune	0	0	$U_0 = 0$ and $U_1 = 0$ and $U_2 = 0$

A particular combination of component causes corresponds to one and only one counterfactual type. However, a particular response type may correspond to several combinations of component causes. For the example, individuals of the "doomed" type may have any combination of component causes including $U_0 = 1$, no matter what the values of U_1 and U_2 are, or any combination including $\{U_1 = 1$ and $U_2 = 1\}$.

Sufficient-component causes can also be used to provide a mechanistic description of exchangeability $Y^a \coprod A$. For a dichotomous treatment and outcome, exchangeability means that the proportion of subjects who would have the outcome under treatment, and under no treatment, is the same in the treated $A = 1$ and the untreated $A = 0$. That is, $\Pr[Y^{a=1} = 1|A = 1] = \Pr[Y^{a=1} = 1|A = 0]$ and $\Pr[Y^{a=0} = 1|A = 1] = \Pr[Y^{a=0} = 1|A = 0]$.

Now the individuals who would develop the outcome if treated are the "doomed" and the "causative", that is, those with $U_0 = 1$ or $U_1 = 1$. The individuals who would get the outcome if untreated are the "doomed" and the "preventative", that is, those with $U_0 = 1$ or $U_2 = 1$. Therefore there will be exchangeability if the proportions of "doomed" + "causative" and of "doomed" + "preventative" are equal in the treated and the untreated. That is, exchangeability for a dichotomous treatment and outcome can be expressed in terms of sufficient-component causes as $\Pr[U_0 = 1$ or $U_1 = 1|A = 1] = \Pr[U_0 = 1$ or $U_1 = 1|A = 0]$ and $\Pr[U_0 = 1$ or $U_2 = 1|A = 1] = \Pr[U_0 = 1$ or $U_2 = 1|A = 0]$.

For additional details see Greenland and Brumback (2002), Flanders (2006), and VanderWeele and Hernán (2006). Some of the above results were generalized to the case of two or more dichotomous treatments by VanderWeele and Robins (2008).

is no need to contemplate the causal mechanisms (physical, chemical, biologic, sociological...) that underlie the presence of interaction.

This section describes a second concept of interaction that perhaps brings us one step closer to the causal mechanisms by which treatments A and E bring about the outcome. This second concept of interaction is not based on counterfactual contrasts but rather on sufficient-component causes, and thus we refer to it as interaction within the sufficient-component-cause framework or, for brevity, *sufficient cause interaction.*

A sufficient cause interaction between A and E exists in the population if A and E occur together in a sufficient cause. For example, suppose individuals with background factors $U_5 = 1$ will develop the outcome when jointly receiving vitamins ($E = 1$) and heart transplant ($A = 1$), but not when receiving only one of the two treatments. Then a sufficient cause interaction between A and E exists if there exists a subject with $U_5 = 1$. It then follows that if there exists a subject with counterfactual responses $Y^{a=1,e=1} = 1$ and $Y^{a=0,e=1} = Y^{a=1,e=0} = 0$, a sufficient cause interaction between A and E is present.

Sufficient cause interactions can be synergistic or antagonistic. There is *synergism* between treatment A and treatment E when $A = 1$ and $E = 1$

Fine Point 5.3

Biologic interaction. In epidemiologic discussions, sufficient cause interaction is commonly referred to as biologic interaction (Rothman et al, 1980). This choice of terminology might seem to imply that, in biomedical applications, there exist biological mechanisms through which two treatments A and E act on each other in bringing about the outcome. However, this may not be necessarily the case as illustrated by the following example proposed by VanderWeele and Robins (2007a).

Suppose A and E are the two alleles of a gene that produces an essential protein. Individuals with a deleterious mutation in both alleles ($A = 1$ and $E = 1$) will lack the essential protein and die within a week after birth, whereas those with a mutation in none of the alleles (i.e., $A = 0$ and $E = 0$) or in only one of the alleles (i.e., $A = 0$ and $E = 1$, $A = 1$ and $E = 0$) will have normal levels of the protein and will survive. We would say that there is synergism between the alleles A and E because there exists a sufficient component cause of death that includes $A = 1$ and $E = 1$. That is, both alleles work together to produce the outcome. However, it might be argued that they do not physically act on each other and thus that they do not interact in any biological sense.

Rothman (1976) described the concepts of synergism and antagonism within the sufficient-component-cause framework.

are present in the same sufficient cause, and *antagonism* between treatment A and treatment E when $A = 1$ and $E = 0$ (or $A = 0$ and $E = 1$) are present in the same sufficient cause. Alternatively, one can think of antagonism between treatment A and treatment E as synergism between treatment A and no treatment E (or between no treatment A and treatment E).

Unlike the counterfactual definition of interaction, sufficient cause interaction makes explicit reference to the causal mechanisms involving the treatments A and E. One could then think that identifying the presence of sufficient cause interaction requires detailed knowledge about these causal mechanisms. It turns out that this not always the case: sometimes we can conclude that sufficient cause interaction exists even if we lack any knowledge whatsoever about the sufficient causes and their components. Specifically, if the inequalities in Fine Point 5.1 hold, then there exists synergism between A and E. That is, one can empirically check that synergism is present without ever giving any thought to the causal mechanisms by which A and E work together to bring about the outcome. This result is not that surprising because of the correspondence between counterfactual response types and sufficient causes (see Fine Point 5.2), and because the above inequality is a sufficient but not a necessary condition, i.e., the inequality may not hold even if synergism exists.

5.6 Counterfactuals or sufficient-component causes?

A counterfactual framework of causation was already hinted by Hume (1748).

The sufficient-component-cause framework and the counterfactual (potential outcomes) framework address different questions. The sufficient component cause model considers sets of actions, events, or states of nature which together inevitably bring about the outcome under consideration. The model gives an account of the causes of a particular effect. It addresses the question, "Given a particular effect, what are the various events which might have been its cause?" The potential outcomes or counterfactual model focuses on one particular cause or intervention and gives an account of the various effects of that cause. In contrast to the sufficient component cause framework, the potential outcomes framework addresses the question, "What would have occurred if a particular factor were intervened upon and thus set to a different level than it in fact

Technical Point 5.3

Monotonicity of causal effects and sufficient causes. When treatment A and E have monotonic effects, then some sufficient causes are guaranteed not to exist. For example, suppose that cigarette smoking ($A = 1$) never prevents heart disease, and that physical inactivity ($E = 1$) never prevents heart disease. Then no sufficient causes including either $A = 0$ or $E = 0$ can be present. This is so because, if a sufficient cause including the component $A = 0$ existed, then some individuals (e.g., those with $U_2 = 1$) would develop the outcome if they were unexposed ($A = 0$) or, equivalently, the outcome could be prevented in those individuals by treating them ($A = 1$). The same rationale applies to $E = 0$. The sufficient component causes that cannot exist when the effects of A and E are monotonic are crossed out in Figure 5.3.

The sufficient-component-cause framework was developed in philosophy by Mackie (1965). He introduced the concept of *INUS* condition for Y: an *I*nsufficient but *N*ecessary part of a condition which is itself *U*nnecessary but exclusively *S*ufficient for Y.

was?" Unlike the sufficient component cause framework, the counterfactual framework does not require a detailed knowledge of the mechanisms by which the factor affects the outcome.

The counterfactual approach addresses the question "what happens?" The sufficient-component-cause approach addresses the question "how does it happen?" For the contents of this book—conditions and methods to estimate the average causal effects of hypothetical interventions—the counterfactual framework is the natural one. The sufficient-component-cause framework is helpful to think about the causal mechanisms at work in bringing about a particular outcome. Sufficient-component causes have a rightful place in the teaching of causal inference because they help understand key concepts like the dependence of the magnitude of causal effects on the distribution of background factors (effect modifiers), and the relationship between effect modification, interaction, and synergism.

Though the sufficient-component-cause framework is useful from a pedagogic standpoint, its relevance to actual data analysis is yet to be determined. In its classical form, the sufficient-component-cause framework is deterministic, its conclusions depend on the coding on the outcome, and is by definition limited to dichotomous treatments and outcomes (or to variables that can be recoded as dichotomous variables). This limitation practically rules out the consideration of any continuous factors, and restricts the applicability of the framework to contexts with a small number of dichotomous factors. However, recent extensions of the sufficient-component-cause framework to stochastic settings and to categorical and ordinal treatments may lead to an increased application of this approach to realistic data analysis. Finally, even allowing for recent extensions of the sufficient-component-cause framework, we may rarely have the large amount of data needed to study the fine distinctions it makes.

To estimate causal effects more generally, the counterfactual framework will likely continue to be the one most often employed. Some apparently alternative frameworks—causal diagrams, decision theory—are essentially equivalent to the counterfactual framework, as described in the next chapter.

Fine Point 5.4

More on the attributable fraction. Fine Point 3.1 defined the excess fraction for treatment A as the proportion of cases attributable to treatment A in a particular population, and described an example in which the excess fraction for A was 75%. That is, 75% of the cases would not have occurred if everybody had received treatment $a = 0$ rather than their observed treatment A. Now consider a second treatment E. Suppose that the excess fraction for E is 50%. Does this mean that a joint intervention on A and E could prevent 125% ($75\% + 50\%$) of the cases? Of course not.

Clearly the excess fraction cannot exceed 100% for a single treatment (either A or E). Similarly, it should be clear that the excess fraction for any joint intervention on A and E cannot exceed 100%. That is, if we were allowed to intervene in any way we wish (by modifying A, E, or both) in a population, we could never prevent a fraction of disease greater than 100%. In other words, no more than 100% of the cases can be attributed to the lack of certain intervention, whether single or joint. But then why is the sum of excess fractions for two single treatments greater than 100%? The sufficient-component-cause framework helps answer this question.

As an example, suppose that Zeus had background factors $U_5 = 1$ (and none of the other background factors) and was treated with both $A = 1$ and $E = 1$. Zeus would not have been a case if either treatment A or treatment E had been withheld. Thus Zeus is counted as a case prevented by an intervention that sets $a = 0$, i.e., Zeus is part of the 75% of cases attributable to A. But Zeus is also counted as a case prevented by an intervention that sets $e = 0$, i.e., Zeus is part of the 50% of cases attributable to E. No wonder the sum of the excess fractions for A and E exceeds 100%: some individuals like Zeus are counted twice!

The sufficient-component-cause framework shows that it makes little sense to talk about the fraction of disease attributable to A and E separately when both may be components of the same sufficient cause. For example, the discussion about the fraction of disease attributable to either genes or environment is misleading. Consider the mental retardation caused by phenylketonuria, a condition that appears in genetically susceptible individuals who eat certain foods. The excess fraction for those foods is 100% because all cases can be prevented by removing the foods from the diet. The excess fraction for the genes is also 100% because all cases would be prevented if we could replace the susceptibility genes. Thus the causes of mental retardation can be seen as either 100% genetic or 100% environmental. See Rothman, Greenland, and Lash (2008) for further discussion.
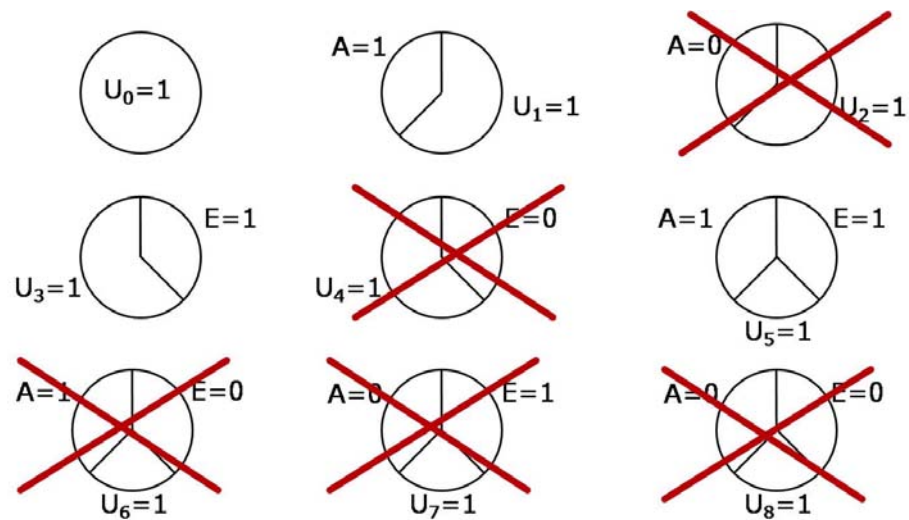
Figure 5.3

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

The modern theory of diagrams for causal inference arose within the disciplines of computer science and artificial intelligence. Comprehensive books on this subject have been written by Pearl (2009) and Spirtes, Glymour and Scheines (2000).

Richardson and Robins (2013) have recently developed a new causal graph—the Single World Intervention Graph (SWIG)—that seamlessly unifies the counterfactual and graphical approaches to causality 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.

This chapter describes graphs, which we will refer to as causal diagrams, to represent key causal concepts. This and the next three chapters are focused on problem conceptualization via causal diagrams. We will use causal diagrams to classify sources of systematic bias and to identify potential problems in study design and analysis. 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. As a result, no estimator can be consistent (see Chapter 1 for a definition of consistent estimator). Chapters 7, 8, and 9 are devoted to three types of systematic bias: confounding, selection bias, and measurement bias, respectively.

The graphical approach to bias has generally been found to be easier to use and more intuitive than the counterfactual approach. However, the two approaches are intimately linked. Specifically, associated with each graph is an underlying counterfactual model. 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 undelying counterfactual variables on the graph. Therefore the the link between graphs and counterfactuals has remained hidden.

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.

The presence of an arrow pointing from a particular variable V to another variable W indicates either that we know there is a direct causal effect (i.e., an

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_m of $V_3 = Y$ are (L, A). We will adopt the notational convention that if $m > j$, V_m is not an ancestor of V_j.

A causal DAG is a DAG in which 1) the lack of an arrow from node V_j to 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. 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

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

L $\longrightarrow$ A $\longrightarrow$ Y
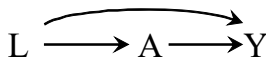
Figure 6.1

effect not mediated through any other variables on the graph) for at least one individual, or that we are unwilling to assume such individual causal effects do not exist. Alternatively, the lack of an arrow means that we know, or are willing to assume, 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 either we know that disease severity affects the probability of receiving a heart transplant or that we are not willing to assume otherwise. 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 we focus on *causal* directed acyclic graphs. Informally, a directed acyclic graph is causal if the common causes of any pair of variables in the graph are also in the graph. 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 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.
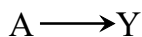
Figure 6.1 may also represent an observational study. Specifically, Figure

A $\longrightarrow$ Y

Figure 6.2

Technical Point 6.2

Counterfactual models associated with a causal DAG. 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 W denote the support (i.e., the set of possible values w) of W. For any $w_1, \ldots, w_m$, 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 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. 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 cause of V_3.

Robins (1986) called this NPSEM a finest causally interpreted structural tree graph (FCISTGs). Pearl (2000) showed how to represent this model with a DAG under the assumption that every variable on the graph is subject to intervention with well-defined causal effects. Robins (1986) also proposed more realistic CISTGs in which only a subset of the variables are subject to intervention. For expositional purposes, we will assume that every variable can be intervened on, even though the statistical methods considered here do not actually require this assumption.

A FCISTG model does not imply that the causal Markov assumption holds; additional statistical independence assumptions are needed. For example, Pearl (2000) 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 the one-step ahead counterfactuals $V_m^{\overline{v}_{m-1}} = f_m(pa_m, \epsilon_m)$ and $V_j^{\overline{v}_{j-1}} = f_j(pa_j, \epsilon_j)$, $j < m$, are jointly independent when $\overline{v}_{j-1}$ is a subvector of the $\overline{v}_{m-1}$, and referred to this as the finest fully randomized causally interpreted structured tree graph (FFRCISTG) model, which was introduced in Chapter 2. Robins (1986) showed this assumption implies that the causal Markov assumption holds. An NPSEM-IE is an FFRCISTGs but not vice-versa because an NPSEM-IE makes stronger assumptions than an FFRCISTG (Robins and Richardson 2010).

A DAG represents an NPSEM but we need to specify which type. For example, the DAG in Figure 6.2 may correspond to either an NPSEM-IE that implies full exchangeability $(Y^{a=0}, Y^{a=1}) \amalg A$, or to an FFRCISTG that only implies marginal exchangeability $Y^a \amalg A$ for both $a = 0$ and $a = 1$. In this book we assume that DAGs represent FFRCISTGs.

6.1 represents an observational study in which we are willing to assume that the assignment of heart transplant A depends on disease severity L *and on 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, $Y^a \amalg A|L$ for all a.

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.
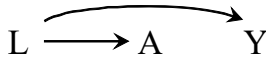
6.2 Causal diagrams and marginal independence

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., $\Pr[Y^{a=1} = 1] \neq \Pr[Y^{a=0} = 1]$. 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.

$L \longrightarrow A \qquad Y$

Figure 6.3

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 graph 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 graph 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.

A path between two variables R and S in a DAG is a route that connects R and S by following a sequence of (nonintersecting) edges. A path is causal if it consists entirely of edges with their arrows pointing in the same direction. Otherwise it is noncausal.

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. Graph theory again confirms our intuition. In graphic

---

Influence diagrams. An alternative approach to causal inference is based on decision theory (Dawid 2000, 2002). This decision-theoretic approach employs a notation that makes no reference to counterfactuals and uses causal diagrams augmented with decision nodes to represent the interventions of interest. Though the 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 were always explicit about the potential interventions on the variable A, the additional nodes (to represent the potential interventions) would be somewhat redundant.
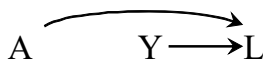
Figure 6.4

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. In graph theory the common effect L is referred to as a *collider* on the path $A - L - Y$ because two arrowheads collide on this node.

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). He 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 $A \amalg Y$. The knowledge that both A and Y cause heart disease L is irrelevant when considering the association between A and Y. Graph 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

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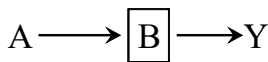
associated because aspirin has a causal effect on heart disease. Now suppose we obtain an additional piece of information: aspirin A affects the risk of death Y because it reduces platelet aggregation B. This new knowledge is translated into the causal diagram of Figure 6.5 that shows platelet aggregation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

Figure 6.5

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, $A \amalg Y|B$. Indeed graph theory states that a box placed around variable B blocks the flow of association through the path $A \rightarrow B \rightarrow Y$.

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.

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 ($L = 1$). In that case, learning that an individual carries a lighter ($A = 0$) 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, $A \amalg Y|L$. 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.

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.

Finally, consider Figure 6.4 again. We concluded in the previous section that having the haplotype A was independent of being a cigarette smoker 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 subjects with heart
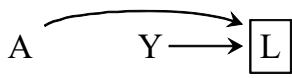
Figure 6.7

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

Figure 6.8

The mathematical theory underlying the graphical rules is known as "d-separation" (Pearl 1995). See Fine Point 6.3.

Figure 6.9

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

disease ($L = 1$). The square around L in Figure 6.7 indicates that they are conditioning on a particular value of L. Knowing that a subject 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. Graph 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.
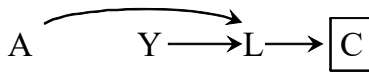
As another example, the causal diagram in Figure 6.8 adds to that in Figure 6.7 a diuretic medication C whose use is a consequence of a diagnosis of heart disease. A and Y are also associated within levels of C because C is a common effect of A and Y. Graph theory shows that conditioning on a variable 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. 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 6.2 for a systematic summary of the graphical rules.
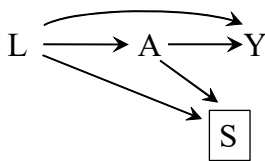
There is another possible source of association between two variables that we have not discussed yet: chance or random variability. Unlike the structural reasons for an association between two variables—causal effect of one on the other, shared common causes, conditioning on common effects—random variability 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 Graphs, counterfactuals, and interventions

Causal diagrams encode qualitative expert knowledge, or assumptions, about the causal structure of a problem and hence about the causal determinant of biases. Though causal diagrams are a useful tool to think conceptually about a causal inference problem, quantitative approaches are needed to compute causal effects. The identification formulas for the effects of interventions given in Chapter 2 can also be derived using the tools of graph theory. Therefore our choice of counterfactual theory in Chapters 1-5 did not really privilege one particular approach but only one particular notation. See also Fine Point 6.1.

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 from A to 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. We would generally expect that in a setting represented by Figure 6.2 there is both an average causal effect of A on Y, $\Pr[Y^{a=1} = 1] \neq \Pr[Y^{a=0} = 1]$, and an association between A and Y, $\Pr[Y = 1|A = 1] \neq \Pr[Y = 1|A = 0]$. 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.

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 can be 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 betweem them are blocked, but might also be independent even if some paths were open. In this book we will assume faithfulness unless we say otherwise.

The causal diagrams in this chapter include the treatment A, the outcome Y, variables that are conditioned on, any other measured variables that are necessary to achieve conditional exchangeability (see Chapter 7), and common causes (whether measured or unmeasured) of any of the above variables. Not all these variables are created equal. For causal inference purposes, one needs to differentiate between variables that are and are not potentially intervened upon.

We have made this distinction throughout the book. For example, the tree graphs introduced in Chapter 2 have a circle around branchings corresponding to nontreatment variables L and Y; our discussion on well-defined interventions in Chapter 3 imposes requirements on the treatment variables A that do not apply to other variables like, say, L; the "pies" representing sufficient causes in Chapter 5 distinguish between potential treatments A and E and background factors U; etc. In contrast, causal diagrams seems to assign the same status to all variables in the diagram—in fact this is the case when causal diagrams are considered as representations of nonparametric structural equations models as

described in Technical Point 6.2.

The apparently equal status of all variables in causal diagrams may be misleading, especially when some of those variables are ill-defined: it may be okay to draw a causal diagram that includes a node for a vaguely defined unmeasured confounder, but it is critical that the nodes for treatment variables are precisely defined so that multiple versions do not exist (see Fine Point 1.2).

For example, suppose that we are interested in the causal effect of the dichotomous treatment R, where $R = 1$ is defined as "exercising at least 30 minutes daily," and $R = 0$ is defined as "exercising less than 30 minutes daily." Individuals who exercise longer than 30 minutes will be classified as $R = 1$, and thus each of the possible durations $30, 31, 32...$ minutes can be viewed as a different version of the treatment $R = 1$. More formally, let $A(r)$ be a version of treatment $R = r$. For each individual with $R = 1$ in the study $A(r = 1)$ can take values $30, 31, 32, ...$ indicating all possible durations of exercise greater or equal than 30 minutes. For each individual with $R = 0$ in the study $A(r = 0)$ can take values $0, 1, 2..., 29$ including all durations of less than 30 minutes. We refer to R as a *compound treatment* because multiple values $a(r)$ can be mapped onto a single value $R = r$. Figure 6.10 shows a causal diagram that includes both the compound treatment R (the decision node), its versions A—a vector including all the variables $A(r)$—, two sets of common causes L and W, and unmeasured common causes U. Being explicit about the compound treatment R of interest and its versions $A(r)$ is an important step towards a well defined causal effect and the identification of confounders.
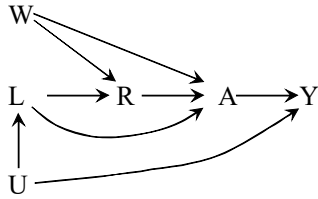
Figure 6.10

6.5 A structural classification of bias

We begin by defining bias due to structural reasons—systematic bias—first for conditional effects, then for marginal effects. For the average causal effects within levels of L, there is conditional bias whenever $\Pr[Y^{a=1}|L = l] - \Pr[Y^{a=0}|L = l]$ differs from $\Pr[Y|L = l, A = 1] - \Pr[Y|L = l, A = 0]$ for at least one stratum l. That is, there is bias whenever the effect measure (e.g., causal risk ratio or difference) and the corresponding association measure (e.g., associational risk ratio or difference) are not equal. As discussed in Section 2.3, conditional exchangeability $Y^a \amalg A|L$ implies the absence of conditional bias. The converse is also true: absence of conditional bias implies conditional exchangeability.

For the average causal effect in the entire population, we say there is (unconditional) bias when $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1] \neq \Pr[Y = 1|A = 1] - \Pr[Y = 1|A = 0]$. Absence of conditional bias implies that we can obtain an unbiased estimate of the average causal effect in the entire population by, say, standardization.

Under faithfulness, the presence of conditional bias implies the presence of unconditional bias since without faithfulness positive bias in one stratum of L might exactly cancel the negative bias in another.

When the null hypothesis of no causal effect of treatment on the outcome holds, but treatment and outcome are associated in the data, we say that there is *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.

Bias under the null can result from two different causal structures:

Bias may also result from (non-structural) random variability. Seel Chapter 10.

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

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

There is another possible source of bias under the null: 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.*

Confounding, selection bias, and measurement bias are described in more detail in Chapters 7, 8, and 9, respectively. Any causal structure that results in bias under the null will also cause bias under the alternative (i.e.,when treatment has an effect on the outcome). However, the converse is not true. For example, conditioning on a descendant of Y may cause bias under the alternative but not under the null. Further as discussed in Chapter 9, some forms of measurement error will cause bias under the alternative but not under the null. In general, we will refer to bias as any structural association between treatment and outcome that does not arise from the causal effect of treatment on outcome. Causal diagrams are helpful to represent different sources of association and thus to sharpen discussions about bias.

The three types of bias—confounding, selection, measurement—may arise in observational studies, but also in randomized experiments. This may not be obvious from previous chapters, in which we conceptualized observational studies as some sort of imperfect randomized experiments, whereas randomized experiments like the one represented in Figure 6.2 were presented as ideal studies in which no participants are lost during the follow-up, all participants adhere to the assigned treatment, and the assigned treatment remains unknown to both study participants and investigators. We might as well have told you a fairy tale or a mythological story. Real randomized experiments rarely look like that. The remaining chapters of Part I will elaborate on the sometimes fuzzy boundary between experimenting and observing. Specifically, in the next three chapters we turn our attention to the use of causal diagrams to represent three classes of biases: bias due to the presence of common causes, bias due to the selection of individuals, and bias due to the measurement of variables. 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.

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
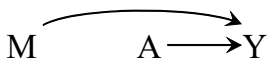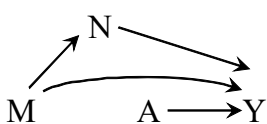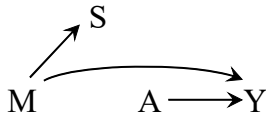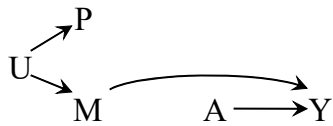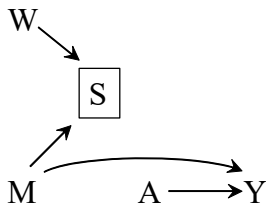

Figure 6.11

Figure 6.12

classify all individuals as receiving high ($M = 1$) or normal ($M = 0$) quality of care, compute the stratified risk differences in each level of M as described in Chapter 4, and indeed confirm that there is effect modification by M on the additive scale. The causal diagram in Figure 6.11 includes the effect modifier M with an arrow into the outcome Y but no arrow into treatment A (which is randomly assigned and thus independent of M). Two important caveats.

First, the causal diagram in Figure 6.11 would still be a valid causal diagram if it did not include M because M is not a common cause of A and Y. It is only because the causal question makes reference to M (i.e., what is the average causal effect of A on Y *within levels of* M?), that M needs to be included in the causal diagram. Other variables measured along the path between "quality of care" M and the outcome Y could also qualify as effect modifiers. For example, Figure 6.12 shows the effect modifier "therapy complications" N, which partly mediates the effect of M on Y.

Second, the causal diagram in Figure 6.11 does not necessarily indicate the presence of effect modification by M. The causal diagram implies that both A and M affect death Y, but it does not distinguish among the following three qualitatively distinct ways that M 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 $M = 1$ and stratum $M = 0$.

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

3. Treatment A has a causal effect on Y in one stratum of M but no causal effect in the other stratum, e.g., A only kills subjects with $M = 0$.

That is, Figure 6.11—as well as all the other figures discussed in this section—is equally valid to depict a setting with or without effect modification by M.

In the above example, the effect modifier M 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 6.13 includes the variable "cost of the treatment" S (1: high, 0: low), which is affected by "quality of care" M but has itself no effect on mortality Y. An analysis stratified by S will generally detect effect modification by S even though the variable that truly modifies the effect of A on Y is M. The variable S is a *surrogate effect modifier* whereas the variable M 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 6.13 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 6.14 includes the variables

Figure 6.13

Figure 6.14

Figure 6.15

See VanderWeele and Robins (2007b) for a finer classification of effect modification via causal diagrams.

"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 M and nationality P. Thus P will behave as a surrogate effect modifier because P is associated with the causal effect modifier M. Another example: Figure 6.15 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 M, 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 M 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.

Fine Point 6.3

D-separation. We now define a graphical relationship between variables on a DAG known as d-separation ('d-' stands for directional). The importance of d-separation is the following result: Given a DAG G and a distribution over its nodes, suppose each variable is independent of its non-descendants conditional on its parents. Then if the two sets of variables are d-separated given a third set, the two sets are conditionally independent given the third (i.e., independent within every joint stratum of the third variables).

To define d-separation, we first define the terms "path" and "blocked path." A path is a sequence of edges connecting two variables on the graph (with each edge occurring only once). 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. For example, 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 noncollider that has been conditioned on is blocked. For example, 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. For example, 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. For example, in Figure 6.8, the path between A and Y is open after conditioning on S, 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 noncollider 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 said to be d-separated if all paths between them are blocked (otherwise they are d-connected). Two sets of variables are said to be 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 marginally independent (d-connected) 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 marginally independent (d-separated) because the only path between them is blocked by the collider L. In Figure 6.5, we conclude that A is conditionally independent of Y, given B. From Figure 6.7 we infer that A is not conditionally independent of Y, given L, and from Figure 6.8 we infer that A is not conditionally independent of Y, given S.

The d-separation rules to infer associational statements from causal diagrams were formalized by Pearl (1995). A mathematically equivalent set of graphical rules, known as "moralization", was developed by Lauritzen et al. (1990).

Chapter 7
CONFOUNDING

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

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

7.1 The structure of confounding

$$L \longrightarrow A \longrightarrow Y$$

Figure 7.1

In a causal DAG, a backdoor path is a noncausal path between treatment and outcome that remains even if all arrows pointing from treatment to other variables (in graph-theoretic terms, the descendants of treatment) are removed. That is, the path has an arrow pointing into treatment.
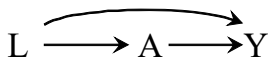
Confounding is the bias that arises when the treatment and the outcome share a cause. The structure of confounding can be represented by using causal diagrams. For example, the diagram in Figure 7.1 (same as Figure 6.1) depicts a treatment A, an outcome Y, and their shared (or common) cause L. This diagram shows two sources of association between treatment and outcome: 1) the path $A \rightarrow Y$ that represents the causal effect of A on Y, and 2) the path $A \leftarrow L \rightarrow Y$ between A and Y that is mediated by the common cause L. In graph theory, the path $A \leftarrow L \rightarrow Y$ that links A and Y through their common cause L is an example of a *backdoor path*.

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

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

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

Figure 7.2

Figure 7.3

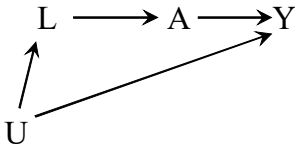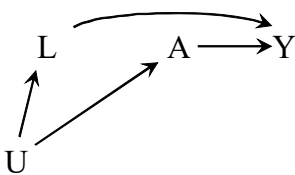bias, depicted in the causal diagram in Figure 7.1, is often referred to as a *healthy worker bias*.

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

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

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

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

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

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

In all these cases, the bias has the same structure: it is due to the presence of a common cause (L or U) of the treatment A and the outcome Y or, equivalently, to the presence of an unblocked backdoor path between A and Y. We refer to the bias caused by common causes as confounding, and we use other names to refer to biases caused by structural reasons other than the presence of common causes. For example, we say that selection bias is the result of conditioning on common effects. For simplicity of presentation, we assume throughout this chapter that other sources of bias (e.g., selection bias, measurement error, and random variability) are absent.

7.2 Confounding and identifiability of causal effects

Pearl (1995) proposed the backdoor criterion for nonparametric identification of causal effects. All backdoor paths are blocked if treatment an outcome are d-separated given the measured covariates in a graph in which the arrows out of A are removed.

With confounding structurally defined as the bias resulting from the presence of common causes of treatment and outcome, the next question is: under what conditions can confounding be eliminated in the analysis? In other words, in the absence of measurement error and selection bias, under what conditions can the causal effect of treatment A on outcome Y be identified? An important result from graph theory, known as *the backdoor criterion*, is that the causal effect of treatment A on the outcome Y is identifiable if all backdoor paths between them can be blocked by conditioning on variables that are not affected by—non-descendants of—treatment A.

Thus the two settings in which causal effects are identifiable are

1. *No common causes.* If, like in Figure 6.2, there are no common causes of treatment and outcome, and hence no backdoor paths that need to be blocked, we say that there is no confounding.

2. *Common causes but enough measured variables (that are non-descendants of treatment) to block all backdoor paths.* If, like in Figure 7.1, the backdoor path through the common cause L can be blocked by conditioning on some measured covariates (in this example, L itself), we say that there is confounding but no unmeasured confounding.

The first setting is expected in marginally randomized experiments in which all subjects have the same probability of receiving treatment. In these experiments confounding is not expected to occur because treatment is solely determined by the flip of a coin—or its computerized upgrade, the random number generator—and the flip of the coin cannot be a cause of the outcome.

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

The second setting is expected in conditionally randomized experiments in which the probability of receiving treatment is the same for all subjects with the same value of risk factor L but, by design, this probability varies across values of L. The design of these experiments guarantees the presence of confounding, because L is a common cause of treatment and outcome, but in these experiments confounding is not expected conditional on—within levels of—the covariates L. This second setting is also what one hopes for in observational studies in which many variables L have been measured.

The backdoor criterion answers three questions: 1) does confounding exist?, 2) can confounding be eliminated?, and 3) what variables are necessary to eliminate the confounding? The answer to the first question is affirmative if there exist unblocked backdoor paths between treatment and outcome; the answer to the second question is affirmative if all those backdoor paths can be blocked using the measured variables; the answer to the third question is any minimal set of variables that, when conditioned on, block all backdoor paths. The backdoor criterion, however, does not answer questions regarding the magnitude or direction of confounding (see Fine Point 7.3 for more on this topic). It is logically possible that some unblocked backdoor paths are weak (e.g., if L does not have a large effect on either A or Y) and thus induce little bias, or that several strong backdoor paths induce bias in opposite directions and thus result in a weak net bias.

In Chapter 4 we described how the causal effect of interest can be identified—via standardization or IP weighting—in the presence of confounding when the appropriate variables are measured. The variables that are used to standardize or IP weight are often referred to as *confounders*. We now review the definition of confounder and some criteria to select them.

7.3 Confounders

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

Confounding is the bias that results from the presence of common causes of—open backdoor paths between—treatment A and outcome Y. It is then natural to define a confounder as a variable that, possibly in conjunction with other variables, can be used to block all backdoor paths between treatment and outcome. To provide a more formal definition of confounder, we first need to define a *sufficient set for confounding adjustment* as a set of non-descendants of treatment $L = (C, F)$ that includes enough variables to block all otherwise open backdoor paths. Then the variable C is a confounder given data on the variables F if $L = (C, F)$ is a sufficient set but F, or any subset of F, is not. If F is the empty set, then we simply say that $L = C$ is a confounder.

In contrast with this structural—or counterfactual—definition, a confounder was traditionally defined as any variable that meets the following three conditions: 1) it is associated with the treatment, 2) it is associated with the outcome conditional on the treatment (with "conditional on the treatment" often replaced by "in the untreated"), and 3) it does not lie on a causal pathway between treatment and outcome. According to this traditional definition, all so defined confounders should be adjusted for in the analysis. However, this traditional definition of confounder may lead to inappropriate adjustment for confounding. To see why, let us compare the structural and traditional definitions of confounder in Figures 7.1-7.4. For simplicity, these four figures depict settings in which investigators need no data beyond the measured variables L for confounding adjustment (with F being the empty set), and in which the variables L are affected by neither the treatment A nor the outcome Y.

In Figure 7.1 there is confounding because the treatment A and the outcome Y share the cause L, i.e., because there is a backdoor path between A and Y through L. However, this backdoor path can be blocked by conditioning on L. Thus, if the investigators collected data on L for all individuals, there is no unmeasured confounding given L. We say that L is a confounder because it is needed to eliminate confounding. Let us now turn to the traditional definition of confounder. The variable L is associated with the treatment (because it has a causal effect on A), is associated with the outcome conditional on the treatment (because it has a direct effect on Y), and it does not lie on the causal pathway between treatment and outcome. Then, according to the traditional definition, L is a confounder and it should be adjusted for. There is no discrepancy between the structural and traditional definitions of confounder under the causal diagram in Figure 7.1.

In Figure 7.2 there is confounding because the treatment A and the outcome Y share the cause U, i.e., there is a backdoor path between A and Y through U. (Unlike the variables L, A, and Y, we suppose that the variable U was not measured by the investigators.) This backdoor path could be theoretically blocked, and thus confounding eliminated, by conditioning on U, had data on this variable been collected. However, this backdoor path can also be blocked by conditioning on L. Thus, there is no unmeasured confounding given L. We say that L is a confounder because it is needed to eliminate confounding, even though the confounding resulted from the presence of U. Let us now turn to the traditional definition of confounder. The variable L is associated with the treatment (because it has a causal effect on A), is associated with the outcome conditional on the treatment (because it shares the cause U with Y), and it does not lie on the causal pathway between treatment and outcome. Then, according to the traditional definition, L is a confounder and it should be adjusted for. Again, there is no discrepancy between the structural and
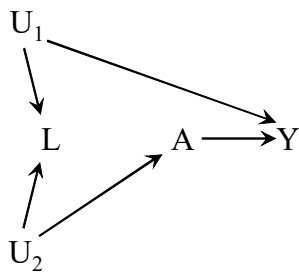
Figure 7.4

The bias induced in Figure 7.4 was described by Greenland et al (1999), and referred to as M-bias (Greenland 2003) because the structure of the variables involved in it—U_2, L, U_1—resembles a letter M lying on its side.
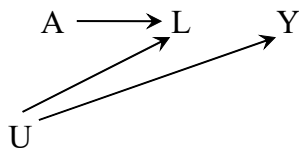

Figure 7.5

Figure 7.5 is another example in which, in the absence of confounding, the traditional criteria lead to selection bias due to adjustment for L. The traditional criteria would not have resulted in bias had condition 3) been replaced by the condition that the variable is not caused by treatment, i.e., it is a non-descendant of A.

traditional definitions of confounder in Figure 7.2.

In Figure 7.3 there is also confounding because the treatment A and the outcome Y share the cause U, and the backdoor path can also be blocked by conditioning on L. Therefore there is no unmeasured confounding given L, and we say that L is a confounder. According to the traditional definition, L is also a confounder and should be adjusted for because L is associated with the treatment (it shares the cause U with A), is associated with the outcome conditional on the treatment (it has a causal effect on Y), and it does not lie on the causal pathway between treatment and outcome. Again, there is no discrepancy between the structural and traditional definitions of confounder for the causal diagram in Figure 7.3.

The key figure is Figure 7.4. In this causal diagram there are no common causes of treatment A and outcome Y, and therefore there is no confounding. The backdoor path between A and Y through L ($A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$) is blocked because L is a collider on that path. Thus all the association between A and Y is due to the effect of A on Y: association is causation. There is no need to adjust for L. (Adjustment for either U_1 or U_2 is impossible, as these are unmeasured variables.) In fact, adjustment for L by stratification would induce bias because conditioning on L would open the otherwise blocked backdoor path between A and Y. This implies that, although there is no unconditional bias, there is conditional bias for at least one stratum of L. We refer to this bias as selection bias because it arises from selecting a particular stratum of L in which the association between A and Y is calculated.

Though there is no confounding, L meets the criteria for a traditional confounder: it is associated with the treatment (it shares the cause U_2 with A), it is associated with the outcome conditional on the treatment (it shares the cause U_1 with Y), and it does not lie on the causal pathway between treatment and outcome. Hence, according to the traditional definition, L is considered a confounder that should be adjusted for, even in the absence of confounding!

The result of trying to adjust for the nonexistent confounding would be selection bias. For example, suppose A represents physical activity, Y cervical cancer, U_1 a pre-cancer lesion, L a diagnostic test (Pap smear) for pre-cancer, and U_2 a health-conscious personality (more physically active, more visits to the doctor). Then, under the causal diagram in Figure 7.4, the effect of physical activity A on cancer Y is unconfounded and there is no need to adjust for L. But let us say that one decides to adjust for L by, for example, restricting the analysis to women with a negative test ($L = 0$). Conditioning on the collider L opens the backdoor path between A and Y ($A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$), which was previously blocked by the collider itself. Thus the association between A and Y would be a mixture of the association due to the effect of A on Y and the association due to the open backdoor path. Association would not be causation any more.

We have described an example in which the standard definition of confounder fails because it misleads investigators into adjusting for a variable when adjustment for such variable is not only superfluous but also harmful. This problem arises because the standard definition treats the concept of confound*er*, rather than that of confound*ing*, as the primary concept. In contrast, the structural definition first establishes the presence of confounding—common causes—and then identifies the confounders that are necessary to adjust for confounding in the analysis. Confounding is an absolute concept—common causes of treatment and outcome either exist or do not exist in a particular region of the universe—whereas confounder is a relative one—L may be needed to block a backdoor path only when U is not measured.

Fine Point 7.1

Surrogate confounders and time-varying confounders. Consider now the causal diagram in Figure 7.6. There is confounding for the effect of A on Y because of the presence of the unmeasured common cause U. The measured variable L is a proxy or surrogate for U. For example, the unmeasured variable socioeconomic status U may confound the effect of physical activity A on the risk of cardiovascular disease Y. Income L is a surrogate for the often ill-defined variable socioeconomic status. Should we adjust for the variable L? On the one hand, L is not a confounder (it does not lie on a backdoor path between A and Y). On the other hand, adjusting for the measured L, which is associated with the unmeasured U, may indirectly adjust for some of the confounding caused by U. In the extreme, if L were perfectly correlated with U then it might make no difference whether one conditions on L or on U. Indeed if L is binary and is a nondiferentially missclassified (see Chapter 9) version of U, conditioning on L will result in a partial blockage of the backdoor path $A \leftarrow U \rightarrow Y$ under some weak conditions (Ogburn and VanderWeele 2012). Therefore we will often prefer to adjust, rather than not to adjust, for L. We refer to nonconfounders that can be used to reduce confounding bias as *surrogate confounders*. A strategy to fight confounding is to measure as many surrogate confounders as possible and adjust for all of them.

Causal diagrams in this chapter include only fixed treatments that do not vary over time, but the structural definitions of confounding and confounders can be generalized to the case of time-varying treatments. When the treatment is time-varying, then so can be the confounders. A *time-varying confounder* is a time-varying variable that can be used to help eliminate confounding for the effect of a time-varying treatment. A *time-varying surrogate confounder* is a time-varying nonconfounder that can be used to reduce confounding for a time-varying treatment. Settings with time-varying confounders and treatments make it even clearer why the traditional definition of confounding, and the conventional methods for confounding adjustment, may result in selection bias. The bias of convenional methods is described in Part III.
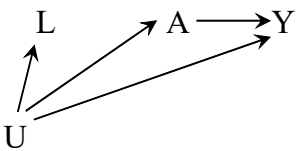

Figure 7.6

Furthermore, our example shows that confounding is a causal concept and that associational or statistical criteria are insufficient to characterize confounding. The standard definition of confounder that relies almost exclusively on statistical considerations may lead, as shown by Figure 7.4, to the wrong advice: adjust for a "confounder" even when confounding does not exist. In contrast, the structural definition of confounding emphasizes that causal inference from observational data requires a priori causal assumptions or beliefs, which are derived from subject-matter knowledge rather than statistical associations detected in the data. One important advantage of the structural definition is that it prevents inconsistencies between beliefs and actions. For example, if you believe Figure 7.4 is the true causal diagram—and therefore that there is no confounding for the effect of A on Y—then you will not adjust for the variable L.

A final note on the traditional definition of confounder. In an attempt to eliminate the problem described for Figure 7.4, some authors have proposed a modified definition of confounder that replaces the traditional condition "2) it is associated with the outcome conditional on the treatment" by the condition "2) it is a cause of the outcome." This modified definition of confounder indeed prevents inappropriate adjustment for L in Figure 7.4, but only to create a new problem by not considering L a confounder—that needs to be adjusted for—in Figure 7.2. Thus this modification of the traditional definition of confounder may lead to lack of adjustment for confounding.

Technical Point 7.1

Fixing the traditional definition of confounder. Figures 7.4 and 7.5 depict two examples in which the traditional definition of confounder misleads investigators into adjusting for a variable when adjustment for such variable is not only superfluous but also harmful. The traditional definition fails because it relies on two incorrect statistical criteria—conditions 1) and 2)—and one incorrect causal criterion—condition 3). To "fix" the traditional definition one needs to do two things:

1. Replace condition 3) by the condition that "there exist variables L and U such that there is conditional exchangeability within their joint levels $Y^a \amalg A|L, U$. If this new condition holds, it will quite generally be the case that L is not on a causal pathway between A and Y.

2. Replace conditions 1) and 2) by the following condition: U can be decomposed into two disjoint subsets U_1 and U_2 (i.e., $U = U_1 \cup U_2$ and $U_1 \cap U_2$ is empty) such that (i) U_1 and A are not associated within strata of L, and (ii) U_2 and Y are not associated within joint strata of A, L, and U_1. The variables in U_1 may be associated with the variables in U_2. U_1 can always be chosen to be the largest subset of U that is unassociated with treatment.

If these two new conditions are met we say L is a confounder and U is a non-confounder given data on L. These conditions were proposed by Robins (1997,Theorem 4.3) and further discussed by Greenland, Pearl, and Robins (1999, pp. 45-46, note the condition that $U = U_1 \cup U_2$ was inadvertently left out). These conditions generalize the traditional definition of confounder to overcome the difficulties found in Figures 7.4 and 7.5. For example, Greenland, Pearl, and Robins applied these conditions to Figure 7.4 to show that there is no confounding.

7.4 Confounding and exchangeability

So far we have defined confounding in terms of common causes of—open backdoor paths between—treatment and outcome. It is also possible to provide a definition of confounding strictly in terms of counterfactuals, with no explicit reference to common causes, in the absence of bias caused by selection (Chapter 8) or measurement (Chapter 9). In fact, that is precisely what we did in previous chapters in which we described the (confounding) bias that results from lack of exchangeability of the treated and the untreated.

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

When the treatment is unconditionally and randomly assigned, the treated and the untreated are expected to be exchangeable because no common causes exist. Marginal exchangeability, i.e., $Y^a \amalg A$, is equivalent to no confounding by either measured or unmeasured covariates. The average causal effect $E[Y^{a=1}] - E[Y^{a=0}]$ is calculated without adjustment for any variables.

When the treatment is assigned at random but conditionally on the prognostic factors L, the treated and the untreated are not expected to be exchangeable because the variables in L become common causes of treatment and outcome. Take our heart transplant study, a conditionally randomized experiment, as an example. Individuals who received a transplant ($A = 1$) are different from the untreated ($A = 0$) because, if the treated had remained untreated, their risk of death Y would have been higher than that of those that were actually untreated—the treated had a higher frequency of severe heart disease L, a common cause of A and Y. Thus the consequence of common causes of treatment and outcome is that the treated and the untreated are not marginally exchangeable. In conditionally randomized experiments, the treated and the untreated are expected to be conditionally exchangeable given L, i.e., $Y^a \amalg A|L$. Then the average causal effect in any stratum l of L is given by the stratum-specific risk difference $E[Y^{a=1}|L = l] - E[Y^{a=0}|L = l]$. There-

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

A formal proof of this result was given by Pearl (2000).

SWIGs overcome the shortcomings of previously proposed twin causal diagrams (Balke and Pearl 1994).

Figure 7.7

Figure 7.8

Figure 7.9

fore the average causal effect $E[Y^{a=1}] - E[Y^{a=0}]$ may be calculated by adjusting for the measured variables L via standardization. We say that there is no residual confounding whose elimination would require adjustment for unmeasured variables. For brevity, we say that there is *no unmeasured confounding*.

If conditioning on a set of variables L (that are non-descendants of A) blocks all backdoor paths, then the treated and untreated are exchangeable within levels of L, i.e., L is a *sufficient set for confounding adjustment* (see the previous section). To a non-mathematician such a result seems rather magical as there appears to be no obvious relationship between counterfactual independences and the absence of back door paths because counterfactuals are not included as variables on a causal graph. A new type of graphs—Single World Intervention Graphs (SWIGs)—seamlessly unify the counterfactual and graphical approaches by explicitly including the counterfactual variables on the graph. The SWIG depicts the variables and causal relations that would be observed in a hypothetical world in which all subjects received treatment level a. That is, a SWIG is a *graph* that represents a counterfactual *world* created by a *single intervention*. In contrast, a standard causal diagram represents the variables and causal relations that are observed in the actual world. A SWIG can be viewed as a function that transforms a given causal diagram under a given intervention. The following examples describe this transformation.
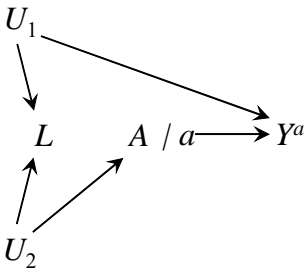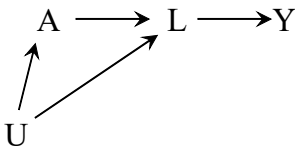
Suppose the causal diagram in Figure 7.2 represents the observed study data. The SWIG in Figure 7.7 is a transformation of Figure 7.2 that represents the data from a hypothetical intervention in which all subjects receive the same treatment level a. The treatment node is split into left and right sides. The right side encodes the treatment value a under the intervention; the left side encodes the value of treatment A that would have been observed in the absence of intervention, i.e., *the natural value of treatment*. Note that A is not a cause—does not have an arrow into—a because the value a is the same for all subjects. The outcome is Y^a, the value of Y in the hypothetical study. The remaining variables are temporally prior to A. Thus these variables and A take the same value as in the observational study. Conditional exchangeability $Y^a \amalg A|L$ holds because, on the SWIG, all paths between Y^a and A are blocked after conditioning on L.

Consider now the causal diagram in Figure 7.4 and the SWIG in Figure 7.8. Marginal exchangeability $Y^a \amalg A$ holds because, on the SWIG, all paths between Y^a and A are blocked (without conditioning on L). In contrast, conditional exchangeability $Y^a \amalg A|L$ does not hold because, on the SWIG, the path $Y^a \longleftarrow U_1 \longrightarrow L \longleftarrow U_2 \longrightarrow A$ is open when the collider L is conditioned on. This is why the marginal A-Y association is causal, but the conditional A-Y association given L is not, and thus any method that adjusts for L results in bias. These examples show how SWIGs unify the counterfactual and graphical approaches. See also Fine Point 7.2.

Knowledge of the causal structure is a prerequisite to determine the existence of confounding and label a variable as a confounder, and thus to decide which variables need to be measured and adjusted for. In observational studies, investigators measure many variables L in an attempt to ensure that the treated and the untreated are conditionally exchangeable given the measured covariates L. The underlying assumption is that, even though common causes may exist (confounding), the measured variables L are sufficient to block all backdoor paths (no unmeasured confounding). Of course, there is no guarantee that the assumption of no unmeasured confounding is true, which makes causal inference from observational data a risky undertaking.

There is a scientific consequence to the potential confounding in observa-

---

Fine Point 7.2

SWIGs verify that confounders cannot be descendants of treatment. Consider the causal DAG in Figure 7.9. L is a descendant of treatment A that blocks all backdoor paths from A to Y. Unlike in Figures 7.4 and 7.5, conditioning on L does not cause selection bias because no collider path is opened. Rather, because the causal effect of A on Y is solely through the intermediate variable L, conditioning on L completely blocks this pathway. This example shows that adjusting for a variable L that blocks all backdoor paths does not eliminates bias when L is a descendant of A.

Since conditional exchangeability $Y^a \amalg A|L$ implies the adjustment for L eliminates all bias, it must be the case thatconditional exchangeability fails to hold and the average treatment effect $E[Y^{a=1}] - E[Y^{a=0}]$ cannot be identified in this example. This failure can be verified by analyzing the SWIG in Figure 7.10, which depicts a counterfactual world in which A has been set to the value a. In this world, the factual variable L is replaced by the counterfactual variable L^a, that is, the value of L that would have been observed if all individuals had received treatment value a. Since L^a blocks all paths from Y^a to A we conclude that $Y^a \amalg A|L^a$ holds, but we cannot conclude that conditional exchangeabilty $Y^a \amalg A|L$ holds as L is not even on the graph. (Under an FFRCISTG, any independence that cannot be read off the SWIG cannot be assumed to hold.) Therefore, we cannot ensure that the average treatment effect $E[Y^{a=1}] - E[Y^{a=0}]$ is identified from data on (L, A, Y).
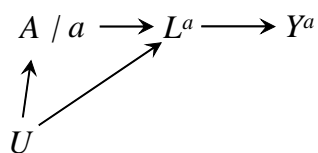
Figure 7.10

Additional conditions (e.g., no bias due to selection or measurement) are required for valid causal inference from observational data. But, unlike the expectation of no unmeasured confounding, these additional conditions may fail to hold in *both* observational studies and randomized experiments.

tional studies. Suppose you conducted an observational study to identify the effect of heart transplant A on death Y and that you assumed no unmeasured confounding given disease severity L. A critic of your study says "the inferences from this observational study may be incorrect because of potential confounding." The critic is not making a scientific statement, but a logical one. Since the findings from *any* observational study may be confounded, it is obviously true that those of your study can be confounded. If the critic's intent was to provide evidence about the shortcomings of your particular study, he failed. His criticism is completely noninformative because he simply restated a characteristic of observational research that you (and apparently he) already knew before the study was conducted.

To appropriately criticize your study, the critic needs to work harder and engage in a truly scientific conversation. For example, the critic may cite experimental or observational findings that contradict your findings, or he can say something along the lines of "the inferences from this observational study may be incorrect because of potential confounding due to cigarette smoking, a common cause through which a backdoor path may remain open". This latter option provides you with a testable challenge to your assumption of no unmeasured confounding. The burden of the proof is again yours. Your next move is to try and adjust for smoking.

The next section reviews the methods to adjust, or control for, confounding when, as in Figures 7.1-7.3, enough confounders L are measured to block all backdoor paths between treatment and outcome.

An important point. We have referred to unmeasured confounding as an "all or nothing" issue: either bias exists or it doesn't. In practice, however, it is important to consider the expected direction and magnitude of the bias. See Fine Point 7.4.

Fine Point 7.3

Identifying assumptions represented in causal diagrams. Exchangeability, positivity, and well-defined interventions are conditions required for causal inference via standardization or IP weighting. Positivity is roughly translated into graph language as the condition that the arrows from the nodes L to the treatment node A are not deterministic. A more precise discussion is given by Richardson and Robins (2013). 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 well-defined interventions are 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, which are then referred to as influence diagrams (Dawid 2002). The different status of treatment nodes compared with other nodes was also explicit in the causal trees introduced in Chapter 2, in which non-treatment branches were enclosed in circles (Robins 1986).

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

7.5 How to adjust for confounding

Randomization is the preferred method to control confounding because a random assignment of treatment is expected to produce exchangeability of the treated and the untreated, either marginally or conditionally. In marginally randomized experiments, no common causes of treatment and outcome are expected to exist and thus the unadjusted association measure is expected to equal the effect measure. In conditionally randomized experiments given covariates L, the common causes (i.e., the covariates L) are measured and thus the adjusted (via standardization or IP weighting) association measure is expected to equal the effect measure. Subject-matter knowledge to identify adjustment variables is unnecessary in ideal randomized experiments.

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

On the other hand, subject-matter knowledge is key in observational studies in order to identify and measure adjustment variables L. Causal inference from observational data relies on the uncheckable assumption that the measured variables L are not caused by treatment and are sufficient to block all backdoor paths—the assumption of no unmeasured confounding or of conditional exchangeability. But, as discussed in Section 4.6, standardization and IP weighting are not the only methods used to adjust for confounding in observational studies. Methods for confounding adjustment can be classified into the two following categories:

The 'g' in g-methods stands for 'generalized'. Unlike conventional stratification-based methods, g-methods can generally be used to estimate the effects of time-varying treatments as described in Part III.

- G-methods: Standardization, IP weighting, G-estimation (see Chapter 14). Methods that exploit conditional exchangeability in subsets defined by L to estimate the causal effect of A on Y in the entire population or in any subset of the population. In our heart transplant study, we used g-methods to adjust for confounding by disease severity L in Sections 2.4 (standardization) and 2.5 (IP weighting). The causal risk ratio in the population was 1.

- Stratification-based methods: Stratification, Restriction, Matching. Methods that exploit conditional exchangeability in subsets defined by L to estimate the association between A and Y in those subsets only. In our heart transplant study, we used stratification-based methods to adjust for confounding by disease severity L in Sections 4.4 (stratification, restriction) and 4.5 (matching). The causal risk ratio was 1 in all the subsets of the population that we studied.

The parametric and semiparametric extensions of g-methods are the parametric g-formula (standardization), IP weighting of marginal structural models, and g-estimation of nested structural models. The parametric and semiparametric extension of stratification is conventional regression. See Part II.

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

Technically, g-estimation requires the slightly weaker assumption that the magnitude of unmeasured confounding given L is known, of which the assumption of no unmeasured confounding is a particular case. See Chapter 14.

Under the assumption of conditional exchangeability given L, g-methods simulate the A-Y association in the population if backdoor paths involving the measured variables L did not exist; the simulated A-Y association can then be entirely attributed to the effect of A on Y. IP weighting achieves this by creating a pseudo-population in which treatment A is independent of the measured confounders L, that is, by "deleting" the arrow from L to A. The practical implications of "deleting" the arrow from measured confounders L to treatment A will become apparent when we discuss time-varying treatments and confounders in Part III.

Stratification-based methods estimate the association between treatment and outcome in one or more subsets of the population in which the treated and the untreated are assumed to be exchangeable. Hence the A-Y association in each subset is entirely attributed to the effect of A on Y. In graph terms, stratification/restriction do not delete the arrow from L to A but rather compute the conditional effect in a subset of the observed population (in which there is an arrow from L to A), which is represented by adding a box around variable L. Matching works by computing the effect in a selected subset of the observed population, which is represented by adding a selection node that is conditioned on (see Fine point 6.1 and Chapter 8).

All the above methods require conditional exchangeability given the measured covariates L to identify the effect of treatment A on outcome Y, i.e., the condition that the investigator has measured enough variables L to block all backdoor paths between A and Y. When interested in the effect in the entire population, conditional exchangeability is required in all strata defined by L; when interested in the effect in a subset of the population, conditional exchangeability is required in that subset only. Achieving conditional exchangeability may be an unrealistic goal in many observational studies but, as discussed in Section 3.2, expert knowledge can be used to get as close as possible to that goal.

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

A final note. The existence of common causes of treatment and outcome, and thus the definition of confounding, does not depend on the adjustment method. We do not say that measured confounding exists simply because the adjusted estimate is different from the unadjusted estimate. In fact, adjustment for measured confounding will generally imply a change in the estimate, but not necessarily the other way around. Changes in estimates may occur for

Fine Point 7.4

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

An approach to predict the direction of confounding bias is the use of signed causal diagrams. Consider the causal diagram in Figure 7.1 with dichotomous L, A, and Y variables. A positive sign over the arrow from L to A is added if L has a positive average causal effect on A (i.e., if the probability of $A = 1$ is greater among those with $L = 1$ than among those with $L = 0$), otherwise a negative sign is added if L has a negative average causal effect on A (i.e., if the probability of $A = 1$ is greater among those with $L = 0$ than among those with $L = 1$). Similarly a positive or negative sign is added over the arrow from L to Y. If both arrows are positive or both arrows are negative, then the confounding bias is said to be positive, which implies that effect estimate will be biased upwards in the absence of adjustment for L. If one arrow is positive and the other one is negative, then the confounding is said to be negative, which implies that the effect estimate will be biased downwards in the absence of adjustment for L. Unfortunately, this simple rule may fail in more complex causal diagrams or when the variables are non dichotomous. See VanderWeele, Hernán, and Robins (2008) for a more detailed discussion of signed diagrams in the context of average causal effects.

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

reasons other than confounding, including the introduction of selection bias when adjusting for nonconfounders (see Chapter 8) and the use of noncollapsible effect measures (see Fine Point 4.3). Attempts to define confounding based on change in estimates have been long abandoned because of these problems.

The next chapter presents another potential source of lack of exchangeability between the treated and the untreated: selection of individuals into the analysis.

Chapter 8
SELECTION BIAS

Suppose an investigator conducted a randomized experiment to answer the causal question "does one's looking up to the sky make other pedestrians look up too?" She found a strong association between her looking up and other pedestrians' looking up. Does this association reflect a causal effect? Well, by definition of randomized experiment, confounding bias is not expected in this study. However, there was another potential problem: The analysis included only those pedestrians that, after having been part of the experiment, gave consent for their data to be used. Shy pedestrians (those less likely to look up anyway) and pedestrians in front of whom the investigator looked up (who felt tricked) were less likely to participate. Thus participating individuals in front of whom the investigator looked up (a reason to decline participation) are less likely to be shy (an additional reason to decline participation) and therefore more likely to lookup. That is, the process of selection of individuals into the analysis guarantees that one's looking up is associated with other pedestrians' looking up, regardless of whether one's looking up actually makes others looking up.

An association created as a result of the process by which individuals are selected into the analysis is referred to as selection bias. Unlike confounding, this type of bias is not due to the presence of common causes of treatment and outcome, and can arise in both randomized experiments and observational studies. Like confounding, selection bias is just a form of lack of exchangeability between the treated and the untreated. This chapter provides a definition of selection bias and reviews the methods to adjust for it.

8.1 The structure of selection bias

Figure 8.1

Sometimes the term "selection bias" is used to refer to lack of generalizability of measures of frequency or effect. That is not the meaning we attribute to the term "selection bias" here. See Chapter 4 for a discussion of generalizability. Pearl (1995) and Spirtes et al (2000) used causal diagrams to describe the structure of bias resulting from selection.
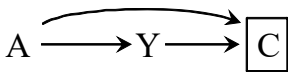
The term "selection bias" encompasses various biases that arise from the procedure by which individuals are selected into the analysis. The structure of selection bias can be represented by using causal diagrams like the one in Figure 8.1, which depicts dichotomous treatment A, outcome Y, and their common effect C. Suppose Figure 8.1 represents a study to estimate the effect of folic acid supplements A given to pregnant women shortly after conception on the fetus's risk of developing a cardiac malformation Y (1: yes, 0: no) during the first two months of pregnancy. The variable C represents death before birth. A cardiac malformation increases mortality (arrow from Y to C), and folic acid supplementation decreases mortality by reducing the risk of malformations other than cardiac ones (arrow from A to C). The study was restricted to fetuses who survived until birth. That is, the study was conditioned on no death $C = 0$ and hence the box around the node C.

The diagram in Figure 8.1 shows two sources of association between treatment and outcome: 1) the open path $A \rightarrow Y$ that represents the causal effect of A on Y, and 2) the open path $A \rightarrow C \leftarrow Y$ that links A and Y through their (conditioned on) common effect C. An analysis conditioned on C will generally result in an association between A and Y. We refer to this induced association between the treatment A and the outcome Y as selection bias due to conditioning on C. Because of selection bias, the associational risk ratio $\Pr[Y = 1|A = 1, C = 0]/\Pr[Y = 1|A = 0, C = 0]$ does not equal the causal risk ratio $\Pr\left[Y^{a=1} = 1\right]/\Pr\left[Y^{a=0} = 1\right]$; association is not causation. If the analysis were not conditioned on the common effect (collider) C, then the only

Figure 8.2

Figure 8.3

Figure 8.4

Figure 8.5

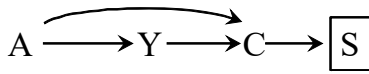
Figure 8.6

open path between treatment and outcome would be $A \rightarrow Y$, and thus the entire association between A and Y would be due to the causal effect of A on Y. That is, the associational risk ratio $\Pr[Y = 1|A = 1]/\Pr[Y = 1|A = 0]$ would equal the causal risk ratio $\Pr\left[Y^{a=1} = 1\right] / \Pr\left[Y^{a=0} = 1\right]$; association would be causation.

The causal diagram in Figure 8.2 shows another example of selection bias. This diagram includes all variables in Figure 8.1 plus a node S representing parental grief (1: yes, 0: no), which is affected by vital status at birth. Suppose the study was restricted to non grieving parents $S = 0$ because the others were unwilling to participate. As discussed in Chapter 6, conditioning on a variable S affected by the collider C also opens the path $A \rightarrow C \leftarrow Y$.
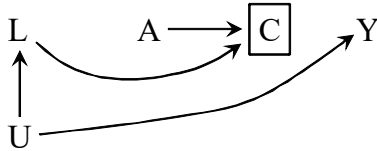
Both Figures 8.1 and 8.2 depict examples of selection bias in which the bias arises because of conditioning on a common effect of treatment and outcome: C in Figure 8.1 and S in Figure 8.2. However, selection bias can be defined more generally as illustrated by Figures 8.3 to 8.6. Consider the causal diagram in Figure 8.3, which represents a follow-up study of HIV-infected individuals to estimate the effect of certain antiretroviral treatment A on the 3-year risk of death Y. The unmeasured variable U represents high level of immunosup-pression (1: yes, 0: no). Patients with $U = 1$ have a greater risk of death. If a patient drops out from the study or is otherwise lost to follow-up before death or the end of the study, we say that he is censored ($C = 1$). Patients with $U = 1$ are more likely to be censored because the severity of their disease prevents them from participating in the study. The effect of U on censoring C is mediated by the presence of symptoms (fever, weight loss, diarrhea, and so on), CD4 count, and viral load in plasma, all included in L, which could or could not be measured. The role of L, when measured, in data analysis is discussed in Section 8.5; in this section, we take L to be unmeasured. Patients receiving treatment are at a greater risk of experiencing side effects, which could lead them to dropout, as represented by the arrow from A to C. For simplicity, assume that treatment A does not cause Y and so there is no arrow from A to Y. The square around C indicates that the analysis is restricted to those patients who remained uncensored ($C = 0$) because those are the only patients in which Y can be assessed.
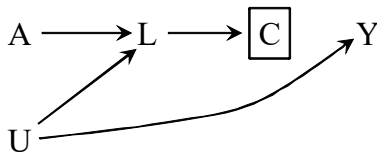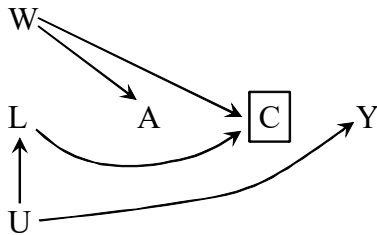
According to the rules of d-separation, conditioning on the collider C opens the path $A \rightarrow C \leftarrow L \leftarrow U \rightarrow Y$ and thus association flows from treatment A to outcome Y, i.e., the associational risk ratio is not equal to 1 even though the causal risk ratio is equal to 1. Figure 8.3 can be viewed as a simple transformation of Figure 8.1: the association between Y and C resulting from a direct effect of Y on C in Figure 8.1 is now the result of U, a common cause of Y and C. Some intuition for this bias: If a treated subject with treatment-induced side effects (and thereby at a greater risk of dropping out) did in fact not drop out ($C = 0$), then it is generally less likely that a second independent cause of dropping out (e.g., $U = 1$) was present. Therefore, an inverse association between A and U would be expected in those who did not dropped out ($C = 0$). Because U is positively associated with the outcome Y, restricting the analysis to subjects who did not drop out of this study induces an inverse association (mediated by U) between A and Y.
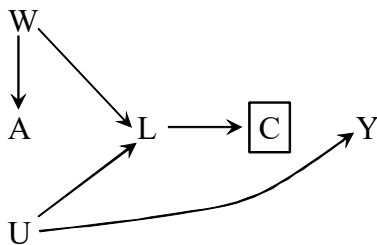
The bias in Figure 8.3 is an example of selection bias that results from conditioning on the censoring variable C, which is a common effect of treatment A and *a cause U of* the outcome Y, rather than of the outcome itself. We now present three additional causal diagrams that could lead to selection bias by differential loss to follow up. In Figure 8.4 prior treatment A has a direct effect on symptoms L. Restricting the study to the uncensored individ-

uals again implies conditioning on the common effect C of A and U, thereby introducing an association between treatment and outcome. Figures 8.5 and 8.6 are variations of Figures 8.3 and 8.4, respectively, in which there is a common cause W of A and another measured variable. W indicates unmeasured lifestyle/personality/educational variables that determine both treatment (arrow from W to A) and either attitudes toward attending study visits (arrow from W to C in Figure 8.5) or threshold for reporting symptoms (arrow from W to L in Figure 8.6).

Figures 8.5 and 8.6 show examples of M-bias.

We have described some different causal structures, depicted in Figures 8.1-8.6, that may lead to selection bias. In all these cases, the bias is the result of selection on a common effect of two other variables in the diagram, i.e., a collider. We will use the term selection bias to refer to all biases that arise from conditioning on a common effect of two variables, one of which is either the treatment or a cause of treatment, and the other is either the outcome or a cause of the outcome. We now describe some examples of selection bias that share this structure.

More generally, selection bias can be defined as the bias resulting from conditioning on the common effect of two variables, one of which is either the treatment or associated with the treatment, and the other is either the outcome or associated with the outcome (Hernán, Hernández-Díaz, and Robins 2004).

8.2 Examples of selection bias

Consider the following examples of bias due to the mechanism by which individuals are selected into the analysis:

- *Differential loss to follow-up*: This is precisely the bias described in the previous section and summarized in Figures 8.3-8.6. It is also referred to as bias due to *informative censoring.*

- *Missing data bias*, *nonresponse bias*: The variable C in Figures 8.3-8.6 can represent missing data on the outcome for any reason, not just as a result of loss to follow up. For example, individuals could have missing data because they are reluctant to provide information or because they miss study visits. Regardless of the reasons why data on Y are missing, restricting the analysis to subjects with complete data ($C = 0$) may result in bias.

- *Healthy worker bias*: Figures 8.3–8.6 can also describe a bias that could arise when estimating the effect of an occupational exposure A (e.g., a chemical) on mortality Y in a cohort of factory workers. The underlying unmeasured true health status U is a determinant of both death Y and of being at work C (1: no, 0: yes). The study is restricted to individuals who are at work ($C = 0$) at the time of outcome ascertainment. (L could be the result of blood tests and a physical examination.) Being exposed to the chemical reduces the probability of being at work in the near future, either directly (e.g., exposure can cause disabling asthma), like in Figures 8.3 and 8.4, or through a common cause W (e.g., certain exposed jobs are eliminated for economic reasons and the workers laid off) like in Figures 8.5 and 8.6.

- *Self-selection bias*, *volunteer bias*: Figures 8.3-8.6 can also represent a study in which C is agreement to participate (1: no, 0: yes), A is cigarette smoking, Y is coronary heart disease, U is family history of heart disease, and W is healthy lifestyle. (L is any mediator between U and C such as heart disease awareness.) Under any of these structures, selection bias

Berkson (1955) described the structure of bias due to self-selection.

Fine Point 8.1

Selection bias in case-control studies. Figure 8.1 can be used to represent selection bias in a case-control study. Suppose certain investigator wants to estimate the effect of postmenopausal estrogen treatment A on coronary heart disease Y. The variable C indicates whether a woman in the study population (the underlying cohort, in epidemiologic terms) is selected for the case-control study (1: no, 0: yes). The arrow from disease status Y to selection C indicates that cases in the population are more likely to be selected than noncases, which is the defining feature of a case-control study. In this particular case-control study, the investigator decided to select controls ($Y = 0$) preferentially among women with a hip fracture. Because treatment A has a protective causal effect on hip fracture, the selection of controls with hip fracture implies that treatment A now has a causal effect on selection C. This effect of A on C is represented by the arrow $A \rightarrow C$. One could add an intermediate node F (representing hip fracture) between A and C, but that is unnecessary for our purposes.

In a case-control study, the association measure (the treatment-outcome odds ratio) is by definition conditional on having been selected into the study ($C = 0$). If subjects with hip fracture are oversampled as controls, then the probability of control selection depends on a consequence of treatment A (as represented by the path from A to C) and "inappropriate control selection" bias will occur. Again, this bias arises because we are conditioning on a common effect C of treatment and outcome. A heuristic explanation of this bias follows. Among subjects selected for the study ($C = 0$), controls are more likely than cases to have had a hip fracture. Therefore, because estrogens lower the incidence of hip fractures, a control is less likely to be on estrogens than a case, and hence the A-Y odds ratio conditional on $C = 0$ would be greater than the causal odds ratio in the population. Other forms of selection bias in case-control studies, including some biases described by Berkson (1946) and incidence-prevalence bias, can also be represented by Figure 8.1 or modifications of it, as discussed by Hernán, Hernández-Díaz, and Robins (2004).

may be present if the study is restricted to those who volunteered or elected to participate ($C = 0$).

Robins, Hernán, and Rotnitzky (2007) used causal diagrams to describe the structure of bias due to the effect of pre-study treatments on selection into the study.

- *Selection affected by treatment received before study entry*: Suppose that C in Figures 8.3-8.6 represents selection into the study (1: no, 0: yes) and that treatment A took place before the study started. If treatment affects the probability of being selected into the study, then selection bias is expected. The case of selection bias arising from the effect of treatment on selection into the study can be viewed as a generalization of self-selection bias. This bias may be present in any study that attempts to estimate the causal effect of a treatment that occurred before the study started or in which treatment includes a pre-study component. For example, selection bias may arise when treatment is measured as the lifetime exposure to certain factor (medical treatment, lifestyle behavior...) in a study that recruited 50 year-old participants. In addition to selection bias, it is also possible that there exists unmeasured confounding for the pre-study component of treatment if confounders were only measured during the study.

For example, selection bias may be induced by certain attempts to eliminate ascertainment bias (Robins 2001) or to estimate direct effects (Cole and Hernán 2002), and by conventional adjustment for variables affected by previous treatment (see Part III).

In addition to the biases described here, as well as in Fine Point 8.1 and Technical Point 8.1, causal diagrams have been used to characterize various other biases that arise from conditioning on a common effect. These examples show that selection bias may occur in *retrospective studies*—those in which data on treatment A are collected *after* the outcome Y occurs—and in *prospective studies*—those in which data on treatment A are collected *before* the outcome Y occurs. Further, these examples show that selection bias may occur both in observational studies and in randomized experiments.

For example, Figures 8.3 and 8.4 could depict either an observational study or an experiment in which treatment A is randomly assigned, because there are

no common causes of A and any other variable. Individuals in *both* randomized experiments and observational studies may be lost to follow-up or drop out of the study before their outcome is ascertained. When this happens, the risk $\Pr[Y = 1|A = a]$ cannot be computed because the value of the outcome Y is unknown for the censored individuals ($C = 1$). Therefore only the risk among the uncensored $\Pr[Y = 1|A = a, C = 0]$ can be computed. This restriction of the analysis to the uncensored individuals may induce selection bias because uncensored subjects who remained through the end of the study ($C = 0$) may not be exchangeable with subjects that were lost ($C = 1$).

Hence a key difference between confounding and selection bias: randomization protects against confounding, but not against selection bias when the selection occurs after the randomization. On the other hand, no bias arises in randomized experiments from selection into the study before treatment is assigned. For example, only volunteers who agree to participate are enrolled in randomized clinical trials, but such trials are not affected by volunteer bias because participants are randomly assigned to treatment only after agreeing to participate ($C = 0$). Thus none of Figures 8.3-8.6 can represent volunteer bias in a randomized trial. Figures 8.3 and 8.4 are eliminated because treatment cannot cause agreement to participate C. Figures 8.5 and 8.6 are eliminated because, as a result of the random treatment assignment, there cannot exist a common cause of treatment and any other variable.
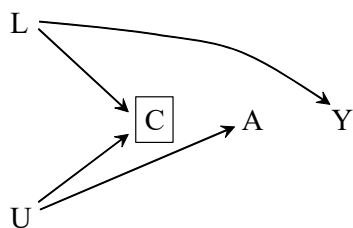
8.3 Selection bias and confounding
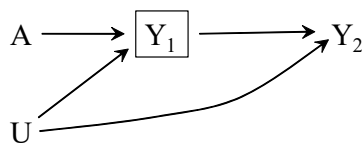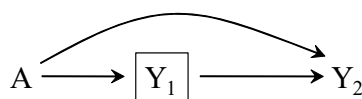

Figure 8.7

Figure 8.8

Figure 8.9

In this and the previous chapter, we describe two reasons why the treated and the untreated may not be exchangeable: 1) the presence of common causes of treatment and outcome, and 2) conditioning on common effects of treatment and outcome (or causes of them). We refer to biases due to the presence of common causes as "confounding" and to those due to conditioning on common effects as "selection bias." This structural definition provides a clear-cut classification of confounding and selection bias, even though it might not coincide perfectly with the traditional, often discipline-specific, terminologies. For instance, the same phenomenon is sometimes named "confounding by indication" by epidemiologists and "selection bias" by statisticians and econometricians. Others use the term "selection bias" when "confounders" are unmeasured. Sometimes the distinction between confounding and selection bias is blurred in the term "selection confounding." Our goal, however, is not to be normative about terminology, but rather to emphasize that, regardless of the particular terms chosen, there are two distinct causal structures that lead to bias.

The end result of both structures is lack of exchangeability between the treated and the untreated. For example, consider a study restricted to firefighters that aims to estimate the causal effect of being physically active A on the risk of heart disease Y as represented in Figure 8.7. For simplicity, we assume that, unknown to the investigators, A does not cause Y. Parental socioeconomic status L affects the risk of becoming a firefighter C and, through childhood diet, of heart disease Y. Attraction toward activities that involve physical activity (an unmeasured variable U) affects the risk of becoming a firefighter and of being physically active (A). U does not affect Y, and L does not affect A. According to our terminology, there is no confounding because there are no common causes of A and Y. Thus, the associational risk ratio $\Pr[Y = 1|A = 1] / \Pr[Y = 1|A = 0]$ is expected to equal the causal risk ratio

Technical Point 8.1

The built-in selection bias of hazard ratios. The causal DAG in Figure 8.8 describes a randomized experiment of the effect of heart transplant A on death at times 1 (Y_1) and 2 (Y_2). The arrow from A to Y_1 represents that transplant decreases the risk of death at time 1. The lack of an arrow from A to Y_2 indicates that A has no direct effect on death at time 2. That is, heart transplant does not influence the survival status at time 2 of any subject who would survive past time 1 when untreated (and thus when treated). U is an unmeasured haplotype that decreases the subject's risk of death at all times. Because of the absence of confounding, the associational risk ratios $aRR_{AY_1} = \frac{\Pr[Y_1=1|A=1]}{\Pr[Y_1=1|A=0]}$ and $aRR_{AY_2} = \frac{\Pr[Y_2=1|A=1]}{\Pr[Y_2=1|A=0]}$ are unbiased measures of the effect of A on death at times 1 and 2, respectively. Note that, even though A has no direct effect on Y_2, aRR_{AY_2} will be less than 1 because it is a measure of the effect of A on total mortality through time 2.

Consider now the time-specific hazard ratio (which, for all practical purposes, is equivalent to the rate ratio). In discrete time, the hazard of death at time 1 is the probability of dying at time 1 and thus the associational hazard ratio is the same as aRR_{AY_1}. However, the hazard at time 2 is the probability of dying at time 2 among those who survived past time 1. Thus, the associational hazard ratio at time 2 is then $aRR_{AY_2|Y_1=0} = \frac{\Pr[Y_2=1|A=1,Y_1=0]}{\Pr[Y_2=1|A=0,Y_1=0]}$. The square around Y_1 in Figure 8.8 indicates this conditioning. Treated survivors of time 1 are less likely than untreated survivors of time 1 to have the protective haplotype U (because treatment can explain their survival) and therefore are more likely to die at time 2. That is, conditional on Y_1, treatment A is associated with a higher mortality at time 2. Thus, the hazard ratio at time 1 is less than 1, whereas the hazard ratio at time 2 is greater than 1, i.e., the hazards have crossed. We conclude that the hazard ratio at time 2 is a biased estimate of the direct effect of treatment on mortality at time 2. The bias is selection bias arising from conditioning on a common effect Y_1 of treatment A and of U, which is a cause of Y_2 that opens the associational path $A \rightarrow Y_1 \leftarrow U \rightarrow Y_2$ between A and Y_2. In the survival analysis literature, an unmeasured cause of death that is marginally unassociated with treatment such as U is often referred to as a *frailty*.

In contrast, the conditional hazard ratio $aRR_{AY_2|Y_1=0,U}$ is 1 within each stratum of U because the path $A \rightarrow Y_1 \leftarrow U \rightarrow Y_2$ is now blocked by conditioning on the noncollider U. Thus, the conditional hazard ratio correctly indicates the absence of a direct effect of A on Y_2. That the unconditional hazard ratio $aRR_{AY_2|Y_1=0}$ differs from the stratum-specific hazard ratios $aRR_{AY_2|Y_1=0,U}$, even though U is independent of A, shows the noncollapsibility of the hazard ratio (Greenland, 1996b). Unfortunately, the unbiased measure $aRR_{AY_2|Y_1=0,U}$ of the direct effect of A on Y_2 cannot be computed because U is unobserved. In the absence of data on U, it is impossible to know whether A has a direct effect on Y_2. That is, the data cannot determine whether the true causal DAG generating the data was that in Figure 8.8 or in Figure 8.9. All of the above applies to both observational studies and randomized experiments.

$\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1] = 1$. However, in a study restricted to firefighters ($C = 0$), the associational and causal risk ratios would differ because conditioning on a common effect C of causes of treatment and outcome induces selection bias resulting in lack of exchangeability of the treated and untreated firefighters. To the study investigators, the distinction between confounding and selection bias is moot because, regardless of nomenclature, they must adjust for L to make the treated and the untreated firefighters comparable. This example demonstrates that a structural classification of bias does not always have consequences for the analysis of a study. Indeed, for this reason, many epidemiologists use the term "confounder" for any variable L on which one has to adjust for, regardless of whether the lack of exchangeability is the result of conditioning on a common effect or the result of a common cause of treatment and outcome.

There are, however, advantages of adopting a structural approach to the classification of sources of non exchangeability. First, the structure of the problem frequently guides the choice of analytical methods to reduce or avoid the bias. For example, in longitudinal studies with time-varying treatments, identifying the structure allows us to detect situations in which adjustment for

confounding via stratification would introduce selection bias (see Part III). In those cases, IP weighting or g-estimation are better alternatives. Second, even when understanding the structure of bias does not have implications for data analysis (like in the firefighters' study), it could still help study design. For example, investigators running a study restricted to firefighters should make sure that they collect information on joint risk factors for the outcome Y and for the selection variable C (i.e., becoming a firefighter), as described in the first example of confounding in Section 7.1. Third, selection bias resulting from conditioning on pre-treatment variables (e.g., being a firefighter) could explain why certain variables behave as "confounders" in some studies but not others. In our example, parental socioeconomic status L would not necessarily need to be adjusted for in studies not restricted to firefighters. Finally, causal diagrams enhance communication among investigators and may decrease the occurrence of misunderstandings.

Selection on pre-treatment factors may introduce bias. Some authors refer to this bias as "confounding", rather than as "selection bias", even if no common causes exist. This choice of terminology usually has no practical consequences: adjusting for L in Figure 7.4 creates bias, regardless of its name. However, disregard for the causal structure when there is selection on post-treatment factors may lead to apparent paradoxes like the so-called Simpson's paradox (1951). See Hernán, Clayton, and Keiding (2011) for details.

As an example of the last point, consider the "*healthy worker bias*". We described this bias in the previous section as an example of a bias that arises from conditioning on the variable C, which is a common effect of (a cause of) treatment and (a cause of) the outcome. Thus the bias can be represented by the causal diagrams in Figures 8.3-8.6. However, the term "healthy worker bias" is also used to describe the bias that occurs when comparing the risk in certain group of workers with that in a group of subjects from the general population. This second bias can be depicted by the causal diagram in Figure 7.1 in which L represents health status, A represents membership in the group of workers, and Y represents the outcome of interest. There are arrows from L to A and Y because being healthy affects job type and risk of subsequent outcome, respectively. In this case, the bias is caused by the common cause L and we would refer to it as confounding. The use of causal diagrams to represent the structure of the "healthy worker bias" prevents any confusions that may arise from employing the same term for different sources of non exchangeability.

8.4 Selection bias and identifiability of causal effects

Suppose an investigator conducted a marginally randomized experiment to estimate the average causal effect of wasabi intake on the one-year risk of death ($Y = 1$). Half of the 60 study participants were randomly assigned to eating meals supplemented with wasabi ($A = 1$) until the end of follow-up or death, whichever occurred first. The other half were assigned to meals that contained no wasabi ($A = 0$). After 1 year, 17 subjects died in each group. That is, the associational risk ratio $\Pr[Y = 1|A = 1] / \Pr[Y = 1|A = 0]$ was 1. Because of randomization, the causal risk ratio $\Pr[Y^{a=1} = 1] / \Pr[Y^{a=0} = 1]$ is also expected to be 1. (If ignoring random variability bothers you, please imagine the study had 60 million patients rather than 60.)

Unfortunately, the investigator could not observe the 17 deaths that occurred in each group because many patients were lost to follow-up, or censored, before the end of the study (i.e., death or one year after treatment assignment). The proportion of censoring ($C = 1$) was higher among patients with heart disease ($L = 1$) at the start of the study and among those assigned to wasabi supplementation ($A = 1$). In fact, only 9 individuals in the wasabi group and 22 individuals in the other group were not lost to follow-up. The investigator observed 4 deaths in the wasabi group and 11 deaths in the other group. That is,

the associational risk ratio $\Pr[Y = 1|A = 1, C = 0] / \Pr[Y = 1|A = 0, C = 0]$ was $(4/9)/(11/22) = 0.89$ among the uncensored. The risk ratio of 0.89 in the uncensored differs from the causal risk ratio of 1 in the entire population: There is selection bias due to conditioning on the common effect C.

The causal diagram in Figure 8.3 depicts the relation between the variables L, A, C, and Y in the randomized trial of wasabi. U represents atherosclerosis, an unmeasured variable, that affects both heart disease L and death Y. Figure 8.3 shows that there are no common causes of A and Y, as expected in a marginally randomized experiment, and thus there is no need to adjust for confounding to compute the causal effect of A on Y. On the other hand, Figure 8.3 shows that there is a common cause U of C and Y. The presence of this backdoor path $C \leftarrow L \leftarrow U \rightarrow Y$ implies that, were the investigator interested in estimating the causal effect of censoring C on Y (which is null in Figure 8.3), she would have to adjust for confounding due to the common cause U. The backdoor criterion says that such adjustment is possible because the measured variable L can be used to block the backdoor path $C \leftarrow L \leftarrow U \rightarrow Y$.

The causal contrast we have considered so far is "the risk if everybody had been treated", $\Pr[Y^{a=1} = 1]$, versus "the risk if everybody had remained untreated", $\Pr[Y^{a=0} = 1]$, and this causal contrast does not involve C at all. Why then are we talking about confounding for the causal effect of C? It turns out that the causal contrast of interest needs to be modified in the presence of censoring or, in general, of selection. Because selection bias would not exist if everybody had been uncensored $C = 0$, we would like to consider a causal contrast that reflects what would have happened in the absence of censoring.

Let $Y^{a=1,c=0}$ be a subject's counterfactual outcome if he had received treatment $A = 1$ *and* he had remained uncensored $C = 0$. Similarly, let $Y^{a=0,c=0}$ be a subject's counterfactual outcome if he had not received treatment $A = 0$ *and* he had remained uncensored $C = 0$. Our causal contrast of interest is now "the risk if everybody had been treated and had remained uncensored", $\Pr[Y^{a=1,c=0} = 1]$, versus "the risk if everybody had remained untreated and uncensored", $\Pr[Y^{a=0,c=0} = 1]$. This causal contrast does involve the censoring variable C and therefore considerations about confounding for C become central. In fact, under this conceptualization of the causal contrast of interest, we can think of censoring C as just another treatment. The goal of the analysis is to compute the causal effect of a joint intervention on A and C. Since censoring C is now viewed as a treatment, we will need to ensure that the identifiability conditions of exchangeability, positivity, and well-defined interventions hold for C as well as for A.

For example, we may want to compute the causal risk ratio $\mathrm{E}\left[Y^{a=1,c=0}\right] / \mathrm{E}\left[Y^{a=0,c=0}\right]$ or the causal risk difference $\mathrm{E}\left[Y^{a=1,c=0}\right] - \mathrm{E}\left[Y^{a=0,c=0}\right]$.

Under these identifiability conditions, selection bias can be eliminated via analytic adjustment and, in the absence of measurement error and confounding, the causal effect of treatment A on outcome Y can be identified. To eliminate selection bias for the effect of treatment A, we need to adjust for confounding for the effect of treatment C. The next section explains how to do so.

8.5 How to adjust for selection bias

Though selection bias can sometimes be avoided by an adequate design (see Fine Point 8.1), it is often unavoidable. For example, loss to follow up, self-selection, and, in general, missing data leading to bias can occur no matter how careful the investigator. In those cases, the selection bias needs to be explicitly corrected in the analysis. This correction can sometimes be accomplished by

IP weighting (or by standardization), which is based on assigning a weight W^C to each selected subject $(C = 0)$ so that she accounts in the analysis not only for herself, but also for those like her, i.e., with the same values of L and A, who were not selected $(C = 1)$. The IP weight W^C is the inverse of the probability of her selection $\Pr[C = 0|L, A]$.
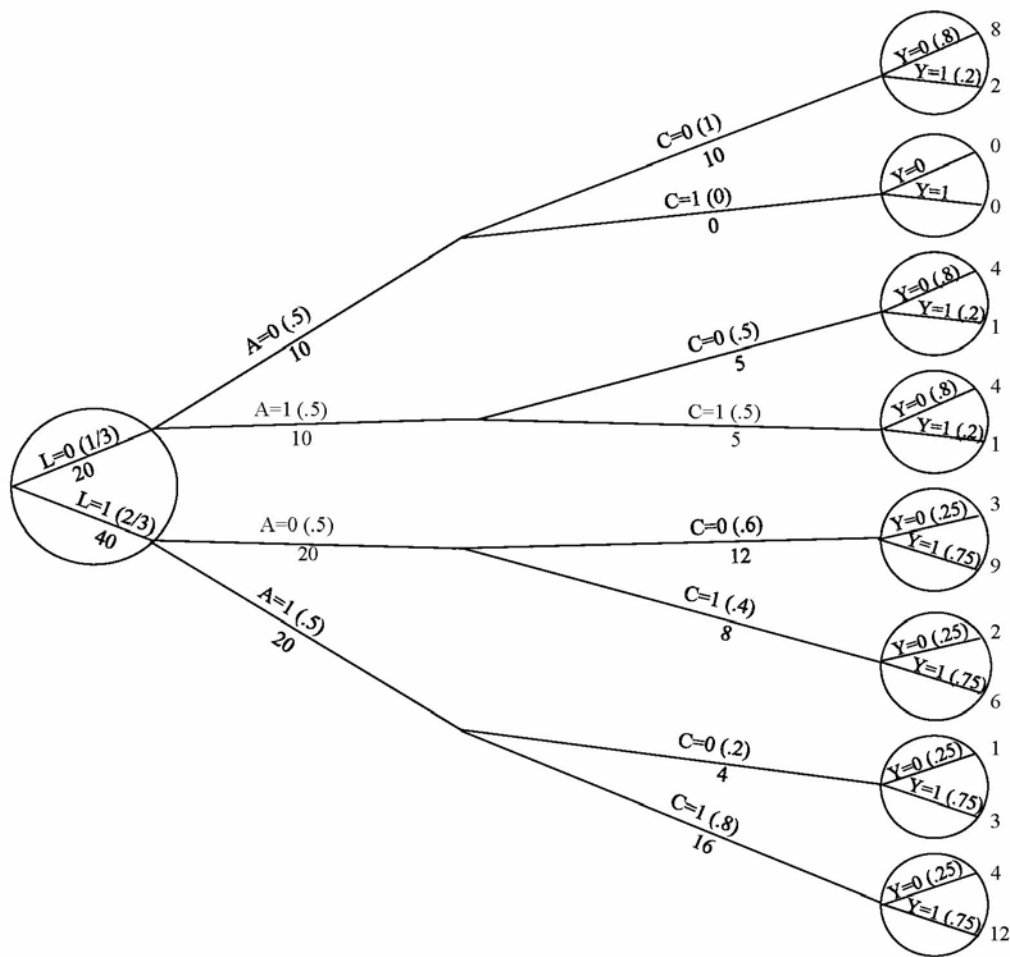

Figure 8.10

We have described IP weights to adjust for confounding, $W^A = 1/f(A|L)$, and selection bias. $W^C = 1/\Pr[C = 0|A, L]$. When both confounding and selection bias exist, the product weight $W^A W^C$ can be used to adjust simultaneously for both biases under assumptions described in Chapter 12 and Part III.

To describe the application of IP weighting for selection bias adjustment consider again the wasabi randomized trial described in the previous section. The tree graph in Figure 8.10 presents the trial data. Of the 60 individuals in the trial, 40 had $(L = 1)$ and 20 did not have $(L = 0)$ heart disease at the time of randomization. Regardless of their L status, all individuals had a 50/50 chance of being assigned to wasabi supplementation $(A = 1)$. Thus 10 individuals in the $L = 0$ group and 20 in the $L = 1$ group received treatment $A = 1$. This lack of effect of L on A is represented by the lack of an arrow from L to A in the causal diagram of Figure 8.3. The probability of remaining uncensored varies across branches in the tree. For example, 50% of the individuals without heart disease that were assigned to wasabi $(L = 0, A = 1)$, whereas 60% of the individuals with heart disease that were assigned to no wasabi $(L = 1, A = 0)$, remained uncensored. This effect of A and L on C is represented

by arrows from A and L into C in the causal diagram of Figure 8.3. Finally, the tree shows how many people would have died ($Y = 1$) both among the uncensored and the censored individuals. Of course, in real life, investigators would never know how many deaths occurred among the censored individuals. It is precisely the lack of this knowledge which forces investigators to restrict the analysis to the uncensored, opening the door for selection bias. Here we show the deaths in the censored to document that, as depicted in Figure 8.3, treatment A is marginally independent on Y, and censoring C is independent of Y within levels of L. It can also be checked that the risk ratio in the entire population (inaccessible to the investigator) is 1 whereas the risk ratio in the uncensored (accessible to the investigator) is 0.89.
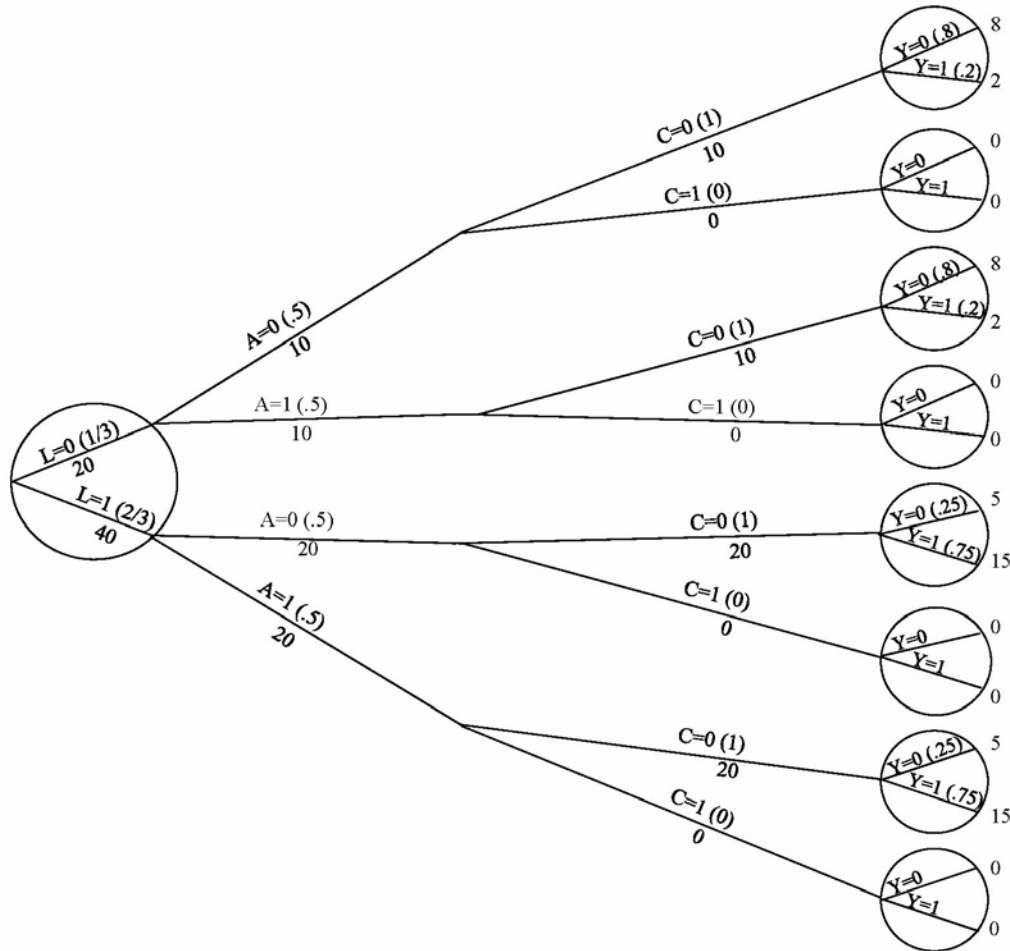
Figure 8.11

Let us now describe the intuition behind the use of IP weighting to adjust for selection bias. Look at the bottom of the tree in Figure 8.10. There are 20 individuals with heart disease ($L = 1$) who were assigned to wasabi supplementation ($A = 1$). Of these, 4 remained uncensored and 16 were lost to follow-up. That is, the conditional probability of remaining uncensored in this group is 1/5, i.e., $\Pr[C = 0|L = 1, A = 1] = 4/20 = 0.2$. In an IP weighted analysis the 16 censored individuals receive a zero weight (i.e., they

do not contribute to the analysis), whereas the 4 uncensored individuals receive a weight of 5, which is the inverse of their probability of being uncensored (1/5). IP weighting replaces the 20 original subjects by 5 copies of each of the 4 uncensored subjects. The same procedure can be repeated for the other branches of the tree, as shown in Figure 8.11, to construct a pseudo-population of the same size as the original study population but in which nobody is lost to follow-up. (We let the reader derive the IP weights for each branch of the tree.) The associational risk ratio in the pseudo-population is 1, the same as the risk ratio $\Pr\left[Y^{a=1,c=0} = 1\right] / \Pr\left[Y^{a=0,c=0} = 1\right]$ that would have been computed in the original population if nobody had been censored.

The association measure in the pseudo-population equals the effect measure in the original population if the following three identifiability conditions are met.

In many applications, it is reasonable to assume that censoring does not have a causal effect on the outcome (an exception would be a setting in which being lost to follow-up prevents people from getting additional treatment). One might then imagine that the definition of causal effect could ignore censoring, i.e., we could omit the superscript $c = 0$. However, omitting the superscript would obscure the fact that estimating the effect of treatment A in the presence of selection bias due to C requires 1) the same assumptions for censoring C as for treatment A, and 2) statistical methods that are identical to those you would have to use if you wanted to estimate the effect of censoring C.

First, the average outcome in the uncensored subjects must equal the unobserved average outcome in the censored subjects with the same values of A and L. This provision will be satisfied if the probability of selection $\Pr[C = 0|L = 1, A = 1]$ is calculated conditional on treatment A and on all additional factors that independently predict both selection and the outcome, that is, if the variables in A and L are sufficient to block all backdoor paths between C and Y. Unfortunately, one can never be sure that these additional factors were identified and recorded in L, and thus the causal interpretation of the resulting adjustment for selection bias depends on this untestable *exchangeability* assumption.

Second, IP weighting requires that all conditional probabilities of being uncensored given the variables in L must be greater than zero. Note this *positivity* condition is required for the probability of being uncensored ($C = 0$) but not for the probability of being censored ($C = 1$) because we are not interested in inferring what would have happened if study subjects had been censored, and thus there is no point in constructing a pseudo-population in which everybody is censored. For example, the tree in Figure 8.10 shows that $\Pr[C = 1|L = 0, A = 0] = 0$, but this zero does not affect our ability to construct a pseudo-population in which nobody is censored.

The third condition is *well-defined interventions*. IP weighting is used to create a pseudo-population in which censoring C has been abolished, and in which the effect of the treatment A is the same as in the original population. Thus, the pseudo-population effect measure is equal to the effect measure had nobody been censored. This effect measure may be relatively well defined when censoring is the result of loss to follow up or nonresponse, but not when censoring is the result of *competing events*. For example, in a study aimed at estimating the effect of certain treatment on the risk of Alzheimer's disease, we might not wish to base our effect estimates on a pseudo-population in which all other causes of death (cancer, heart disease, stroke, and so on) have been removed, because it is unclear even conceptually what sort of medical intervention would produce such a population. A more pragmatic reason is that no feasible intervention could possibly remove just one cause of death without affecting the others as well.

A *competing event* is an event that prevents the outcome of interest from happening. A typical example of competing event is death because, once an individual dies, no other outcomes can occur.

Finally, one could argue that IP weighting is not necessary to adjust for selection bias in a setting like that described in Figure 8.3. Rather, one might attempt to remove selection bias by stratification (i.e., by estimating the effect measure conditional on the L variables) rather than by IP weighting. Stratification could yield unbiased conditional effect measures within levels of L because conditioning on L is sufficient to block the backdoor path from C to Y. That is, the conditional risk ratio

$$\Pr\left[Y = 1 | A = 1, C = 0, L = l\right] / \Pr\left[Y = 1 A = 0, C = 0, L = l\right]$$

can be interpreted as the effect of treatment among the uncensored with $L = l$. For the same reason stratification would work (i.e., it would provide an unbiased conditional effect measure) under the causal structure depicted in Figure 8.5. Stratification, however, would not work under the structure depicted in Figures 8.4 and 8.6. Take Figure 8.4. Conditioning on L blocks the backdoor path from C to Y but also opens the path $A \rightarrow L \leftarrow U \rightarrow Y$ from A to Y because L is a collider on that path. Thus, even if the causal effect of A on Y is null, the conditional (on L) risk ratio would be generally different from 1. And similarly for Figure 8.6. In contrast, IP weighting appropriately adjusts for selection bias under Figures 8.3-8.6 because this approach is not based on estimating effect measures conditional on the covariates L, but rather on estimating unconditional effect measures after reweighting the subjects according to their treatment and their values of L. This is the first time we discuss a situation in which stratification cannot be used to validly compute the causal effect of treatment, even if the three conditions of exchangeability, positivity, and well-defined interventions hold. We will discuss other situations with a similar structure in Part III when estimating direct effects and the effect of time-varying treatments.
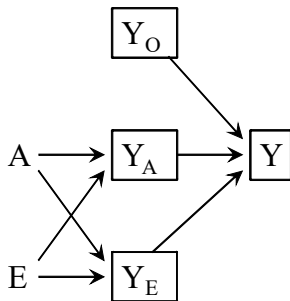

Figure 8.12

Figure 8.13

8.6 Selection without bias

Figure 8.14

The causal diagram in Figure 8.12 represents a hypothetical study with dichotomous variables surgery A, certain genetic haplotype E, and death Y. According to the rules of d-separation, surgery A and haplotype E are (i) marginally independent, i.e., the probability of receiving surgery is the same for people with and without the genetic haplotype, and (ii) associated conditionally on Y, i.e., the probability of receiving surgery varies by haplotype when the study is restricted to, say, the survivors ($Y = 0$).

Indeed conditioning on the common effect Y of two independent causes A and E always induces a conditional association between A and E in at least one of the strata of Y (say, $Y = 1$). However, there is a special situation under which A and E remain conditionally independent within the other stratum (say, $Y = 0$).
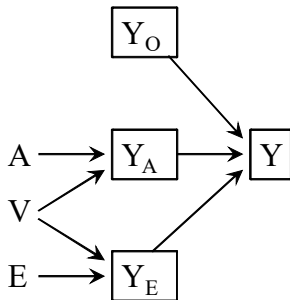
Suppose A and E affect survival through totally independent mechanisms in such a way that E cannot possibly modify the effect of A on Y, and vice versa. For example, suppose that the surgery A affects survival through the removal of a tumor, whereas the haplotype E affects survival through increasing levels of low-density lipoprotein-cholesterol levels resulting in an increased risk of heart attack (whether or not a tumor is present). In this scenario, we can consider 3 cause-specific mortality variables: death from tumor Y_A, death from heart attack Y_E, and death from any other causes Y_O. The observed mortality variable Y is equal to 1 (death) when Y_A or Y_E or Y_O is equal to 1, and Y is equal to 0 (survival) when Y_A and Y_E and Y_O equal 0. The causal diagram in Figure 8.13, an expansion of that in Figure 8.12, represents a causal structure linking all these variables. We assume data on underlying cause of death (Y_A, Y_E, Y_O) are not recorded and thus the only measured variables are those in Figure 8.12 (A, E, Y).

Figure 8.15

Technical Point 8.2

Multiplicative survival model. When the conditional probability of survival $\Pr[Y = 0 | E = e, A = a]$ given A and E is equal to a product $g(e)h(a)$ of functions of e and a, we say that a multiplicative survival model holds. A multiplicative survival model

$$\Pr[Y = 0 | E = e, A = a] = g(e)h(a)$$

is equivalent to a model that assumes the survival ratio $\Pr[Y = 0 | E = e, A = a] / \Pr[Y = 0 | E = e, A = 0]$ does not depend on e and is equal to $h(a)$. The data follow a multiplicative survival model when there is no interaction between A and E on the multiplicative scale as depicted in Figure 8.13. Note that if $\Pr[Y = 0 | E = e, A = a] = g(e)h(a)$, then $\Pr[Y = 1 | E = e, A = a] = 1 - g(e)h(a)$ does not follow a multiplicative mortality model. Hence, when A and E are conditionally independent given $Y = 0$, they will be conditionally dependent given $Y = 1$.
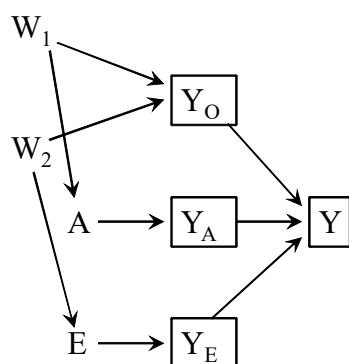

Figure 8.16

Augmented causal DAGs, introduced by Hernán, Hernández-Díaz, and Robins (2004), can be extended to represent the sufficient causes described in Chapter 5 (VanderWeele and Robins, 2007c).

Because the arrows from Y_A, Y_E and Y_O to Y are deterministic, conditioning on observed survival $(Y = 0)$ is equivalent to simultaneously conditioning on $Y_A = 0$, $Y_E = 0$, and $Y_O = 0$ as well, i.e., conditioning on $Y = 0$ implies $Y_A = Y_E = Y_O = 0$. As a consequence, we find by applying d-separation to Figure 8.13 that A and E are conditionally independent given $Y = 0$, i.e., the path, between A and E through the conditioned on collider Y is blocked by conditioning on the noncolliders Y_A, Y_E and Y_O. On the other hand, conditioning on death $Y = 1$ does not imply conditioning on any specific values of Y_A, Y_E and Y_O as the event $Y = 1$ is compatible with 7 possible unmeasured events: $(Y_A = 1, Y_E = 0, Y_O = 0)$, $(Y_A = 0, Y_E = 1, Y_O = 0)$, $(Y_A = 0, Y_E = 0, Y_O = 1)$, $(Y_A = 1, Y_E = 1, Y_O = 0)$, $(Y_A = 0, Y_E = 1, Y_O = 1)$, $(Y_A = 1, Y_E = 0, Y_O = 1)$, and $(Y_A = 1, Y_E = 1, Y_O = 1)$. Thus, the path between A and E through the conditioned on collider Y is not blocked: A and E are associated given $Y = 1$.

In contrast with the situation represented in Figure 8.13, the variables A and E will not be independent conditionally on $Y = 0$ when one of the situations represented in Figures 8.14-8.16 occur. If A and E affect survival through a common mechanism, then there will exist an arrow either from A to Y_E or from E to Y_A, as shown in Figure 8.14. In that case, and A and E will be dependent within both strata of Y. Similarly, if Y_A and Y_E are not independent because of a common cause V as shown in Figure 8.15, A and E will be dependent within both strata of Y. Finally, if the causes Y_A and Y_O, and Y_E and Y_O, are not independent because of common causes W_1 and W_2 as shown in Figure 8.16, then A and E will also be dependent within both strata of Y. When the data can be summarized by Figure 8.12, we say that the data follow a *multiplicative survival model* (see Technical Point 8.2).

What is interesting about Figure 8.13 is that by adding the unmeasured variables Y_A, Y_E and Y_O, which functionally determine the observed variable Y, we have created an augmented causal diagram that succeeds in representing both the conditional independence between A and E given $Y = 0$ and the their conditional dependence given $Y = 1$.

In summary, conditioning on a collider always induces an association between its causes, but this association could be restricted to certain levels of the common effect. In other words, it is theoretically possible that selection on a common effect does not result in selection bias when the analysis is restricted to a single level of the common effect.

Fine Point 8.2

The strength and direction of selection bias. We have referred to selection bias as an "all or nothing" issue: either bias exists or it doesn't. In practice, however, it is important to consider the expected direction and magnitude of the bias.

The direction of the conditional association between 2 marginally independent causes A and E within strata of their common effect Y depends on how the two causes A and E interact to cause Y. For example, suppose that, in the presence of an undiscovered background factor U that is unassociated with A or E, having either $A = 1$ or $E = 1$ is sufficient and necessary to cause death (an "or" mechanism), but that neither A nor E causes death in the absence of U. Then among those who died ($Y = 1$), A and E will be negatively associated, because it is more likely that an individual with $A = 0$ had $E = 1$ because the absence of A increases the chance that E was the cause of death. (Indeed, the logarithm of the conditional odds ratio $OR_{AE|Y=1}$ will approach minus infinity as the population prevalence of U approaches 1.0.) This "or" mechanism was the only explanation given in the main text for the conditional association of independent causes within strata of a common effect; nonetheless, other possibilities exist.

For example, suppose that in the presence of the undiscovered background factor U, having both $A = 1$ and $E = 1$ is sufficient and necessary to cause death (an "and" mechanism) and that neither A nor E causes death in the absence of U. Then, among those who die, those with $A = 1$ are more likely to have $E = 1$, i.e., A and E are positively correlated. A standard DAG such as that in Figure 8.12 fails to distinguish between the case of A and E interacting through an "or" mechanism from the case of an "and" mechanism. Causal DAGs with sufficient causation structures (VanderWeele and Robins, 2007c) overcome this shortcoming.

Regardless of the direction of selection bias, another key issue is its magnitude. Biases that are not large enough to affect the conclusions of the study may be safely ignored in practice, whether the bias is upwards or downwards. Generally speaking, a large selection bias requires strong associations between the collider and both treatment and outcome. Greenland (2003) studied the magnitude of selection bias, which he referred to as *collider-stratification bias*, under several scenarios.

Chapter 9
MEASUREMENT BIAS

Suppose an investigator conducted a randomized experiment to answer the causal question "does one's looking up to the sky make other pedestrians look up too?" She found a weak association between her looking up and other pedestrians' looking up. Does this weak association reflect a weak causal effect? By definition of randomized experiment, confounding bias is not expected in this study. In addition, no selection bias was expected because all pedestrians' responses—whether they did or did not look up—were recorded. However, there was another problem: the investigator's collaborator who was in charge of recording the pedestrians' responses made many mistakes. Specifically, the collaborator missed half of the instances in which a pedestrian looked up and recorded these responses as "did not look up." Thus, even if the treatment (the investigator's looking up) truly had a strong effect on the outcome (other people's looking up), the misclassification of the outcome will result in a dilution of the association between treatment and the (mismeasured) outcome.

We say that there is measurement bias when the association between treatment and outcome is weakened or strengthened as a result of the process by which the study data are measured. Since measurement errors can occur under any study design—including randomized experiments and observational studies—measurement bias need always be considered when interpreting effect estimates. This chapter provides a description of biases due to measurement error.
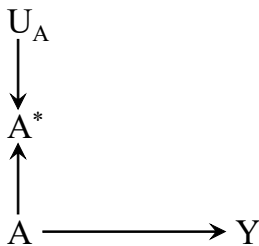
9.1 Measurement error

Figure 9.1

The term "misclassification" is synonymous for "measurement error" for discrete variables

In previous chapters we made the implicit assumption that all variables were perfectly measured. This assumption is generally unrealistic. For example, consider an observational study designed to estimate the effect of a cholesterol-lowering drug A on the risk of liver disease Y in which the information on drug use is obtained by medical record abstraction. We expect that treatment A will be measured imperfectly: the abstractor may make a mistake when transcribing the data, the physician may forget to write down that the patient was prescribed the drug, the patient may not take the prescribed treatment... Thus, the treatment variable in our analysis data set will not be the *true* use of the drug, but rather the *measured* use of the drug. We will refer to the measured treatment as A^* (read A-star), which will not necessarily equal the true treatment A for a given individual.

The causal diagram in Figure 9.1 depicts the variables A, A^*, and Y. For simplicity, we chose a diagram that represents no confounding or selection bias for the causal effect of A on Y. The true treatment A affects both the outcome Y and the measured treatment A^*. The causal diagram also includes the node U_A to represent all factors other than A that determine the value of A^*. We refer to U_A as the *measurement error* for A. Note that the node U_A is unnecessary in discussions of confounding (it is not part of a backdoor path) or selection bias (no variables are conditioned on) and therefore we omitted it from the causal diagrams in Chapters 7 and 8. For the same reasons, the determinants of the variables A and Y are not included in Figure 9.1. The psychological literature sometimes refers to A as the "construct" and to A^* as the "measure" or "indicator." The challenge in observational disci-
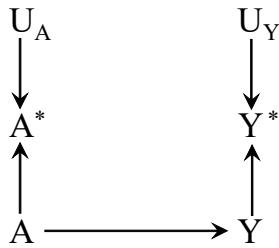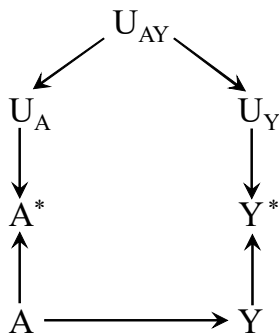
plines is making inferences about the unobserved construct (e.g., intelligence, cholesterol-lowering drug use) by using data on the observed measure (e.g., intelligence quotient estimated from questionnaire responses; information on statin use from medical records).

Besides treatment A, the outcome Y can be measured with error too. The causal diagram in Figure 9.2 includes the measured outcome Y^*, and the measurement error U_Y for Y. Figure 9.2 illustrates a common situation in practice. One wants to compute the average causal effect of the treatment A on the outcome Y, but these variables have not been, or cannot be, measured correctly. Rather, only the mismeasured versions A^* and Y^* are available to the investigator who aims at identifying the causal effect of A on Y.

Figure 9.2

Figure 9.2 represents a setting in which there is neither confounding nor selection bias for the causal effect of treatment A on outcome Y. According to our reasoning in previous chapters, association is causation in this setting. We can compute any association measure and endow it with a causal interpretation. For example, the associational risk ratio $\Pr[Y = 1|A = 1]/\Pr[Y = 1|A = 0]$ is equal to the causal risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$. Our implicit assumption in previous chapters, which we now make explicit, was that perfectly measured data on A and Y were available. We now consider the more realistic setting in which treatment or outcome are measured with error. Then there is no guarantee that the measure of association between A^* and Y^* will equal the measure of causal effect of A on Y. The associational risk ratio $\Pr[Y^* = 1|A^* = 1]/\Pr[Y^* = 1|A^* = 0]$ will generally differ from the causal risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$. We say that there is *measurement bias* or *information bias*. In the presence of measurement bias, the identifiability conditions of exchangeability, positivity, and well-defined interventions are insufficient to compute the causal effect of treatment A on outcome Y.

Figure 9.3

9.2 The structure of measurement error

The causal structure of confounding can be summarized as the presence of common causes of treatment and outcome, and the causal structure of selection bias can be summarized as conditioning on common effects of treatment and outcome (or of their causes). Measurement bias arises in the presence of measurement error, but there is no single structure to summarize measurement error. This section classifies the structure of measurement error according to two properties: *independence* and *nondifferentiality*.
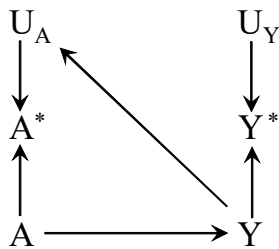
The causal diagram in Figure 9.2 depicts the measurement errors U_A and U_Y for both treatment A and outcome Y, respectively. According to the rules of d-separation, the measurement errors U_A and U_Y are independent because the path between them is blocked by colliders (either A^* or Y^*). Independent errors will arise if, for example, information on both drug use A and liver toxicity Y was obtained from electronic medical records in which data entry errors occurred haphazardly. In other settings, however, the measurement errors for exposure and outcome may be dependent, as depicted in Figure 9.3. For example, dependent measurement errors will occur if the information on both A and Y were obtained retrospectively by phone interview and a subject's ability to recall her medical history (U_{AY}) affected the measurement of both A and Y.

Both Figures 9.2 and 9.3 represent settings in which the error for treatment U_A is independent of the true value of the outcome Y, and the error for the
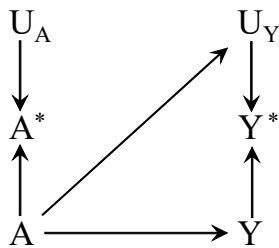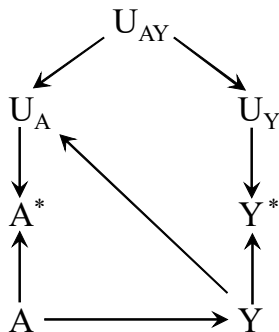

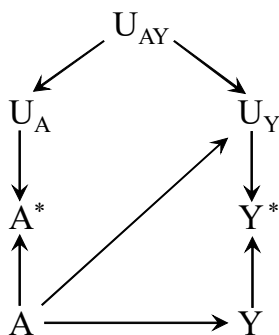Figure 9.4

Figure 9.5

Figure 9.6

Figure 9.7

outcome U_Y is independent of the true value of treatment. We then say that the measurement error for treatment is nondifferential with respect to the outcome, and that the measurement error for the outcome is nondifferential with respect to the treatment. The causal diagram in Figure 9.4 shows an example of independent but differential measurement error in which the true value of the outcome affects the measurement of the treatment (i.e., an arrow from Y to U_A). Some examples of differential measurement error of the treatment follow.

Suppose that the outcome Y were dementia rather than liver toxicity, and that drug use A were ascertained by interviewing study participants. Since the presence of dementia affects the ability to recall A, one would expect an arrow from Y to U_A. Similarly, one would expect an arrow from Y to U_A in a study to compute the effect of alcohol use during pregnancy A on birth defects Y if alcohol intake is ascertained by recall after delivery—because recall may be affected by the outcome of the pregnancy. The resulting measurement bias in these two examples is often referred to as *recall bias*. A bias with the same structure might arise if blood levels of drug A^* are used in place of actual drug use A, and blood levels are measured after liver toxicity Y is present—because liver toxicity affects the measured blood levels of the drug. The resulting measurement bias is often referred to as *reverse causation bias*.

The causal diagram in Figure 9.5 shows an example of independent but differential measurement error in which the true value of the treatment affects the measurement of the outcome (i.e., an arrow from A to U_Y). A differential measurement error of the outcome will occur if physicians, suspecting that drug use A causes liver toxicity Y, monitored patients receiving drug more closely than other patients. Figures 9.6 and 9.7 depict measurement errors that are both dependent and differential, which may result from a combination of the settings described above.

In summary, we have discussed four types of measurement error: independent nondifferential (Figure 9.2), dependent nondifferential (Figure 9.3), independent differential (Figures 9.4 and 9.5), and dependent differential (Figures 9.6 and 9.7). The particular structure of the measurement error determines the methods that can be used to correct for it. For example, there is a large literature on methods for measurement error correction when the measurement error is independent nondifferential. In general, methods for measurement error correction rely on a combination of modeling assumptions and validation samples, i.e., subsets of the data in which key variables are measured with little or no error. The description of methods to measurement error correction is beyond the scope of this book. Rather, our goal is to highlight that the act of measuring variables (like that of selecting subjects) may introduce bias. Realistic causal diagrams of observational studies need to simultaneously represent biases arising from confounding, selection, and measurement. The best method to fight bias due to mismeasurement is, obviously, to improve the measurement procedures.

9.3 Mismeasured confounders

Besides the treatment A and the outcome Y, the confounders L may also be measured with error. Mismeasurement of confounders will result in measurement bias even if both treatment and outcome are perfectly measured. To see this, consider the causal diagram in Figure 9.8, which includes the variables drug use A, liver disease Y, and history of hepatitis L. Individuals with

Technical Point 9.1

Independence and nondifferentiality. Let $f(\cdot)$ denote a probability density function (PDF). The measurement errors U_A for treatment and U_Y for outcome are independent if their joint PDF equals the product of their marginal PDFs, i.e., $f(U_Y, U_A) = f(U_Y)f(U_A)$. The measurement error U_A for the treatment is nondifferential if its PDF is independent of the outcome Y, i.e., $f(U_A|Y) = f(U_A)$. Analogously, the measurement error U_Y for the outcome is nondifferential if its PDF is independent of the treatment A, i.e., $f(U_Y|A) = f(U_Y)$.
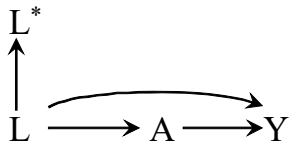

Figure 9.8

Figure 9.9

prior hepatitis L are less likely to be prescribed drug A and more likely to develop liver disease Y. As discussed in Chapter 7, there is confounding for the effect of the treatment A on the outcome Y because there exists a back-door path $A \leftarrow L \rightarrow Y$, but there is no unmeasured confounding given L because the backdoor path $A \leftarrow L \rightarrow Y$ can be blocked by conditioning on L. That is, there is exchangeability of the treated and the untreated conditional on the confounder L, and one can apply IP weighting or standardization to compute the average causal effect of A on Y. The standardized, or IP weighted, risk ratio based on L, Y, and A will equal the causal risk ratio $\Pr\left[Y^{a=1} = 1\right] / \Pr\left[Y^{a=0} = 1\right]$.

Again the implicit assumption in the above reasoning is that the confounder L was perfectly measured. Suppose investigators did not have access to the study participants' medical records. Rather, to ascertain previous diagnoses of hepatitis, investigators had to ask participants via a questionnaire. Since not all participants provided an accurate recollection of their medical history—some did not want anyone to know about it, others had memory problems or simply made a mistake when responding to the questionnaire—the confounder L was measured with error. Investigators had data on the mismeasured variable L^* rather than on the variable L. Unfortunately, the backdoor path $A \leftarrow L \rightarrow Y$ cannot be generally blocked by conditioning on L^*. The standardized (or IP weighted) risk ratio based on L^*, Y, and A will generally differ from the causal risk ratio $\Pr\left[Y^{a=1} = 1\right] / \Pr\left[Y^{a=0} = 1\right]$. We then say that there is *measurement bias* or *information bias*.
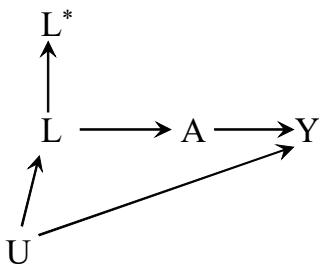
The causal diagram in Figure 9.9 shows an example of confounding of the causal effect of A on Y in which L is not the common cause shared by A and Y. Here too mismeasurement of L leads to measurement bias because the backdoor path $A \leftarrow L \leftarrow U \rightarrow Y$ cannot be generally blocked by conditioning on L^*. (Note that Figures 9.8 and 9.9 do not include the measurement error U_L because the particular structure of this error is not relevant to our discussion.)

Alternatively, one could view the bias due to mismeasured confounders in Figures 9.8 and 9.9 as a form of unmeasured confounding rather than as a form of measurement bias. In fact the causal diagram in Figure 9.8 is equivalent to that in Figure 7.5. One can think of L as an unmeasured variable and of L^* as a surrogate confounder (see Fine Point 7.1). The particular choice of terminology—unmeasured confounding versus bias due to mismeasurement of the confounders—is irrelevant for practical purposes.

Mismeasurement of confounders may also result in apparent effect modification. As an example, suppose that all study participants who reported a prior diagnosis of hepatitis ($L^* = 1$) and half of those who reported no prior diagnosis of hepatitis ($L^* = 0$) did actually have a prior diagnosis of hepatitis ($L = 1$). That is, the true and the measured value of the confounder coincide in the stratum $L^* = 1$, but not in the stratum $L^* = 0$. Suppose further

The strength and direction of measurement bias. In general, measurement error will result in bias. A notable exception is the setting in which A and Y are unassociated and the measurement error is independent and nondifferential: If the arrow from A to Y did not exist in Figure 9.2, then both the A-Y association and the A^*-Y^* association would be null. In all other circumstances, measurement bias may result in an A^*-Y^* association that is either further from or closer to the null than the A-Y association. Worse, for non-dichotomous treatments, measurement bias may result in A^*-Y^* and A-Y associations in opposite directions. This association or trend reversal may occur even under the independent and nondifferential measurement error structure of Figure 9.2 when the mean of A^* is a nonmonotonic function of A. See Dosemeci, Wacholder, and Lubin (1990) and Weinberg, Umbach, and Greenland (1994) for details. VanderWeele and Hernán (2009) described a more general framework using signed causal diagrams.

The magnitude of the measurement bias depends on the magnitude of the measurement error. That is, measurement bias generally increases with the strength of the arrows from U_A to A^* and from U_Y to Y^*. Causal diagrams do not encode quantitative information, and therefore they cannot be used to describe the magnitude of the bias.

that treatment A has no effect on any subject's liver disease Y, i.e., the sharp null hypothesis holds. When investigators restrict the analysis to the stratum $L^* = 1$, there will be no confounding by L because all participants included in the analysis have the same value of L (i.e., $L = 1$). Therefore they will find no association between A and Y in the stratum $L^* = 1$. However, when the investigators restrict the analysis to the stratum $L^* = 0$, there will be confounding by L because the stratum $L^* = 0$ includes a mixture of individuals with both $L = 1$ and $L = 0$. Thus the investigators will find an association between A and Y as a consequence of uncontrolled confounding by L. If the investigators are unaware of the fact that there is mismeasurement of the confounder in the stratum $L^* = 0$ but not in the stratum $L^* = 1$, they could naively conclude that both the association measure in the stratum $L^* = 0$ and the association measure in the stratum $L^* = 1$ can be interpreted as effect measures. Because these two association measures are different, the investigators will say that L^* is a modifier of the effect of A on Y even though no true effect modification exists.
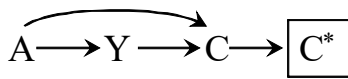
Finally, it is also possible that a collider C is measured with error as represented in Figure 9.10. In this setting, conditioning on the mismeasured collider C^* will still introduce selection bias because C^* is a common effect of the treatment A and the outcome Y.

Figure 9.10

9.4 Adherence to treatment in randomized experiments

Consider a marginally randomized experiment to compute the causal effect of heart transplant on 5-year mortality Y. So far in this book we have used the notation $A = 1$ to refer to the study participants who were assigned and therefore received the treatment of interest (heart transplant in this example), and $A = 0$ to the others. This choice of notation was appropriate for ideal randomized experiments in which all participants assigned to treatment actually received treatment, and in which all participants assigned to no treatment actually did not receive treatment. This notation, however is not detailed enough for real randomized experiments in which participants may not comply with the assigned treatment. In real randomized experiments we need to distinguish

between two versions of treatment: the assigned treatment Z and the received treatment A. Let A indicate, as usual, the treatment received (1 if the person receives a transplant, 0 otherwise). Let Z indicate treatment assigned (1 if the person is assigned to transplant, 0 otherwise). For a given individual, the value of Z and A may differ. For example, a subject randomly assigned to receive a heart transplant ($Z = 1$) may not receive it ($A = 0$) because he refuses to undergo the surgical procedure, or a subject assigned to medical treatment only ($Z = 0$) may still obtain a transplant ($A = 1$) outside of the study. The assigned treatment Z is a misclassified version of the treatment A that was truly received by the study participants. *Noncompliance* or lack of *adherence* with the assigned treatment is a special case of treatment misclassification that occurs in randomized experiments.

But there is a key difference between the assigned treatment Z and the misclassified treatments A^* that we have considered so far. The mismeasured treatment A^* considered in Figures 9.1-9.7 does not have a causal effect on the outcome Y. The association between A^* and Y is entirely due to their common cause A. In observational studies, one generally expects no causal effect of the measured treatment A^* on the outcome, even if the true treatment A has a causal effect. However, this is not the case for the special mismeasured treatment Z in randomized experiments (and that is why we elected to use the notation Z rather than A^* to refer to it). As shown in Figure 9.11, the assigned treatment Z may have a causal effect on the outcome Y through two different pathways. (Next section discusses the variable U in Figure 9.11.)
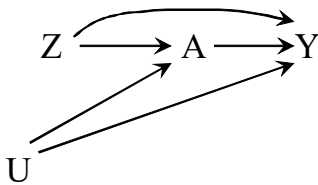
First, treatment assignment Z may affect the outcome Y through the effect of assigned treatment on received treatment A. Individuals assigned to heart transplant are more likely to receive a heart transplant, as represented by the arrow from Z to A. If receiving a heart transplant has a causal effect on mortality, as represented by the arrow from A to Y, then we conclude that assigned treatment Z has a causal effect on the outcome Y through the pathway $Z \to A \to Y$.

Second, treatment assignment Z may affect the outcome Y through pathways that are not mediated by received treatment A. That is, awareness of the assigned treatment might lead to changes in the study participants' behavior. For example, suppose those who are aware of receiving a transplant tend to spontaneously change their diet in an attempt to keep their new heart healthy, or that doctors take special care of patients who were not assigned to a heart transplant. These behavioral changes are represented by the direct arrow from Z to Y.

Thus the causal effect of the assigned treatment Z on Y combines the effect of Z via received treatment A and the concurrent behavioral changes. But suppose for a moment that the investigators were only interested in the causal effect of Z on Y that is mediated by A, uncontaminated by changes in behavior or care. Then investigators would withhold knowledge of the assigned treatment Z from participants and their doctors. For example, if Z were aspirin the investigators would administer an aspirin pill to those randomly assigned to $Z = 1$, and an identical pill (except that it does not contain aspirin but only some inert substance) to those assigned to $Z = 0$. A pill that looks like the active treatment but contains no active treatment is a *placebo*. Because neither a participant nor her doctor know whether the pill is the active treatment or a placebo, we would say that both participants and doctors are "blinded" and that the study is a *double-blind placebo-controlled* randomized experiment. The goal of this design is to ensure that the whole effect, if any, of the treatment assignment Z is solely due to the received treatment A. When placebo has

Figure 9.11

Technical Point 9.2

The exclusion restriction. Let $Y^{z,a}$ be the counterfactual outcome under randomized treatment assignment z and actual treatment received a. We say that the exclusion restriction holds when $Y^{z=0,a} = Y^{z=1,a}$ for all subjects and all values a and, specifically, for the value A observed for each subject.
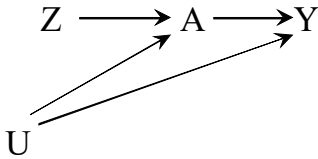

Figure 9.12

no effect, there is no direct arrow from Z to Y (Figure 9.12) and we say that the *exclusion restriction* holds. A blind treatment assignment, however, is sometimes unfeasible. For example, in our heart transplant study, there is no practical way of administering a convincing placebo for open heart surgery and lifelong immunosuppressive therapy. Other studies cannot be effectively blinded because the well known side effects of a treatment make apparent who is taking it.

Since there is no confounding, selection bias or measurement bias for Z in Figures 9.11 and 9.12, the association between Z and Y can be appropriately interpreted as the causal effect of Z on Y. The associational risk ratio $\Pr[Y = 1|Z = 1]/\Pr[Y = 1|Z = 0]$ equals the causal risk ratio $\Pr[Y^{z=1} = 1]/\Pr[Y^{z=0} = 1]$. When the exclusion restriction holds as in Figure 9.12, this effect measure quantifies the causal effect of Z on Y mediated by A—otherwise the effect measure quantifies a mixture of effects. But why would one be interested in the effect of assigned treatment Z rather than in the effect of the received treatment A? The next section provides some answers to this question.

9.5 The intention-to-treat effect and the per-protocol effect

The *per-protocol effect* is the causal effect of treatment if all individuals had adhered to their assigned treatment as indicated in the protocol of the randomized experiment. If all study participants adhered to the assigned treatment, the values of assigned treatment Z and received treatment A coincide for all participants, and therefore the per-protocol effect can be equivalently defined as either the average causal effect of Z or of A. Full adherence was implicit in Chapter 2, where we explained that, in ideal randomized experiments with perfect adherence to treatment, the treated $(A = 1)$ and the untreated $(A = 0)$ are exchangeable, $Y^a \amalg A$, and association is causation. The associational risk ratio $\Pr[Y = 1|A = 1]/\Pr[Y = 1|A = 0]$ is expected to equal the causal risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$, which measures the per-protocol effect on the risk ratio scale.

In general, the per-protocol effect cannot be validly estimated via a naïve "as treated" analysis, i.e., an analysis that estimates the unadjusted association between A and Y.

Consider now a setting in which not all subjects adhere to the assigned treatment so that the values of assigned treatment Z and received treatment A differ for some participants. Under imperfect adherence there is no guarantee that those who obtained a heart transplant $A = 1$ are a random sample of all subjects assigned to no heart transplant $Z = 0$. If, for example, the most severely ill subjects in the $Z = 0$ group seek a heart transplant $(A = 1)$ outside of the study, then the group $A = 1$ would include a higher proportion of severely ill subjects than the group $A = 0$. The groups $A = 1$ and $A = 0$ would not be exchangeable, and thus association between A and Y would not be causation. The associational risk ratio $\Pr[Y = 1|A = 1]/\Pr[Y = 1|A = 0]$ would not equal the per-protocol (causal) risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$.

Fine Point 9.2

Pseudo-intention-to-treat analysis. The ITT effect can only be computed in the absence of loss to follow-up or other forms of censoring. When some subjects do not complete the follow-up, their outcomes are unknown as thus the analysis needs to be restricted to subjects with complete follow-up. Thus, we can only compute the *pseudo-ITT effect* $\Pr[Y = 1|Z = 1, C = 0]/\Pr[Y = 1|Z = 0, C = 0]$ where $C = 0$ indicates that a subject remained uncensored until the measurement of Y. As described in Chapter 8, censoring may induce selection bias and thus the pseudo-ITT effect may be biased in either direction when compared with the true ITT effect. In the presence of loss to follow-up or other forms of censoring, the analysis of randomized experiments requires appropriate adjustment for selection bias even to compute the ITT effect.

In general, the per-protocol effect cannot be validly estimated via a naïve "per protocol" analysis, i.e., an analysis that estimates the un-adjusted association between A and Y among those who complied with their assigned treatment.
Hernán and Hernández-Díaz (2012) discuss the bias of naive as treated and per protocol analyses using causal diagrams.

The previous paragraph describes an example of *nonrandom noncompliance*, which arises when the reasons why participants receive treatment ($A = 1$) are not random but rather associated with the participants' prognosis U. Figure 9.11 depicts this situation with U representing severe illness (1: yes, 0: no). Nonrandom noncompliance implies confounding for the effect of A on Y, as indicated by the backdoor path $A \leftarrow U \rightarrow Y$ in Figure 9.11. In the presence of nonrandom noncompliance, unbiased estimation of the per-protocol effect may require adjustment for confounding, as if we were dealing with an observational study rather than with a randomized experiment.

On the other hand, as stated in the previous section, the effect of assigned treatment Z is not confounded. Because Z is randomly assigned, exchangeability $Y^z \perp\!\!\!\perp Z$ holds for the assigned treatment Z even if it does not hold for the received treatment A. Association between Z and Y implies a causal effect of Z on Y, whether or not all subjects adhere to the assigned treatment. However, effect measures for Z do not measure "the effect of treating with A" but rather "the effect of assigning participants to being treated with A" or "the effect of having the intention of treating with A." We therefore refer to the causal effect of randomized assignment Z as the *intention-to-treat effect*, or the ITT effect. Leaving aside random variability, the ITT effect on the risk ratio scale, $\Pr[Y^{z=1} = 1]/\Pr[Y^{z=0} = 1]$, is equal to the associational risk ratio $\Pr[Y = 1|Z = 1]/\Pr[Y = 1|Z = 0]$.

In general, the intention-to-treat effect can be validly estimated via a naïve "intention-to-treat" analysis, i.e., an analysis that estimates the unadjusted association between Z and Y.

Interestingly, the ITT effect is usually the primary, if not the only, effect measure reported in randomized trials. What is the justification to prefer the ITT effect over the per-protocol effect?

A frequent answer is that, in double-blind placebo-controlled randomized experiments, imperfect adherence results in an attenuation—not an exaggeration—of the effect. That is, the value of the ITT effect is guaranteed to be closer to the null than the value of the per-protocol effect For example, the ITT risk ratio $\Pr[Y = 1|Z = 1]/\Pr[Y = 1|Z = 0]$ will have a value between 1 and that of the causal risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$. The ITT effect can thus be interpreted as a lower bound for the per-protocol effect, i.e., as a conservative effect estimate.

There are, however, three problems with this answer. First, the answer assumes monotonicity of effects (see Technical Point 5.2), that is, that the treatment effect is in the same direction for all individuals. If this were not the case and the degree of noncompliance were high, then the per-protocol effect may be closer to the null, or even in the opposite direction, compared with the ITT effect. For example, suppose that 50% of the individuals assigned to treatment did not comply, and that the direction of the effect in those who

Under the sharp causal null hypothesis and the exclusion restriction, $\Pr[Y=1|Z=1]/\Pr[Y=1|Z=0] = \Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1] = 1$. That is, the ITT analysis preserves the null in double-blind placebo-controlled randomized experiments.In statistical terms, the ITT analysis provides a valid—though perhaps underpowered—α-level test of the null hypothesis of no average treatment effect.

Null preservation is a key property because it ensures no effect will be declared when no effect exists.

When treatment can vary over time, we define the per-protocol effect as the effect that would have been observed if everyone had adhered to their assigned treatment strategy throughout the follow-up. See Toh and Hernán (2008) for an example of adherence-adjustment in a randomized clinical trial with a time-varying treatment.

complied is opposite to that in those who did not comply. Then the ITT effect would be anti-conservative.

Second, even if the direction of the effect is the same for all individuals, the conservativeness of the ITT effect makes it a dangerous effect measure when the goal is evaluating a treatment's safety: one could naïvely conclude that a treatment A is safe because the ITT effect of Z on the adverse outcome is close to null, even if treatment A causes the adverse outcome in a fraction of the patients. The explanation may be that most subjects assigned to $Z=1$ did not take, or stopped taking, the treatment before developing the adverse outcome.

Third, even if the ITT effect is conservative in placebo-controlled experiments, it may not be when subjects are assigned to two active treatments. Suppose subjects with a chronic and painful disease were randomly assigned to either aspirin ($Z=1$) or ibuprofen ($Z=0$). The goal was to determine which drug results in a lower risk of severe pain Y after 1 year of follow-up. Unknown to the investigators, both drugs are equally effective to reduce pain, that is, the per-protocol (causal) risk ratio $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$ is 1. However, in this particular study, adherence to aspirin happened to be lower than adherence to ibuprofen. As a result, the ITT risk ratio $\Pr[Y=1|Z=1]/\Pr[Y=1|Z=0]$ was greater than 1, and the investigators wrongly concluded that aspirin was less effective than ibuprofen to reduce severe pain. See Fine Point 9.3 for more on this issue.

Thus the reporting of ITT effects as the primary findings from a randomized experiment is hard to justify for experiments that are not double-blinded placebo-controlled, and for those aiming at estimating the effect of a treatment's safety as opposed to a treatment's efficacy. Unfortunately, computing the per-protocol effect requires adjustment for confounding under the assumption of exchangeability conditional on the measured covariates, or via instrumental variable estimation (a particular case of g-estimation, see Chapter 16) under alternative assumptions.

In summary, in the analysis of randomized experiments there is trade-off between bias due to potential unmeasured confounding—when choosing the per-protocol effect—and misclassification bias—when choosing the ITT effect. Reporting only the ITT effect implies preference for misclassification bias over confounding, a preference that needs to be justified in each application.

Fine Point 9.3

Effectiveness versus efficacy. Some authors refer to the per-protocol effect, e.g., $\Pr[Y^{a=1}=1]/\Pr[Y^{a=0}=1]$ as the treatment's "efficacy," and to the ITT effect, e.g., $\Pr[Y^{z=1}=1]/\Pr[Y^{z=0}=1]$, as the treatment's "effectiveness." A treatment's "efficacy" closely corresponds to what we have referred to as the average causal effect of treatment A in an ideal randomized experiment. In contrast, a treatment's "effectiveness" would correspond to the effect of assigning treatment Z in a setting in which the interventions under study will no be optimally implemented, typically because a fraction of study subjects will not comply. Using this terminology, it is often argued that "effectiveness" is the most realistic measure of a treatment's effect because "effectiveness" includes any effects of treatment assignment Z not mediated through the received treatment A, and already incorporates the fact that people will not perfectly adhere to the assigned treatment. A treatment's "efficacy," on the other hand, does not reflect a treatment's effect in real conditions. Thus one is justified to report the ITT effect as the primary finding from a randomized experiment not only because it is easy to compute, but also because "effectiveness" is the truly interesting effect measure.

Unfortunately, the above argumentation is problematic. First, the ITT effect measures the effect of assigned treatment under the adherence conditions observed in a particular experiment. The actual adherence in real life may be different (e.g., participants in a study may comply better if they are closely monitored), and may actually be affected by the findings from that particular experiment (e.g., people will be more likely to comply with a treatment after they learn it works). Second, the above argumentation implies that we should refrain from conducting double-blind randomized clinical trials because, in real life, both patients and doctors are aware of the received treatment. Thus a true "effectiveness" measure should incorporate the effects stemming from assignment awareness (e.g., behavioral changes) that are eliminated in double-blind randomized experiments. Third, individual patients who are planning to adhere to the treatment prescribed by their doctors will be more interested in the per-protocol effect—the "efficacy" of treatment—than in the ITT effect. For more details, see the discussion by Hernán and Hernández-Díaz (2012).

Chapter 10
RANDOM VARIABILITY

Suppose an investigator conducted a randomized experiment to answer the causal question "does one's looking up to the sky make other pedestrians look up too?" She found an association between her looking up and other pedestrians' looking up. Does this association reflect a causal effect? By definition of randomized experiment, confounding bias is not expected in this study. In addition, no selection bias was expected because all pedestrians' responses—whether they did or did not look up—were recorded, and no measurement bias was expected because all variables were perfectly measured. However, there was another problem: the study included only 4 pedestrians, 2 in each treatment group. By chance, 1 of the 2 pedestrians in the "looking up" group, and neither of the 2 pedestrians in the "looking straight" group, was blind. Thus, even if the treatment (the investigator's looking up) truly had a strong average effect on the outcome (other people's looking up), half of the subjects in the treatment group happened to be immune to the treatment. The small size of the study population led to a dilution of the estimated effect of treatment on the outcome.

There are two qualitatively different reasons why causal inferences may be wrong: systematic bias and random variability. The previous three chapters described three types of systematic biases: selection bias, measurement bias—both of which may arise in observational studies and in randomized experiments—and unmeasured confounding—which is not expected in randomized experiments. So far we have disregarded the possibility of bias due to random variability by restricting our discussion to huge study populations. In other words, we have operated as if the only obstacles to identify the causal effect were confounding, selection, and measurement. It is about time to get real: the size of study populations in etiologic research rarely precludes the possibility of bias due to random variability. This chapter discusses random variability and how we deal with it.

10.1 Identification versus estimation

The first nine chapters of this book are concerned with the computation of causal effects in study populations of near infinite size. For example, when computing the causal effect of heart transplant on mortality in Chapter 2, we only had a twenty-subject study population but we regarded each subject in our study as representing 1 billion identical subjects. By acting as if we could obtain an unlimited number of individuals for our studies, we could ignore random fluctuations and could focus our attention on systematic biases due to confounding, selection, and measurement. Statisticians have a name for problems in which we can assume the size of the study population is effectively infinite: *identification* problems.

Thus far we have reduced causal inference to an identification problem. Our only goal has been to identify (or, as we often said, to compute) the average causal effect of treatment A on the outcome Y. The concept of identifiability was first described in Section 3.1—and later discussed in Sections 7.2 and 8.4—where we also introduced some conditions required to identify causal effects even if the size of the study population could be made arbitrarily large. These so-called identifying conditions were exchangeability, positivity, and well-defined interventions.

Our ignoring random variability may have been pedagogically convenient to introduce systematic biases, but also extremely unrealistic. In real research

projects the study population is not effectively infinite and hence, we cannot ignore the possibility of random variability. To this end let us return to our twenty-subject study of heart transplant and mortality in which 7 of the 13 treated subjects died.

Suppose our study population of 20 can be conceptualized as being a random sample from a *super-population* so large compared with the study population that we can effectively regard it as infinite. Then it is natural to want to make inferences about the super-population. For example, we may want to make inferences about the super-population probability (or proportion) $\Pr[Y = 1|A = a]$. We refer to the parameter of interest in the super-population, the probability $\Pr[Y = 1|A = a]$ in this case, as the *estimand*. An *estimator* is a rule that takes the data from any sample from the super-population and produces a numerical value for the estimand. This numerical value for a particular sample is the *estimate* from that sample. The sample proportion of subjects that develop the outcome among those receiving treatment level a, $\widehat{\Pr}[Y = 1 \mid A = a]$, is an estimator of the super-population probability $\Pr[Y = 1|A = a]$. The estimate from our sample is $\widehat{\Pr}[Y = 1 \mid A = a] = 7/13$. More specifically, we say that $7/13$ is a *point estimate*. The value of the estimate will depend on the particular 20 subjects randomly sampled from the super-population.

As informally defined in Chapter 1, an estimator is *consistent* for a particular estimand if the estimates get (arbitrarily) closer to the parameter as the sample size increases (see Technical Point 10.1 for the formal definition). Thus the sample proportion $\widehat{\Pr}[Y = 1 \mid A = a]$ consistently estimates the super-population probability $\Pr[Y = 1|A = a]$, i.e., the larger the number n of subjects in our study population, the smaller the magnitude of $\Pr[Y = 1|A = a] - \widehat{\Pr}[Y = 1 \mid A = a]$ is expected to be. Previous chapters were exclusively concerned with identification; from now on we will be concerned with statistical *estimation*.

For an introduction to statistics, see the book by Wasserman (2004). For a more detailed introduction, see Casella and Berger (2002).

Even consistent estimators may result in point estimates that are far from the super-population value. Large differences between the point estimate and the super-population value are much more likely to happen when the size of the study population is small compared with that of the super-population. Therefore it makes sense to have more confidence in estimates that originate from larger study populations. Statistical theory allows one to quantify this confidence in the form of a confidence interval around the point estimate. The larger the size of the study population, the narrower the confidence interval. A common way to construct a 95% confidence interval for a point estimate is to use a 95% Wald confidence interval centered at a point estimate. It is computed as follows.

First, estimate the standard error of the point estimate under the assumption that our study population is a random sample from a much larger super-population. Second, calculate the upper limit of the 95% Wald confidence interval by adding 1.96 times the estimated standard error to the point estimate, and the lower limit of the 95% confidence interval by substracting 1.96 times the estimated standard error from the point estimate. For example, consider our estimator $\widehat{\Pr}[Y = 1 \mid A = a] = \hat{p}$ of the super-population parameter $\Pr[Y = 1|A = a] = p$. Its standard error is $\sqrt{\frac{p(1-p)}{n}}$ (the standard error of a binomial) and thus its estimated standard error is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(7/13)(6/13)}{13}} = 0.138$. Recall that the Wald 95% confidence interval for a parameter θ based on an estimator $\widehat{\theta}$ is $\widehat{\theta} \pm 1.96 \times \widehat{se}\left(\widehat{\theta}\right)$ where $\widehat{se}\left(\widehat{\theta}\right)$ is an estimate of the (exact or

large sample) standard error of $\widehat{\theta}$. Therefore the 95% Wald confidence interval for our estimate is 0.27 to 0.81. The length and centering of the 95% Wald confidence interval will vary from sample to sample.

A 95% confidence interval is *calibrated* if the estimand is contained in the interval in 95% of random samples, conservative if the estimand is contained in more than 95% of samples, and anticonservative if contained in less than 95%. We will say that a confidence interval is *valid* if it is either calibrated or conservative, i.e. it covers the true parameter at least 95% of the time. We would like to choose the valid interval whose length is narrowest.

A Wald confidence interval centered at $\hat{p}$ can only be guaranteed to be valid in large samples. For simplicity, here we assume that our sample size was sufficiently large for the validity of our Wald interval.

The validity of confidence intervals is based on the variability of estimates over samples of the super-population, but we only see one of those samples when we conduct a study. Why should we care about what would have happened in other samples that we did not see? One answer is that the definition of confidence interval also implies the following. Suppose we and all of our colleagues keep conducting research studies for the rest of our lifetimes. In each new study, we construct a valid 95% confidence interval for the parameter of interest. Then, at the end of our lives, we can look back at all the studies we conducted, and conclude that the parameters of interest were trapped in—or covered by—the confidence interval in at least 95% of our studies. Unfortunately, we will have no way of identifying the 5% of our studies in which the confidence interval failed to include the super-population quantity.

Importantly, the 95% confidence interval from a single study does not imply that there is a 95% probability that the estimand is in the interval. In our example, we cannot conclude that the probability that the estimand lies between the values 0.27 and 0.81 is 95%. The estimand is fixed, which implies that either it is or it is not included in the interval (0.27, 0.81). The probability that the estimand is included in that interval is either 0 or 1. A confidence interval only has a *frequentist* interpretation. Its level (e.g., 95%) refers to the frequency with which the interval will *trap* the unknown super-population quantity of interest over a collection of studies (or in hypothetical repetitions of a particular study).

In contrast with a frequentist 95% confidence interval, a Bayesian 95% credible interval can be interpreted as "there is a 95% probability that the estimand is in the interval", but probability is defined as degree-of-belief. For the relation between confidence intervals and credible intervals, see Fine Point 11.1

Confidence intervals are often classified as either *small-sample* or *large-sample* (equivalently, asymptotic) confidence intervals. A small-sample valid (conservative or calibrated) confidence interval is one that is valid at all sample sizes for which it is defined. Small-sample calibrated confidence intervals are sometimes called *exact* confidence intervals. A large-sample valid confidence interval is one that is valid only in large samples. A large-sample exact 95% confidence interval is one whose coverage becomes arbitrarily close to 95% as the sample size increases. The Wald confidence interval for $\Pr[Y = 1 | A = a] = p$ mentioned above is a large-sample valid confidence interval, but not a small-sample valid interval. (There do exist small-sample valid confidence intervals for p, but they are not often used in practice.) When the sample size is small, a valid large-sample confidence interval, such as the Wald 95% confidence interval of our example above, may not be valid. In this book, when we use the term 95% confidence interval, we mean a large-sample valid confidence interval, like a Wald interval, unless stated otherwise. See also Fine Point 10.1.

There are many valid large-sample confidence intervals other than the Wald interval (Casella and Berger, 2002). One of these might be preferred over the Wald interval, which can be badly anti-conservative in small samples (Brown et al, 2001).

However, not all estimators can be used to center a valid Wald confidence interval, even in large samples. Most users of statistics will consider an estimator unbiased if it can center a valid Wald interval and biased if it cannot (see Technical Point 10.1 for details). For now, we will equate the term bias with the inability to center Wald confidence intervals.

10.2 Estimation of causal effects

Suppose our heart transplant study was a marginally randomized experiment, and that the 20 subjects were a random sample of all subjects in a nearly infinite super-population of interest. Suppose further that all subjects in the super-population were randomly assigned to either $A = 1$ or $A = 0$, and that all of them adhered to their assigned treatment. Exchangeability of the treated and the untreated would hold in the super-population, i.e., $\Pr[Y^a = 1] = \Pr[Y = 1|A = a]$, and therefore the causal risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$ equals the associational risk ratio $\Pr[Y = 1|A = 1]/\Pr[Y = 1|A = 0]$ in the super-population.

Because our study population is a random sample of the super-population, the sample proportion of subjects that develop the outcome among those with observed treatment value $A = a$, $\widehat{\Pr}[Y = 1 \mid A = a]$, is a consistent estimator or the super-population probability $\Pr[Y = 1|A = a]$. Because of exchangeability in the super-population, the sample proportion $\widehat{\Pr}[Y = 1 \mid A = a]$ is also a consistent estimator of $\Pr[Y^a = 1]$. Thus testing the causal null hypothesis $\Pr[Y^{a=1} = 1] = \Pr[Y^{a=0} = 1]$ boils down to comparing, via standard statistical procedures, the sample proportions $\widehat{\Pr}[Y = 1 \mid A = 1] = 7/13$ and $\widehat{\Pr}[Y = 1 \mid A = 0] = 3/7$. Standard statistical methods can also be used to compute 95% confidence intervals for the causal risk ratio and risk difference in the super-population, which are estimated by $(7/13)/(3/7)$ and $(7/13) - (3/7)$, respectively. Slightly more involved, but standard, statistical procedures are used in observational studies to obtain confidence intervals for standardized, IP weighted, or stratified association measures.

There is an alternative way to think about sampling variability in randomized experiments. Suppose only subjects in the study population, not all subjects in the super-population, are randomly assigned to either $A = 1$ or $A = 0$. Because of the presence of random sampling variability, we do not expect that exchangeability will exactly hold in our sample. For example, suppose that only the 20 subjects in our study were randomly assigned to either heart transplant ($A = 1$) or medical treatment ($A = 0$). Each subject can be classified as good or bad prognosis at the time of randomization. We say that the groups $A = 0$ and $A = 1$ are exchangeable if they include exactly the same proportion of subjects with bad prognosis. By chance, it is possible that 2 out of the 13 subjects assigned to $A = 1$ and 3 of the 7 subjects assigned to $A = 0$ had bad prognosis. However, if we increased the size of our sample then there is a high probability that the *relative* imbalance between the groups $A = 1$ and

Technical Point 10.1

Bias and consistency in statistical inference. We have discussed systematic bias (due to unknown sources of confounding, selection, or measurement error) and consistent estimators in earlier chapters. Here we discuss these and other concepts of bias, and describe how they are related.

To provide a formal definition of consistent estimator for an estimand θ, suppose we observe n independent, identically distributed (i.i.d.) copies of a vector-valued random variable whose distribution P lies in a set $\mathcal{M}$ of distributions (our model). Then the estimator $\widehat{\theta}_n$ is consistent for $\theta = \theta(P)$ if $\widehat{\theta}_n$ converges to θ in probability under P, i.e., $P \in \mathcal{M}$

$$| \Pr{}_P \left[\left\| \widehat{\theta}_n - \theta(P) \right\| > \varepsilon \right] | \to 0 \text{ as } n \to \infty \text{ for every } \varepsilon > 0.$$

The estimator $\widehat{\theta}_n$ is exactly unbiased under P if $\mathrm{E}_P \left[\widehat{\theta}_n \right] = \theta(P)$. The *exact bias* under P is the difference $\mathrm{E}_P \left[\widehat{\theta}_n \right] - \theta(P)$. Note that we denote the estimator by $\widehat{\theta}_n$ rather than by simply $\widehat{\theta}$ to emphasize that the estimate depends on the sample size n. On the other hand, the parameter $\theta(P)$ is a fixed, though unknown, quantity.

Systematic bias precludes both consistency and exact unbiasedness of an estimator. Because most studies have some degree of unknown systematic bias, we cannot actually expect that the 95% confidence intervals around the estimate $\hat{\theta}_n$ will really cover the parameter θ in at least 95% of the studies. That is, in reality, our actual intervals will generally be anti-conservative.

Consistent estimators are not guaranteed to center a valid Wald confidence interval. Most researchers (e.g., epidemiologists) will declare an estimator unbiased only if it can center a valid Wald confidence interval. As argued by Robins (1987), this definition of bias is essentially equivalent to the definition of uniform asymptotic unbiasedness because in general only uniformly asymptotic unbiased estimators can center a valid Wald interval. All inconsistent estimators (such as those resulting from unknown systematic bias), and some consistent estimators, are biased under this definition, which is the one we use in the main text.

$A = 0$ would decrease.

Under this conceptualization, there are two possible targets for inference. First, investigators may be agnostic about the existence of a super-population and restrict their inference to the sample that was actually randomized. This is referred to as *randomization-based inference*, and requires taking into account some technicalities that are beyond the scope of this book. Second, investigators may still be interested in making inferences about the super-population from which the study sample was randomly drawn. From an inference standpoint, this latter case turns out to be mathematically equivalent to the conceptualization of sampling variability described at the start of this section in which the entire super-population was randomly assigned to treatment. That is, randomization followed by sampling is equivalent to sampling followed by randomization.

See Robins (1988) for a discussion of randomization-based inference.

In many cases we are not interested in the first target. To see why, consider a study that compares the effect of two first-line treatments on the mortality of cancer patients. After the study ends, we may determine that it is better to initiate one of the two treatments, but this information is now irrelevant to the actual study participants. The purpose of the study was not to guide the choice of treatment for patients in the study but rather for a group of individuals similar to—but larger than—the studied sample. Heretofore we have assumed that there is a larger group—the super-population—from which the study participants were randomly sampled. We now turn our attention to the concept of the super-population.

10.3 The myth of the super-population

As discussed in Chapter 1, there are two sources of randomness: sampling variability and nondeterministic counterfactuals. Consider our estimate $\widehat{\Pr}[Y = 1 \mid A = 1] = \hat{p} = 7/13$ of the super-population risk $\Pr[Y = 1|A = a] = p$. Nearly all investigators would report a binomial confidence $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 7/13 \pm 1.96\sqrt{\frac{(7/13)(6/13)}{13}}$ for the probability p. If asked why these intervals, they would say it is to incorporate the uncertainty due to random variability. But these intervals are valid only if $\hat{p}$ has a binomial sampling distribution. So we must ask when would that happen. In fact there are two scenarios under which $\hat{p}$ has a binomial sampling distribution.

Robins (1988) discussed these two scenarios in more detail.

- *Scenario 1.* The study population is sampled at random from an essentially infinite super-population, sometimes referred to as the source or target population, and our estimand is the proportion $p = \Pr[Y = 1|A = 1]$ of treated subjects who developed the outcome in the super-population. It is then mathematically true that, in repeated random samples of size 13 from the treated subjects in the super-population, the number of subjects who develop the outcome among the 13 is a binomial random variable with success probability $\Pr[Y = 1|A = 1]$. As a result, the 95% Wald confidence interval calculated in the previous section is asymptotically exact for $\Pr[Y = 1|A = 1]$. This is the model we have considered so far.

- Scenario 2. The study population is not sampled from any super-population. Rather (i) each subject i among the 13 treated subjects has an individual nondeterministic (stochastic) counterfactual probability $p_i^{a=1}$ (ii) the observed outcome $Y_i = Y_i^{a=1}$ for subject i occurs with probability $p_i^{a=1}$ and (iii) $p_i^{a=1}$ takes the same value, say p, for each of the 13 treated subjects. Then the number of subjects who develop the outcome among the 13 treated is a binomial random variable with success probability p. As a result, the 95% confidence interval calculated in the previous section is asymptotically exact for p.

Scenario 1 assumes a hypothetical super-population. Scenario 2 does not. However, Scenario 2 is untenable because the probability $p_i^{a=1}$ of developing the outcome when treated will almost certainly vary among the 13 treated subjects due to between-individual differences in risk. For example we would expect the probability of death $p_i^{a=1}$ to have some dependence on a subject's genetic make-up. If the $p_i^{a=1}$ are nonconstant then the estimand of interest in the actual study population would generally be the average, say p, of the 13 $p_i^{a=1}$. But in that case the number of treated who develop the outcome is not a binomial random variable with success probability p, and the 95% confidence interval for p calculated in the previous section is not asymptotically exact (but rather asymptotically conservative.)

Therefore, any investigator who reports a binomial confidence interval for $\Pr[Y = 1|A = a]$, and who acknowledges that there exists between-individual variation in risk, must be implicitly assuming Scenario 1: the study subjects were sampled from a near-infinite super-population and that all inferences are concerned with quantities from that super-population. Under Scenario 1, the number with the outcome among the 13 treated is a binomial variable regardless of whether the underlying counterfactual is deterministic or stochastic.

Fine Point 10.2

Quantitative bias analysis. The width of the usual Wald-type confidence intervals is a function of the standard error of the estimator and thus reflects only uncertainty due to random error. However, the possible presence of systematic bias due to confounding, selection, or measurement is another important source of uncertainty around effect estimates. This uncertainty due to systematic bias is well recognized by investigators and usually a central part of the discussion section of scientific articles. However, most discussions revolve around informal judgments about the potential direction and magnitude of the systematic bias. Some authors argue that quantitative methods need to be used to produce intervals around the effect estimate that integrate random and systematic sources of uncertainty. These methods, referred to as quantitative bias analysis. See the book by Lash, Fox, and Fink (2009). Bayesian alternatives are discussed by Greenland and Lash (2008), and Greenland (2009a, 2009b).

An advantage of working under the hypothetical super-population scenario is that nothing hinges on whether the world is deterministic or nondeterministic. On the other hand, the super-population is generally a fiction; in most studies subjects are not randomly sampled from any near-infinite population. Why then has the myth of the super-population endured? One reason is that it leads to simple statistical methods.

A second reason has to do with generalization. As we mentioned in the previous section, investigators generally wish to generalize their findings about treatment effects from the study population (e.g., the 20 individuals in our heart transplant study) to some large target population (e.g., all immortals in the Greek pantheon). The simplest way of doing so is to assume the study population is a random sample from a large population of subjects who are potential recipients of treatment. Since this is a fiction, a 95% confidence interval computed under Scenario 1 should be interpreted as covering the super-population parameter had, often contrary to fact, the study subjects been sampled randomly from a near infinite super-population. In other words, confidence intervals obtained under Scenario 1 should be viewed as a what-if statements.

It follows from the above that an investigator might not want to entertain Scenario 1 if the size of the pool of potential recipients is not much larger than the size of the study population, or if there is selection bias, i.e., the target population of potential recipients is believed to differ from the study population to an extent that cannot be accounted for by sampling variability (see Fine Point 10.2).

We will accept that subjects were randomly sampled from a super-population, and explore the consequences of random variability for causal inference in that context. We first explore this question in a simple randomized experiment.

10.4 The conditionality "principle"

Table 10.1 summarizes the data from a randomized experiment to estimate the average causal effect of treatment A (1: yes, 0: no) on the 1-year risk of death Y (1: yes, 0: no). The experiment included 240 subjects, 120 in each treatment group. The associational risk ratio is $\Pr[Y = 1|A = 1]/\Pr[Y = 1|A = 0] = 24/42 = 0.57$. Suppose the experiment had been conducted in a super-population of near-infinite size, the treated and the untreated would

Table 10.1

	$Y = 1$	$Y = 0$
$A = 1$	24	96
$A = 0$	42	78

Table 10.2

L = 1	$Y = 1$	$Y = 0$
$A = 1$	4	76
$A = 0$	2	38

L = 0	$Y = 1$	$Y = 0$
$A = 1$	20	20
$A = 0$	40	40

be exchangeable, i.e., $Y^a \amalg A$, and the associational risk ratio would equal the causal risk ratio $\Pr\left[Y^{a=1} = 1\right] / \Pr\left[Y^{a=0} = 1\right]$. Suppose the study investigators computed a 95% confidence interval $(0.37, 0.88)$ around the point estimate 0.57 and published an article in which they concluded that treatment was beneficial because it reduced the risk of death by 43%.

However, the study population had only 240 individuals and is therefore likely that, due to chance, the treated and the untreated are not perfectly exchangeable. Random assignment of treatment does not guarantee exact exchangeability within subjects in the trial; it only guarantees that any departures from exchangeability are due to random variability rather than to a systematic bias. In fact, one can view the uncertainty resulting from our ignorance of the chance correlation between unmeasured baseline risk factors and the treatment A in the study sample as contributing to the length 0.51 of the confidence interval.

A few months later the investigators remember that information on a third variable, cigarette smoking L (1: yes, 0: no), had also been collected and decide to take a look at it. The study data, stratified by L, is shown in Table 10.2. Unexpectedly, the investigators find that the probability of receiving treatment for smokers $(80/120)$ is twice that for nonsmokers $(40/120)$, which suggests that the treated and the untreated are not exchangeable and thus that some form of adjustment for smoking is necessary. When the investigators adjust via stratification, he associational risk ratio in smokers, $\Pr[Y = 1|A = 1, L = 1]/\Pr[Y = 1|A = 0, L = 1]$, is equal to 1. The associational risk ratio in nonsmokers, $\Pr[Y = 1|A = 1, L = 0]/\Pr[Y = 1|A = 0, L = 0]$, is also equal to 1. Treatment has no effect in both smokers and nonsmokers, even though the marginal risk ratio 0.57 suggested a net beneficial effect in the study population.

These new findings are disturbing to the investigators. Either someone did not assign the treatment at random (malfeasance) or randomization did not result in approximate exchangeability (very very bad luck). A debate ensues among the investigators. Should they retract their article and correct the results? They all agree that the answer to this question would be affirmative if the problem were due to malfeasance. If that were the case, there would be confounding by smoking and the effect estimate should be adjusted for smoking. But they all agree that malfeasance is impossible given the study's quality assurance procedures. It is therefore clear that the association between smoking and treatment is entirely due to bad luck. Should they still retract their article and correct the results?

One investigator says that they should not retract the article. His argument goes as follows: "OK, randomization went wrong for smoking, but why should we privilege the adjusted over the unadjusted estimator? It is likely that imbalances on other unmeasured factors U cancelled out the effect of the chance imbalance on L, so that the unadjusted estimator is still the closer to the true value in the super-population." A second investigator says that they should retract the article and report the adjusted null result. Her argument goes as follows: "We should adjust for L because the strong association between L and A introduces confounding in our effect estimate. Within levels of L, we have mini randomized trials and the confidence intervals around the corresponding point estimates will reflect the uncertainty due to the possible U-A associations conditional on L."

To determine which investigator is correct, here are the facts of the matter. Suppose, for simplicity, the true causal risk ratio is constant across strata of L, and suppose we could run the randomized experiment trillions of times.

We then select only (i.e., condition on) those runs in which smoking L and treatment A are as strongly positively associated as in the observed data. We would find that the fraction of these runs in which any given risk factor U for Y was positively associated with A essentially equals the number of runs in which it was negatively associated. [This is true even if U and L are highly correlated in both the super-population and in the study data, and further-more both are correlated with A in the study data.] As a consequence, the adjusted estimate of the treatment effect is unbiased but the unadjusted esti-mate is greatly biased when averaged over these runs. Unconditionally—over all the runs of the experiment—both the unadjusted and adjusted estimates are unbiased but the variance of the adjusted estimate is smaller than that of the unadjusted estimate. That is, the adjusted estimator is both conditionally unbiased and unconditionally more efficient. Hence either from the conditional or unconditional point of view, the Wald interval centered on the adjusted esti-mator is the correct analysis and the article needs to be retracted. The second investigator is correct.

The idea that one should condition on the observed L-A association is an example of what is referred to in the statistical literature as *the conditionality principle*. In statistics, the observed L-A association is said to be an ancillary statistic for the causal risk ratio. The conditionality principle states that infer-ence on a parameter should be performed conditional on all ancillary statistics (see Technical Point 10.2 for details). The discussion in the preceding para-graph then implies that many researchers intuitively follow the conditionality principle when they consider an estimator to be biased if it cannot center a valid Wald confidence interval conditional on any ancillary statistics. That is, our previous definition of bias was not sufficiently restrictive. From now on, we will say that an estimator is unbiased if and only if it can center a valid Wald interval conditional on all ancillary statistics.

When confronted with the frequentist argument that "Adjustment for L is unnecessary because unconditionally—over all the runs of the experiment—the unadjusted estimate is unbiased, investigators that intuitively apply the conditionality principle would aptly respond "Why should the various L-A associations in other hypothetical studies affect what I do in my study? In my study L is a confounder and adjustment is needed to eliminate confounding bias." This is a convincing argument for both randomized experiments and observational studies when, as above, the number of measured confounders is not large. When the number of measured variables is large however, following the conditionality principle is no longer a wise strategy.

10.5 The curse of dimensionality

If the investigators had measured 100 pre-treatment variables rather than only one, then the pre-treatment variable L formed by combining the 100 variables $L = (L_1, ..., L_{100})$ has 2^{100} strata. When, as in this case, there are many possible combinations of values of the pretreatment variables, we say that the data is of *high dimensionality*. For simplicity, suppose that there is no multiplicative effect modification by L, i.e., the super-population risk ratio $\Pr[Y = 1|A = 1, L = l]/\Pr[Y = 1|A = 0, L = l]$ is constant across the 2^{100} strata. In particular, suppose that the constant stratum-specific risk ratio is 1.

The investigators debate again whether to retract the article and report

Technical Point 10.2

A formal statement of the conditionality principle. The likelihood for the observed data has three factors: the density of Y given A and L, the density of A given L, and the marginal density of L. Consider a simple example with one dichotomous L, exchangeability given L, and in which the parameter of interest is the stratum-specific causal risk ratio $sRR = \Pr(Y = 1|L = l, A = 1) / \Pr(Y = 1|L = l, A = 0)$ known to be constant across strata of L. Then the likelihood of the data is

$$\prod_{i=1}^{N}\prod_{i=1}^{N} f(Y_i|L_i, A_i; sRR, \mathbf{p}_0) \times f(A_i|L_i; \alpha) \times f(L_i; \rho)$$

where $p_0 = (p_{01}, p_{02})$ with $p_{0l} = \Pr(Y = 1|L = l, A = 0)$, α, and ρ are nuisance parameters associated with the conditional density of Y given A and L, the conditional density of A given L, and the marginal density of L, respectively. See, for example, Casella and Berger (2002).

The data on A and L are said to be exactly ancillary for the parameter of interest when, as in this case, the distribution of the data conditional on these variables depends on the parameter of interest, but the joint density of A and L does not share parameters with $f(Y_i|L_i, A_i; sRR, \mathbf{p}_0)$. The conditionality principle states that one should always perform inference on the parameter of interest conditional on any ancillary statistics.

their estimate of the stratified risk ratio. They have by now agreed that they should follow the conditionality principle because the marginal risk ratio 0.57 is biased. However, they notice that, when there are 2^{100} strata, a 95% confidence interval for the conditional risk ratio is much less precise than the marginal risk ratio. This is exactly the opposite of what was found when L had only 2 strata. In fact, the 95% confidence interval may be so wide as to be completely uninformative.

To see why, note that, because 2^{100} is much larger than the number of individuals (240), there will at most only a few strata of L that will contain both a treated and an untreated subject. Suppose only one of 2^{100} strata contains a single treated subject and and a single untreated subject, and no other stratum contains both a treated and untreated subject. Then the 95% confidence interval for the risk ratio conditional on the observed distribution of A within the 2^{100} strata of L is $(0, \infty)$ because in the single stratum with both a treated and an untreated subject, the empirical risk ratio could be ∞, 0, or 1 depending on the value of Y in each subject.

What should the investigators do? By trying to do the right thing— following the conditionality principle—in the simple setting with one dichotomous variable, they put themselves in a corner for the high-dimensional setting. This is the *curse of dimensionality*: conditional on all 100 covariates the marginal estimator is still biased, but now the conditional estimator is

Robins and Wasserman (1999) provide a technical description of the curse of dimensionality.

uninformative. This shows that, just because conditionality is compelling in simple examples, it should not be raised to a principle since it cannot be carried through high-dimensional models. Though we have discussed this issue in the context of a randomized experiment, our discussion applies equally to observational studies.

Finding a solution to the curse of dimensionality is not straightforward. One approach is to reduce the dimensionality of the data by excluding some variables from the analysis. Many procedures to eliminate variables from the analysis are *ad hoc*. For example, investigators often exclude variables in L that they believe to be unimportant or that happen to be weakly associated with the treatment A or the outcome Y in the study sample, where "weak association" is defined by using some arbitrary threshold (e.g., a p-value greater than 0.10).

Multiple authors have studied the problems of *ad hoc* or automatic variable selection. See Greenland (2008) for a list of citations.

Many software packages use automatic procedures to select the covariates to include in a model such as forward selection, backward selection, stepwise selection. These procedures do not preserve the interpretation of frequentist confidence intervals (see Chapter REF). When *ad hoc* or automatic procedures are employed, 95% confidence intervals tend to be too narrow and thus invalid: they fail to cover the causal parameter of interest at least 95% of the time. The degree of undercoverage will be greater when there is some degree of confounding in the super-population since, in that case, Wald confidence intervals will not be centered on an unbiased estimator of the causal effect.

Unfortunately, there is not much we can do about the curse of dimensionality because the statistical theory to provide correct (honest) confidence intervals for high-dimensional data is still under development. In practice, the most common approach to deal with the curse of dimensionality is to specify low-dimensional, parsimonious statistical models. Using such models results in increased precision of the estimates at the expense of potential bias if the models are incorrect. Part II of this book is devoted to models for causal inference.

Technical Point 10.3

Comparison between adjusted and unadjusted estimators. Consider a setting in which the marginal risk ratio RR and the stratified risk ratios sRR within any level of the variables L are equal. This would be the case in a randomized experiment, in which A and L are known to be independent in the super-population, with no multiplicative effect modification by L. The maximum likelihood estimator (MLE) $\widehat{sRR}_{MLE}$ of the stratified risk ratio sRR, which corresponds to the conditional estimator discussed in the text, is an unconditionally efficient (i.e., the most precise) estimator when the sample size n is much greater than the dimension of the nuisance parameter $\mathbf{p}_0$ (see Technical Point 10.2).

Because of the likelihood factorization , the MLE $\widehat{sRR}_{MLE}$ depends only on the first factor of the likelihood

$$\prod_{i=1}^{N} f\left(Y_i|L_i, A_i; sRR, \mathbf{p}_0\right).$$ That is, the MLE does not care about how L and A were generated. In particular, it does not matter whether α is known as in a randomized experiment, or unknown as in an observational study. Since the MLE is more efficient than the marginal risk ratio estimator $\widehat{RR}$ that ignores data on L, even statisticians who do not accept the conditionality principle will still prefer the stratified over the marginal estimator.

The reason that MLE is both unconditionally more efficient and conditionally less biased than the marginal estimator is not a coincidence. In fact, both properties of the MLE are logically equivalent. To show this we use the facts that $\widehat{RR}$ and $\widehat{sRR}_{MLE}$ have the same conditional variance, i.e., $var\left(\widehat{RR}|\mathbf{A}, \mathbf{L}\right) = var\left(\widehat{sRR}_{MLE}|\mathbf{A}, \mathbf{L}\right)$ and that the MLE is unbiased conditional on $(\mathbf{A}, \mathbf{L}) = (A_i, L_i)$, $i = 1, ..., n$, i.e., $\mathrm{E}\left\{\widehat{sRR}_{MLE}|\mathbf{A}, \mathbf{L}\right\} = sRR$. It then follows from the identities

$$var\left(\widehat{RR}\right) = \mathrm{E}\left[var\left(\widehat{RR}|\mathbf{A}, \mathbf{L}\right)\right] + var\left[\mathrm{E}\left\{\widehat{RR}|\mathbf{A}, \mathbf{L}\right\}\right]$$
$$var\left(\widehat{sRR}_{MLE}\right) = \mathrm{E}\left[var\left(\widehat{sRR}_{MLE}|\mathbf{A}, \mathbf{L}\right)\right]$$

that $var\left(\widehat{RR}\right) > var\left(\widehat{sRR}_{MLE}\right)$ if and only if $\mathrm{E}\left\{\widehat{RR}|\mathbf{A}, \mathbf{L}\right\} > 0$ with positive probability. The above expectations and variances are asymptotic; a more precise statement was provided by Robins and Morgenstern (1987).

But this argument breaks down with high-dimensional data. To see this, consider the case where L has 2^{100} joint strata so the dimension of the nuisance parameter $\mathbf{p}_0$ is 2^{100}. Because the MLE needs to estimate each of the 2^{100} nuisance parameters, little or no information is left in the data to estimate the parameter of interest RR so the unconditional variance of the MLE will be very large, even infinite. (Also, the MLE will fail to be asymptotically unbiased conditional on $\mathbf{A}, \mathbf{L}$). The variance of the marginal estimator is essentially unaffected by the dimension of $\mathbf{p}$ and thus will be more efficient than the MLE. The MLE is only guaranteed to be more efficient than the marginal estimator when the ratio of number of subjects to the number of parameters is large (a frequently used rule of thumb is a minimum ratio of 10, though the minimum ratio depends on the characteristics of the data). Note the marginal estimator uses prior information not used by the conditional estimator. In our example, the marginal estimator uses the information that A and L are known not to be associated in the super-population. Without this prior information the marginal estimator would not be an unbiased estimator of the sRR.