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 individuals. 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 Heterogeneity of treatment effects

Table 4.1

	V	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
Polyphemus	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 V for sex: $V = 1$ for females (referred to as women in this book) and $V = 0$ for males (referred to as men). For convenience, we have rearranged the table so that women occupy the first 10 rows, and men 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 $V = 1$. In this subset of the population, the risk of death under treatment is $\Pr[Y^{a=1} = 1|V = 1] = 6/10 = 0.6$ and the risk of death under no treatment is $\Pr[Y^{a=0} = 1|V = 1] = 4/10 = 0.4$. 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.

Our use of the terms "man" and "woman" in this chapter can be viewed as a slight abuse of notation because these deities are gods and goddesses, not men and 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 $V = 0$. In this subset of the population, the risk of death under treatment is $\Pr[Y^{a=1} = 1|V = 0] = 4/10 = 0.4$ and the risk of death under no treatment is $\Pr[Y^{a=0} = 1|V = 0] = 6/10 = 0.6$. 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.

See Section 6.6 for a structural classification of effect modifiers.

We are now ready to provide a definition of effect modifier. We say that V is a modifier of the effect of A on Y when the average causal effect of A on Y varies across levels of V. 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 V 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 V. Sex V 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 V. We only consider variables V that are not affected by treatment A as effect modifiers.

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

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

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.

In Table 4.1 the causal risk ratio is greater than 1 in women ($V = 1$) and less than 1 in men ($V = 0$). Similarly, the causal risk difference is greater than 0 in women ($V = 1$) and less than 0 in men ($V = 0$). That is, there is *qualitative effect modification* because the average causal effects in the subsets $V = 1$ and $V = 0$ 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 study, there is no additive effect modification by V because the causal risk difference among individuals with $V = 1$ equals that among individuals with $V = 0$, i.e., $0.9 - 0.8 = 0.1 = 0.2 - 0.1$. However, in this study there is multiplicative effect modification by V because the causal risk ratio among individuals with $V = 1$ differs from that among individuals with $V = 0$, i.e., $0.9/0.8 = 1.1 \neq 0.2/0.1 = 2$. 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.

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

4.2 Stratification to identify effect modification

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

A stratified analysis is the natural way to identify effect modification by measured variables (see also Fine Point 4.1). To determine whether V modifies the causal effect of A on Y, one computes the effect of A on Y in each level (stratum) of the variable V. 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 V. Because the 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 V 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 V because, if treatment assignment was random and unconditional, exchangeability is expected in every subset of the population. Thus the causal risk difference in women, $\Pr[Y^{a=1} = 1|V = 1] - \Pr[Y^{a=0} = 1|V = 1]$, is equal to the associational risk difference in women, $\Pr[Y = 1|A = 1, V = 1] - \Pr[Y = 1|A = 0, V = 1]$. And similarly for men. Thus, to identify effect modification by V in an ideal experiment with unconditional randomization, one just needs to conduct a stratified analysis, i.e., to compute the association measure in each level of the variable V. Stratification can be used to compute average causal effects in subsets of the population, but not individual effects (see Fine Points 2.1 and 3.2).

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 according to their passports: 20 are Greek ($V = 1$) and 20 are Roman ($V = 0$). 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.

Table 4.2
Stratum $V = 0$

	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

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 V modifies the effect of A on Y. The goal is to compute the causal effect of A on Y in the Greeks, $\Pr[Y^{a=1} = 1|V = 1] - \Pr[Y^{a=0} = 1|V = 1]$, and in the Romans, $\Pr[Y^{a=1} = 1|V = 0] - \Pr[Y^{a=0} = 1|V = 0]$. If these two causal risk differences differ, we will say that there is additive effect modification by

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 individuals does not equal the counterfactual risk had the treated individuals 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 pre-treatment variable V even if V 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 confusing. The effect modification is by unidentified variables that have a different distribution between the treatment groups.

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).

V. And similarly for the causal risk ratios if interested in multiplicative effect modification.

The procedure to compute the conditional risks $\Pr[Y^{a=1} = 1|V = v]$ and $\Pr[Y^{a=0} = 1|V = v]$ in each stratum v has two stages: 1) stratification by V, and 2) standardization by L (or, equivalently, IP weighting with weights depending on L). We computed the standardized risks in the Greek stratum ($V = 1$) in Chapter 2: the causal risk difference was 0 and the causal risk ratio was 1. Using the same procedure in the Roman stratum ($V = 0$), we can compute the risks $\Pr[Y^{a=1} = 1|V = 0] = 0.6$ and $\Pr[Y^{a=0} = 1|V = 0] = 0.3$. (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 $V = 0$. Because these effect measures differ from those in the stratum $V = 1$, we say that there is both additive and multiplicative effect modification by nationality V 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 $V = 0$ and $V = 1$.

Step 2 can be ignored when V is equal to the variables L that are needed for conditional exchangeability (see Section 4.4).

We have shown that, in our study population, nationality V modifies the effect of heart transplant A on the risk of death Y. However, we have made no claims about the causal mechanisms involved in such effect modification. In fact, it is possible that nationality is simply a marker for the causal factor that is truly responsible for the modification of the effect. For example, suppose that the quality of heart surgery is better in Greece than in Rome. One would then find effect modification by nationality. An intervention to improve the quality of heart surgery in Rome could eliminate the modification of the causal effect by passport-defined nationality. Whenever we want to emphasize this distinction, we will refer to nationality as a *surrogate effect modifier*, and to quality of care as a *causal effect modifier*.

See Section 6.6 for a representation of surrogate and causal effect modifiers using causal graphs.

Therefore, our use of the term effect modification by V does not necessarily

imply that V plays a causal role in the modification of the effect. To avoid potential confusions, some authors prefer to use the more neutral term "effect heterogeneity across strata of V" rather than "effect modification by V." The next chapter introduces "interaction," a concept related to effect modification, that does attribute a causal role to the variables involved.

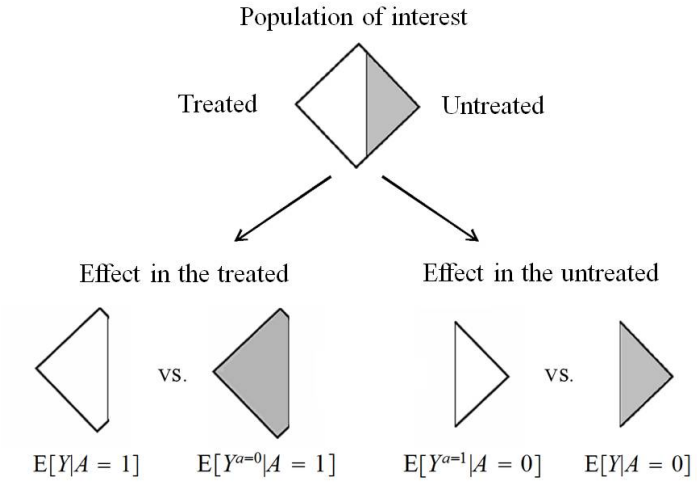

Figure 4.1

4.3 Why 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 V even in randomized experiments.

First, if a factor V modifies the effect of treatment A on the outcome Y then the average causal effect will differ between populations with different prevalence of V. For example, the average causal effect in the population of Table 4.1 is harmful in women and beneficial in men, i.e., there is qualitative effect modification. Because there are 50% of individuals 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. In the presence of non-qualitative effect modification, the magnitude, but not the direction, of the average causal effect may vary across populations. As examples of non-qualitative effect modification, consider the effects of asbestos exposure (which differ between smokers and nonsmokers) and of universal health care (which 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 $Y^a \perp\!\!\!\perp A|L$ for both $a = 0$ and $a = 1$. Computing the average causal effect in the treated only requires *partial exchangeability* $Y^{a=0} \perp\!\!\!\perp A|L$. 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 $Y^{a=1} \perp\!\!\!\perp A|L$.

We now describe how to compute counterfactual means of the form $\mathrm{E}\left[Y^a|A = a'\right]$ under the above assumptions of partial exchangeability. We do so via standardization and via IP weighting:

- Standardization: $\mathrm{E}[Y^a|A = a']$ is equal to $\sum_l \mathrm{E}\left[Y|A = a, L = l\right] \Pr\left[L = l|A = a'\right]$. See Miettinen (1972) and Greenland and Rothman (2008) for a discussion of standardized risk ratios.
- IP weighting: $\mathrm{E}[Y^a|A = a']$ is equal to the IP weighted mean $\dfrac{\mathrm{E}\left[\dfrac{I\left(A=a\right)Y}{f\left(A|L\right)} \Pr\left[A = a'|L\right]\right]}{\mathrm{E}\left[\dfrac{I\left(A=a\right)}{f\left(A|L\right)} \Pr\left[A = a'|L\right]\right]}$ with weights $\dfrac{\Pr\left[A = a'|L\right]}{f\left(A|L\right)}$. For dichotomous A, this equality was derived by Sato and Matsuyama (2003). See Hernán and Robins (2006a) for further details.

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

Hernán and VanderWeele (2011), Pearl and Bareinboim (2014), Dahabreh and Hernán (2019), and others have discussed effect modification in relation to transporting inferences across populations.

A setting in which transportability 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.

The extrapolation of causal effects computed in one population to a second population is referred to as *transportability* of causal inferences across populations (see Fine Point 4.2). In our example, the causal effect of heart transplant A on risk of death Y differs between men and women, and between Romans and Greeks. Thus the average causal effect in this population may not be transportable to other populations with a different distribution of effect modifiers such as sex and nationality.

Conditional causal effects in the strata defined by the effect modifiers may be more transportable than the causal effect in the entire population, but there is no guarantee that the conditional effect measures in one population equal the conditional effect measures in another population. This is so because there could be other unmeasured, or unknown, causal effect modifiers whose conditional distributions vary between the two populations (or for other reasons described in Fine Point 4.2). These unmeasured effect modifiers are not variables needed to achieve exchangeability, but just risk factors for the outcome. Therefore, transportability of effects across populations is a more difficult problem than the identification of causal effects in a single population: one would need to stratify not just on all those things required to achieve exchangeability (which you might have information about, say, by interviewing those who decide how to allocate the treatment) but on unmeasured causes of the outcome for which there is much less information.

Hence, transportability of causal effects is an unverifiable 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 transported to the Netherlands, but most experts would agree that the health effects of use of cholesterol-lowering drugs in Europeans can be transported to Canadians.

Second, evaluating the presence of effect modification is helpful to identify

the groups of individuals 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 ($V = 0$), and a harmful effect in women ($V = 1$). For example, if physicians knew that there is qualitative effect modification by sex, then, in the absence of additional information, they 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 individuals with $V = 0$ and also by 10% in individuals with $V = 1$, i.e., there is no additive effect modification by V because the causal risk difference is 0.1 in all levels of V. Thus, an intervention to treat all patients would be equally effective in reducing risk in both strata of V, despite the fact that there is multiplicative effect modification. In fact, if there is a nonzero causal effect in at least one stratum of V and the counterfactual risk $\Pr[Y^{a=0} = 1|V = v]$ varies with v, then effect modification is guaranteed on either the additive or the multiplicative scale.

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.

Additive, but not multiplicative, effect modification is the appropriate scale to identify the groups that will benefit most from intervention. In the absence of additive effect modification, learning that there is multiplicative effect modification may not be very helpful for decision making.

In our second example, the presence of multiplicative effect modification is expected because the risk under no treatment in the stratum $V = 1$ equals 0.8. Thus, the maximum possible causal risk ratio in the $V = 1$ stratum is $1/0.8 = 1.25$, which is guaranteed to differ from the causal risk ratio of 2 in the $V = 0$ stratum. In these situations, multiplicative effect modification arises from the differences in risk under no treatment $\Pr[Y^{a=0} = 1|V = v]$ across levels of V. Therefore, as a general rule, it is more informative to report the (absolute) counterfactual risks $\Pr[Y^{a=1} = 1|V = v]$ and $\Pr[Y^{a=0} = 1|V = v]$ in every level v of V, rather than simply their ratio or difference.

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 also be a first step towards characterizing the interactions between two treatments. The terms "effect modification" and "interaction" are sometimes used as synonymous in the scientific literature. This chapter focused on "effect modification." 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, i.e., 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 factor L via both standardization and IP weighting.

The present chapter adds another potential goal to the analysis: to identify effect modification by variables V. To achieve this goal, we need to stratify by

Fine Point 4.2

Transportability. Effects estimated in one population are often intended to make decisions in another population—the target population. Can we "transport" the effect from the study population to the target population? The answer depends on the characteristics of both populations. Specifically, transportability of causal effects across populations may be justified if the following characteristics are similar between the two populations:

- Effect modification: The causal effect of treatment 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. If the distribution of effect modifiers differs between populations, then the magnitude of the causal effect of treatment will differ too.
- Versions of treatment: The causal effect of treatment 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 will differ too (Hernán and VanderWeele, 2011).
- Interference: In the main text we focus on settings with no interference (Fine Point 1.1). Interference exists when treating one individual affects the outcome of others in the population. For example, a socially active individual may convince his friends to join him while exercising, and thus an intervention on that individual's physical activity may be more effective than an intervention on a socially isolated individual. Therefore, different contact patterns between populations will translate into causal effects of different magnitude.

A growing literature considers transportability methods that use data from the study population to estimate the causal effect in the target population in the presence of effect modification (e.g., Westreich et al. 2017, Rudolph and van der Laan 2017, Dahabreh et al. 2020b).

The transportability of causal inferences across populations may sometimes be improved by restricting our attention to the average causal effects in the strata defined by the effect modifiers, 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 individuals 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 between-population differences in the distribution of unmeasured effect modifiers, interference patterns, and distribution of versions of treatment.

V in addition to adjusting for L. For example, in this chapter we stratified by nationality V and adjusted for 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 V.

But stratification is not always used to identify effect modification by V. 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 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 individuals with $L = 0$, the other to individuals 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.

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 $Y^a \perp\!\!\!\perp A|L$, then $\Pr[Y=1|A=a,L=0]=\Pr[Y^a=1|L=0]$

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 $V=0$ because nationality V, by itself, is insufficient to guarantee conditional exchangeability.

When considering time-varying treatments, stratum-specific effect measures may not have a causal interpretation even under exchangeability, positivity, and well-defined interventions (Robins 1986, 1987). See Chapter 20.

That is, 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 V 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 like the odds ratio (see Fine Point 4.3) and inappropriate adjustment that leads to bias 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*. For causal inference, stratification is simply the application of restriction to several comprehensive and mutually exclusive subsets of the population, with exchangeability within each of these subsets. When positivity fails in some strata of the population, restriction is used to limit causal inference to those strata of the original population in which positivity holds (see Chapter 3).

4.5 Matching as another form of adjustment

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).

Our discussion on matching applies to cohort studies only. In case-control designs (briefly discussed in Chapter 8), we often match cases and non-cases (i.e., controls) rather than the treated and the untreated. Even if the matching factors suffice for conditional exchangeability, matching in cases and controls does not achieve unconditional exchangeability of the treated and the untreated in the matched population. Adjustment for the matching factors via stratification is required to estimate conditional (stratum-specific) effect measures.

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 individuals (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. See Stuart (2010) for an introduction.

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 individuals 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

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 weight-

Technical Point 4.2

Pooling of stratum-specific effect measures. Until Chapter 10, we avoid statistical considerations by assuming that we work with the entire population rather than with a sample. Thus we talk about computing causal effects rather than about (consistently) estimating them. In practice, however, we can rarely compute causal effects in the population. We estimate them from samples and wish to obtaining reasonably narrow confidence intervals around our effect estimates.

When dealing with stratum-specific effect measures, a common approach 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 there is no effect-measure modification, the pooled effect measure will be a more precise estimate of the common effect measure than each of the stratum-specific effect measures. Pooling methods (e.g., Woolf, Mantel-Haenszel, maximum likelihood) sometimes compute 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. Pooled effect measures can also be computed using regression models that include all possible product terms between all covariates L, but no product terms between treatment A and covariates L, i.e., models 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. The same issues arise in meta-analyses of studies with heterogeneous treatment effects (Dahabreh et al. 2020a).

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
Polyphemus	1	1	0
Persephone	1	1	0
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

ing 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 and positivity 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 and positivity in that subset only; to estimate the marginal effect in the entire population, exchangeability and positivity are required in all levels of L.

In the absence of effect modification, the effect measures (risk ratio or risk difference) 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 noncritical 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 $Z^a \perp\!\!\!\perp A|L$ and positivity hold. We use the risk ratio scale for no particular reason.

Standardization and IP weighting yield the average causal effect in the

Technical Point 4.3

Relation between marginal and conditional causal risk ratios. Suppose we wish to determine under which conditions the marginal risk ratio $\Pr\left[Y^{a=1}=1\right]/\Pr\left[Y^{a=0}=1\right]$ will be less than 1 given that we know the values of the conditional risk ratios $\Pr\left[Y^{a=1}=1|L=l\right]/\Pr\left[Y^{a=0}=1|L=l\right]$ for each stratum l. To do so, note that $\Pr\left[Y^{a=1}=1\right]/\Pr\left[Y^{a=0}=1\right] = \sum_l \left\{\Pr\left[Y^{a=1}=1|L=l\right]/\Pr\left[Y^{a=0}=1|L=l\right]\right\} w(l)$, with $w(l) = \left\{\Pr\left[Y^{a=0}=1|L=l\right]\Pr\left[L=l\right]\right\}/\Pr\left[Y^{a=0}=1\right]$ and $\sum_l w(l) = 1$. Substituting for $w(1)$ and $w(0)$ followed by some algebraic manipulations will provide the condition under which the inequality $\Pr\left[Y^{a=1}=1\right]/\Pr\left[Y^{a=0}=1\right] < 1$ holds.

In our data example, $\Pr\left[Y^{a=1}=1|L=l\right]/\Pr\left[Y^{a=0}=1|L=l\right]$ is 0.5 for $L=1$ and 2.0 for $L=0$. Therefore the marginal risk ratio will be less than 1 if and only if $\Pr\left[Y^{a=0}=1|L=1\right]/\Pr\left[Y^{a=0}=1|L=0\right] > 2\Pr\left[L=0\right]/\Pr\left[L=1\right]$.

Table 4.4

	V	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
Polyphemus	0	1	1
Hermes	0	1	0
Dionysus	0	1	1

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.

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 the ratio $\Pr\left[Z^{a=0}=1|L=1\right]/\Pr\left[Z^{a=0}=1|L=0\right]$ of the counterfactual risk under no treatment in the critical group to that in the noncritical group exceeds 2 times the odds $\Pr\left[L=0\right]/\Pr\left[L=1\right]$ of being in the noncritical group (see Technical Point 4.3). 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 investi-

gator 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. As we have discussed earlier, individual causal effects can only be identified under extreme assumptions. See Fine Points 2.1 and 3.2.

Fine Point 4.3

Collapsibility and the odds ratio. In the absence of multiplicative effect modification by V, 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 $\Pr[Y^{a=1}=1|V=v]/\Pr[Y^{a=0}=1|V=v]$ in every stratum v of V. 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 $V=1$ and $V=0$ 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 who were not depressed at baseline. The treatment A is 1 if the individual moved to a high altitude residence (on the top of Mount Olympus), 0 otherwise; the outcome Y is 1 if the individual subsequently developed depression, 0 otherwise; and V is 1 if the individual was a woman, 0 if a man. The decision to move was random, i.e., those more prone to develop depression were as likely to move as the others; effectively $Y^a \perp\!\!\!\perp A$. 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 $\Pr[Y=1|V=v,A=1]/\Pr[Y=1|V=v,A=0]$ is 2 for men $(V=0)$ and 3 for women $(V=1)$. The (conditional) causal odds ratio $\dfrac{\Pr[Y=1|V=v,A=1]/\Pr[Y=0|V=v,A=1]}{\Pr[Y=1|V=v,A=0]/\Pr[Y=0|V=v,A=0]}$ is 6 for men $(V=0)$ and 6 for women $(V=1)$. 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. The population causal odds ratio can be closer to the null value than the non-null stratum-specific causal odds ratio when V is an independent risk factor for Y and, as in our randomized experiment, A is independent of V (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.