Chapter 22
TARGET TRIAL EMULATION

As discussed in Part I, causal inference from observational data can be viewed as an attempt to emulate a hypothetical randomized trial, which we refer to as the target trial. However, Parts I and II only referred to simplistic target trials that compared time-fixed treatments. Since we now have all the tools that are needed to tackle causal inferences with time-varying treatments, we are now ready to discuss realistic target trials that compare sustained treatment strategies. This chapter generalizes the concept of the target trial to sustained treatment strategies and outlines a unified framework for causal inference, regardless of whether the data arose from a randomized experiment or an observational study.

This chapter also describes a taxonomy of causal effects that may be of interest when emulating a target trial, including observational analogs of intention-to-treat and per-protocol effects. Valid estimation of those causal effects generally requires data on time-varying prognostic factors and treatments, as well as appropriate adjustment for those time-varying factors using g-methods. It is precisely the development of g-methods that makes the concepts discussed here something more than a formal exercise: if data are available on all important fixed and time-varying confounders, the effects of interest can now be validly estimated.

22.1 Intention-to-treat effect and per-protocol effect

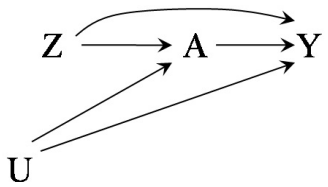

Figure 22.1

Consider a randomized trial in which individuals at risk of being infected by a dangerous virus are randomly assigned to a joint treatment of immediate vaccination plus an experimental antiviral therapy in case of being infected ($Z = 1$) or to standard of care ($Z = 0$), which includes no vaccination and no antiviral therapy. Figure 22.1 represents this trial with assigned treatment Z, received treatment A, and outcome (death) Y. For a given individual, the value of Z and A may differ because of lack of adherence to the assigned treatment some individuals assigned to vaccine ($Z = 1$) may not receive it ($A = 0$) because they refuse to be vaccinated, some individuals assigned to no vaccine ($Z = 0$) may still obtain a vaccine ($A = 1$) outside of the study. The variable U represents the unmeasured risk factors that influence an individual's decision to get vaccinated.

As shown in Figure 22.1, the assigned treatment Z can have a causal effect on the outcome Y through two different pathways. First, treatment assignment Z may affect the outcome Y simply because it affects the received treatment A. Individuals assigned to vaccine are more likely to receive a vaccine, as represented by the arrow from Z to A. If receiving a vaccine has a causal effect on mortality, as represented by the arrow from A to Y, then assignment to vaccine has a causal effect on the outcome Y through the pathway $Z \rightarrow A \rightarrow Y$.

Second, treatment assignment Z may affect the outcome Y through pathways that are not mediated by received treatment A. For example, awareness of the assigned treatment might lead to changes in the participants' behavior individuals aware of having been assigned to vaccination plus a promising antiviral therapy may become less careful about being infected. These behavioral changes are represented by the direct arrow from Z to Y.

Fine Point 22.1

The exclusion restriction (again). The existence of the arrow $Z \rightarrow Y$ in Figure 22.1 represents a direct effect of assignment on the outcome not through treatment. When this arrow exists, we say that the *exclusion restriction* does not hold. See Technical Point 16.1 for a formal discussion of the exclusion restriction.

Often investigators try to partly "de-contaminate" the effect of Z by eliminating the arrow $Z \rightarrow Y$ as shown in Figure 22.2 (same as Figure 16.1), which depicts the *exclusion restriction* of no direct arrow from Z to Y. To do so, they withhold knowledge of the assigned treatment Z from participants and their doctors. For example, investigators would administer the vaccine to those randomly assigned to $Z = 1$, and a *placebo* (an identical injection except that it does not contain vaccine) to those assigned to $Z = 0$. Because participants and their doctors do not know whether the injection they are given is the active treatment or a placebo, they are said to be "blinded" and the study is referred to as a *double-blind placebo-controlled* randomized trial. In Chapter 16, we used the concept of double-blind placebo-controlled randomized trial to motivate the concept of instrumental variable.

A double-blind treatment assignment is often unfeasible. Many studies cannot be effectively blinded because there is no practical way of administering a convincing placebo (e.g., for open heart surgery), because side effects of a treatment will make apparent who is taking it, etc. Also, blinding (and placebo control) is not advised when investigators are interested in quantifying the treatment effect in the real world, in which no blinding (or placebo) exists.

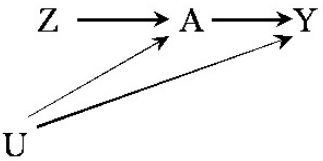

Figure 22.2

Hence, the causal effect of the assigned treatment Z depends not only on the strength of the arrow $A \longrightarrow Y$ (the effect of the received treatment), but also on the strength of the arrows $Z \longrightarrow A$ (the degree of adherence to the assigned treatment in the study) and $Z \longrightarrow Y$ (the concurrent behavioral changes). The effect of Z is not "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," which is why the effect of randomized assignment Z is often referred to as the *intention-to-treat effect*.

No confounding is expected for the effect of assigned treatment because Z is randomly assigned. Exchangeability $Y^z \perp\!\!\!\perp Z$ is expected to hold for the assigned treatment Z because there are no backdoor paths from Z to Y in Figure 22.1. Association between Z and Y implies a causal effect of Z on Y, whether or not all individuals adhered to the assigned treatment. The associational risk ratio $\Pr[Y = 1|Z = 1]/\Pr[Y = 1|Z = 0]$ equals the causal intention-to-treat risk ratio $\Pr[Y^{z=1} = 1]/\Pr[Y^{z=0} = 1]$. The analysis that estimates the unadjusted association between Z and Y to estimate the intention-to-treat effect is referred to as an *intention-to-treat analysis*. See Fine Point 22.2 for common variations of the intention-to-treat analysis that are generally biased.

Now consider the causal effect of treatment that would have been observed if all individuals had adhered to their assigned treatment as specified in the protocol of the experiment, which we refer to as the *per-protocol effect*. Throughout most of this book, we have assumed perfect adherence to the assigned treatment so that the values of assigned treatment Z and received treatment A coincide for all participants. that is, we assumed that U does not exist and thus the treated ($A = 1$) and the untreated ($A = 0$) are exchangeable, $Y^a \perp\!\!\!\perp A$.

The per-protocol effect is defined by the contrast $\Pr[Y^{z=1,a=1} = 1]$ vs. $\Pr[Y^{z=0,a=0} = 1]$ or, under the exclusion restriction, by the contrast $\Pr[Y^{a=1} = 1]$ vs. $\Pr[Y^{a=0} = 1]$. In the text we use the latter for notational simplicity.

Consider now a setting in which U represents high risk of infection (1: yes, 0: no) and in which individuals at high risk of infection ($U = 1$) in the $Z = 0$ group tend to seek vaccination ($A = 1$) outside of the study. If that occurs, then the group $A = 1$ would include a higher proportion of high-risk individuals than the group $A = 0$: the groups $A = 1$ and $A = 0$ would not be exchangeable, and thus the associational risk ratio $\Pr[Y = 1|A = 1]/\Pr[Y = 1|A = 0]$ would not equal the (causal) per-protocol risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$. As

Fine Point 22.2

Pseudo-intention-to-treat analysis and modified intention-to-treat analysis. An intention-to-treat analysis is unbiased for the intention-to-treat effect because it includes all randomized individuals. Therefore, variations of the intention-to-treat analysis that only include a subset of the randomized individuals may be biased.

When some individuals do not complete the follow-up, their outcomes are unknown and thus the analysis needs to be restricted to individuals with complete follow-up. Thus, we can only conduct a *pseudo-intention-to-treat analysis* $\Pr[Y = 1|Z = 1, C = 0]/\Pr[Y = 1|Z = 0, C = 0]$ where $C = 0$ indicates that an individual remained uncensored until the measurement of Y. As described in Chapter 8, censoring may induce selection bias and thus the pseudo-intention-to-treat estimate may be a biased estimate, in either direction, of the intention-to-treat effect. In the presence of loss to follow-up or other forms of censoring, the intention-to-treat analysis of randomized experiments requires appropriate adjustment for selection bias. See Section 21.5 and Little et al. (2012) for additional discussion.

For sustained treatment strategies, a common approach is to restrict the intention-to-treat analysis to individuals who at least initiated their assigned strategy (e.g., took at least one pill). This approach, known as a *modified intention-to-treat analysis*, includes only a subset of randomized individuals and may therefore be biased for the intention-to-treat effect. A modified intention-to-treat analysis generally requires adjustment for the risk factors that affect adherence.

indicated by the backdoor path $A \leftarrow U \rightarrow Y$, there is confounding for the effect of A on Y and estimating the per-protocol effect requires adjustment. That is, estimation of the per-protocol effect requires viewing the randomized experiment as an observational study. Fine Point 22.3 describes conventional approaches to quantify the per-protocol effect that missed this point.

The lack of confounding largely explains why the intention-to-treat effect is privileged in many randomized experiments: "the effect of having the intention of treating with A" may not be the effect that we want—"the effect of treating with A" or the per-protocol effect—but it is easier to compute. As often occurs when a less interesting quantity is easier to compute than a more interesting quantity, we tend to come up with arguments to justify the use of the less interesting quantity. The intention-to-treat effect is no exception. We now discuss why several well-known justifications for the intention-to-treat effect need to be taken with a grain of salt.

In statistical terms, the intention-to-treat analysis provides a valid—though perhaps underpowered—α-level test of the null hypothesis of no average treatment effect in double-blind placebo-controlled randomized experiments.

A common justification for the intention-to-treat effect is that it preserves the null. That is, if treatment A has a null effect on Y, then assigned treatment Z will also have a null effect on Y. *Null preservation* is a key property because it ensures no effect will be declared when no effect exists. More formally, under the sharp causal null hypothesis and the exclusion restriction, it can be shown that $\Pr[Y = 1|Z = 1]/\Pr[Y = 1|Z = 0] = \Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1] = 1$. However, this equality is not true when the exclusion restriction does not hold, as represented in Figure 22.1. In those cases—experiments that are not double-blind placebo-controlled—the effect of A may be null while the effect of Z is non-null. To see that, mentally erase the arrow $A \longrightarrow Y$ in Figure 22.1: there is still an arrow from Z to Y.

A related justification for the intention-to-treat effect is that its value is "closer to the null than the value of the per-protocol effect". The intuition is that, if imperfect adherence results in an attenuation—not an exaggeration—of the effect, the intention-to-treat risk ratio $\Pr[Y = 1|Z = 1]/\Pr[Y = 1|Z = 0]$ will have a value between 1 and that of the per-protocol risk ratio $\Pr[Y^{a=1} = 1]/\Pr[Y^{a=0} = 1]$. The intention-to-treat effect could thus be interpreted as a lower bound for the per-protocol effect, i.e., the intention-to-treat effect is a conservative estimate of the per-protocol effect. Unfortunately, the intention-

Fine Point 22.3

Naïve per-protocol analyses. In randomized trials, two common approaches to attempt to estimate the per-protocol effect of treatment A are "as treated" and so-called "per protocol" analyses.

A conventional *as-treated analysis* compares the distribution of the outcome Y in those who received treatment ($A = 1$) versus those who did not receive treatment ($A = 0$), regardless of their treatment assignment Z. Clearly, a conventional as-treated comparison will be confounded if the reasons that moved participants to take treatment were associated with prognostic factors U that were not measured, as in Figures 22.1 and 22.2. On the other hand, consider a setting in which all backdoor paths between A and Y can be blocked by conditioning on measured factors L, as in Figure 22.3. Then an as-treated analysis needs to adjust for the factors L.

A conventional per-protocol analysis—sometimes referred to as an *on-treatment analysis*—only includes individuals who adhered to the study protocol: the so-called per-protocol population of participants with $A = Z$. The analysis then compares, in the per-protocol population only, the distribution of the outcome Y in those who were assigned to treatment ($Z = 1$) versus those who were not assigned to treatment ($Z = 0$). That is, a conventional per-protocol analysis is just an intention-to-treat analysis restricted to the per-protocol population. This restriction will generally result in a biased estimate of the per-protocol effect. To see why, consider the causal diagram in Figure 22.4, which includes an indicator of selection S into the per-protocol population: $S = 1$ if $A = Z$ and $S = 0$ otherwise. Unless the per-protocol analysis appropriately measures and adjusts for the factors L, selection bias will arise because conditioning on $S = 1$ opens the noncausal path $Z \rightarrow A \leftarrow L \leftarrow U \rightarrow Y$.

That is, as-treated and per-protocol analyses are observational analyses of a randomized experiment and, like any observational analysis, require appropriate adjustment for confounding and selection bias to obtain valid estimates of the per-protocol effect. For examples and additional discussion, see Hernán and Hernández-Díaz (2012).

The argument against conservative intention-to-treat analyses applies to non-inferiority trials, in which the goal is to show that one treatment is not inferior to the other.

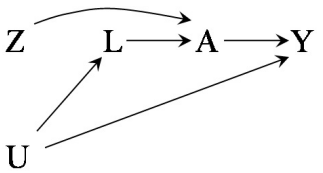

Figure 22.3

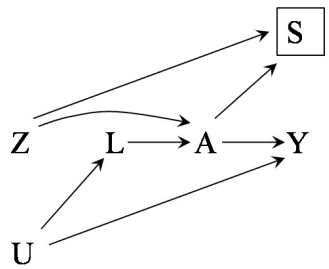

Figure 22.4

to-treat effect is not always conservative, because an attenuated effect is not guaranteed. See Fine Point 22.4

Even in settings in which the intention-to-treat is conservative, that may not be a good thing. Suppose that the goal is evaluating a treatment's safety: one could naïvely conclude that a treatment A is safe because the intention-to-treat effect of Z on the adverse outcome is close to null, even if treatment A causes the adverse outcome in a significant fraction of patients. The explanation may be that many individuals assigned to $Z = 1$ did not take, or stopped taking, treatment before developing the adverse outcome. Then the intention-to-treat effect would be a dangerous way to define the effect of treatment.

In summary, exclusive reliance on intention-to-treat effect estimates is hard to justify for randomized trials with substantial non-adherence and for those evaluating harms rather than benefits. The per-protocol effect is often a more natural estimand for researchers and decision makers (e.g., clinicians, patients). Estimating the per-protocol effect requires adjustment for confounding under the assumption of exchangeability conditional on the measured covariates, or under alternative assumptions such as those required for instrumental variable estimation (see Chapter 16).

The above discussion revolved largely around time-fixed treatments. When, as often happens, the randomized trial studies sustained strategies under which treatment can vary over time, the probability of non-adherence increases greatly. Then the intention-to-treat becomes increasingly noninformative compared with the per-protocol effect, defined as the effect that would have been observed if everyone had adhered to their assigned treatment strategy throughout the follow-up. Estimating the per-protocol effect for sustained strategies, both in a true randomized trial and in an observational analysis that emulates it, generally requires g-methods.

Fine Point 22.4

More misunderstandings about the intention-to-treat effect. A commonly heard argument is that the intention-to-treat effect measures treatment's *effectiveness* in the real world because it incorporates the fact that people will not perfectly adhere to the assigned treatment. In contrast, the per-protocol effect would measure treatment's *efficacy* under perfect adherence to treatment. Using this terminology, it is often argued that "efficacy" does not reflect a treatment's effect in real conditions, and thus one is justified to report the intention-to-treat effect as the primary finding from a randomized experiment because "effectiveness" is the most realistic measure of a treatment's effect.

This reasoning is problematic for several reasons. First, the intention-to-treat 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 adhere 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 adhere to a treatment after they learn it works). Second, if effectiveness is the goal, we should refrain from conducting double-blind placebo-controlled randomized clinical trials because, in real life, both patients and doctors are aware of the received treatment and no placebos are used. 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, individuals who are planning to adhere to the treatment prescribed by their doctors will be more interested in the per-protocol effect than in the intention-to-treat effect.

Another common argument is that the intention-to-treat effect is guaranteed to be conservative. This is not true in all settings. If the per-protocol effect of treatment is not monotonic (i.e, not in the same direction for all individuals; see Technical Point 5.2) and the degree of non-adherence is high, then the per-protocol effect may be closer to the null than the intention-to-treat effect. Even for monotonic effects, the intention-to-treat effect is not necessarily conservative in head-to-head trials in which individuals are assigned to one of two active treatments. Suppose individuals with a painful disease were randomly assigned to either an expensive drug ($Z = 1$) 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, i.e., the per-protocol risk ratio is 1. However, adherence to ibuprofen happened to be lower than adherence to the expensive drug because of a mild side effect that could be easily palliated. As a result, the intention-to-treat risk ratio was greater than 1, and the investigators wrongly concluded that ibuprofen was less effective than the expensive drug to reduce severe pain. For more details, see the discussion by Robins (1998b) and Hernán and Hernández-Díaz (2012).

22.2 A target trial with sustained treatment strategies

We are now ready to discuss target trials that compare sustained treatment strategies. Because the ultimate goal is to emulate these trials using real world observational data, we will only consider *pragmatic trials* with features that resemble the real world. In particular, participants and their treating physicians need to be aware of the treatment they receive (i.e., the treatment assignment is not blinded), nobody receives a placebo (i.e., both strategies g and g' involve either active treatments or no treatment), and participants are monitored as frequently and intensely as regular patients outside of the study. A trial with pragmatic features is preferable when the goal is quantifying the effects of treatment strategies under realistic conditions.

To fix ideas, consider a randomized trial to estimate the effect of antiretroviral therapy on the 5-year risk of death among individuals with HIV infection. Eligible participants—18 years and older, no AIDS, no previous use of antiretroviral therapy—are randomly assigned to either treatment strategy g or treatment strategy g' at the start of follow-up $k = 0$ (baseline). Their follow-up starts at the time of assignment and ends at death (the outcome of interest), loss to follow-up, or 60 months after baseline, whichever occurs earlier.

In previous chapters we considered the causal effect of treatment on an outcome Y measured at the end of follow-up. In this trial, the outcome is a failure time, i.e., time to death (see Technical Point 21.10).

Let A_k take value 1 if the individual receives therapy at time k and 0 otherwise, for $k = 0, 1, 2...K$ with $K = 59$. Our trial will assign eligible individuals to either the strategy g_1 "receive treatment $A_k = 1$ continuously during the follow-up unless a contraindication or toxicity arises" or the strategy g_0 "receive treatment $A_k = 0$ continuously during the follow-up". Let the assignment indicator Z takes value 1 if the individual is assigned to g_1 and 0 if assigned to g_0. Let D_k be an indicator for death (1: yes, 0: no) and C_k an indicator for censoring (1: yes, 0: no) by month $k = 1, 2...K + 1$.

Let us now define the intention-to-treat and per-protocol effects in a randomized trial with sustained treatment strategies. Additional contrasts of sustained strategies—referred to as *direct effects*—are described in Technical Point 22.1.

The *intention-to-treat effect* is contrast of the static strategies

- $(z = 1, \bar{c}_K = \bar{0})$: be assigned to strategy g_1 at baseline and remain under study until the end of follow-up
- $(z = 0, \bar{c}_K = \bar{0})$: be assigned to strategy g_0 at baseline and remain under study until the end of follow-up

The intention-to-treat effect at time k can then be expressed as the contrast of the counterfactual risks of death $\Pr\left[D_k^{z=1,\bar{c}_k=\bar{0}} = 1\right] - \Pr\left[D_k^{z=0,\bar{c}_k=\bar{0}} = 1\right]$ under assignment to strategy g_1 versus g_0 if nobody had been lost to follow-up through time k $(\bar{c}_k = \bar{0})$.

In some randomized trials, assignment to and initiation of the treatment strategies occur simultaneously. That is, all individuals assigned to strategy g_1 start to receive treatment at time 0, regardless of whether they continue taking it after baseline, and no individuals assigned to strategy g_0 receive treatment at time 0, regardless of whether they start taking it after baseline. In those cases, the intention-to-treat effect is not only the effect of assignment but also the effect of initiation of treatment $\Pr\left[D_k^{a_0=1,\bar{c}_k=\bar{0}} = 1\right] - \Pr\left[D_k^{a_0=0,\bar{c}_k=\bar{0}} = 1\right]$.

Like in any randomized trial, some participants will deviate from the protocol by not adhering to their assigned strategy. During the follow-up, some individuals assigned to g_1 will stop treatment for no clinical reason, some individuals assigned to g_0 will start treatment, some individuals will use non-approved concomitant treatments, etc. The intention-to-treat effect is agnostic about these protocol deviations, which are the result of decisions made after baseline. This agnosticism implies that the magnitude of the intention-to-treat effect may heavily depend on the particular patterns of protocol deviations that occur during the conduct of each trial. Two studies with the same protocol but conducted in different settings may have different intention-to-treat effect estimates and neither of them is biased. Due to the limitations of the intention-to-treat effect, we want to complement it with the per-protocol effect.

The *per-protocol effect* is defined by a contrast of the outcome distribution under the interventions:

- receive treatment strategy g_1 continuously between baseline $k = 0$ and end of follow-up
- receive treatment strategy g_0 continuously between baseline $k = 0$ and end of follow-up

The per-protocol effect at time k can then be expressed as the contrast of the counterfactual risks of death $\Pr\left[D_k^{g_1,\bar{c}_k=\bar{0}} = 1\right] - \Pr\left[D_k^{g_0,\bar{c}_k=\bar{0}} = 1\right]$ under full

Technical Point 22.1

Controlled direct effects. Consider the average causal effect of a treatment A on an outcome Y when a mediator M is set to a particular value. We refer to this quantity as the *direct effect* of A on Y not through M. If the mediator M could take two values (0 or 1), then we can define the direct effect of A on Y when M is set to 1 and the direct effect of A on Y when M is set to 0. On the additive scale, these two direct effects are defined by the counterfactual differences $\mathrm{E}\left[Y^{a=1,m=1}\right] - \mathrm{E}\left[Y^{a=0,m=1}\right]$ and $\mathrm{E}\left[Y^{a=1,m=0}\right] - \mathrm{E}\left[Y^{a=0,m=0}\right]$, respectively. These direct effects, which are often referred to as average *controlled direct effects*, could, in principle, be identified by conducting an experiment with sequential randomization for both treatment A and mediator M, or by emulating such target experiment using observational data. Technical Point 22.2 describes other types of direct effects for which no target experiment exists.

Suppose we conduct a randomized experiment in which participants are randomly assigned at baseline to either treatment $A = 1$ or $A = 0$ and one month after baseline to either treatment $M = 1$ or $M = 0$. Thus all individuals will be placed in one of four groups: $(A = 1, M = 1)$, $(A = 1, M = 0)$, $(A = 0, M = 1)$, or $(A = 0, M = 0)$. The outcome of interest Y is measured at 3 months in all individuals (for simplicity, suppose no individuals were lost to follow-up or died). This study design allows us to consistently estimate the controlled direct effects because the randomization of both A and M ensures that the counterfactual quantities $\mathrm{E}\left[Y^{a,m}\right] = \Pr\left[Y^{a,m} = 1\right]$ are consistently estimated by the observed risks $\Pr\left[Y = 1 | A = a, M = m\right]$.

The controlled direct effects can also be validly estimated in observational studies as long as the identifiability conditions of consistency, positivity, and exchangeability hold for both A and M. A precise characterization of these identifiability conditions was actually provided in Chapter 19 because a controlled direct effect is just a particular case of a contrast of treatment strategies sustained over time. To see so, simply replace A and M by A_0 and A_1 in the above expressions. More generally, both the treatment A and the mediator M can be time-varying themselves.

adherence to strategy g_1 versus g_0 if nobody had been lost to follow-up through time k ($\bar{c}_k = \bar{0}$).

Sensible trial protocols will not mandate that treatment be continued no matter what happens to the individual. For example, our strategy g_1 of continuous treatment mandates treatment discontinuation when a contraindication or toxicity arises. That is, the per-protocol effect generally involves the comparison of dynamic strategies ("do this, if X happens then do this other thing") rather than static strategies ("do this, no matter what happens").

Sometimes the study protocol is not explicit about the dynamic nature of the treatment strategies. For example, the protocol may simplify the description of strategy g_1 as "receive treatment $A_k = 1$ continuously during the follow-up" without explicitly stating that the therapy must be discontinued "when a contraindication or toxicity arises". This simplified description of strategy g_1 may lead to misunderstandings. Specifically, an individual assigned to g_1 who discontinues therapy because of toxicity should not be labeled as someone who is not adhering to strategy g_1. In fact, that person is perfectly adhering to strategy g_1 as (it should have been) stated in the protocol. When doing otherwise is not an option in the real world, discontinuation of the originally assigned treatment or initiation of other medically indicated treatments cannot possibly be considered a deviation from protocol. Because the per-protocol effect is defined by a contrast of realistic strategies, it is particularly relevant for causal inference research which seeks to provide evidence for decisions in the real world.

Ideally, to avoid confusions about what should or should not be deemed as nonadherence throughout the follow-up, the protocol would fully specify the treatment strategies of interest. Then the per-protocol effect would be well-defined (Hernán and Robins, 2017).

In fact, the per-protocol effect is often the implicit target of inference. For example, often investigators question the fidelity of the interventions implemented in the study to the interventions described in the protocol, and say that there is "bias". This language indicates that the investigators are really

Technical Point 22.2

Pure direct effects and principal stratum direct effects. Besides the controlled direct effects described in Technical Point 22.1, there exist other definitions of the average direct effect of a treatment A on an outcome Y when a potential mediator M is set to a particular value.

The *pure direct effect* (also known as *natural direct effect*) of A on Y not through M is the average causal effect of A on Y if the value of M had been set to the value that M would have taken if A had been set to 0, i.e., if M had been set to the value $M^{a=0}$ (which is 1 for some individuals and 0 for others). The pure direct effect, defined by the contrast $\mathrm{E}\left[Y^{a=1,M^{a=0}}\right] - \mathrm{E}\left[Y^{a=0,M^{a=0}}\right]$, is a cross-world quantity because $\mathrm{E}\left[Y^{a=1,M^{a=0}}\right]$ includes a counterfactual outcome simultaneously indexed by both $a = 1$ and $a = 0$. Therefore, the pure direct effect cannot be identified from a randomized experiment on A, M, or both, and cannot be identified from observational data under an FFRCISTG model (see Technical Point 6.2). Nonetheless, estimation of pure direct effects is often the goal of causal mediation analyses because total treatment effects can be decomposed into pure direct and total indirect effects. Pure direct effects were introduced by Robins and Greenland (1992); Pearl (2001) renamed them as natural direct effects and showed that, for certain causal graphs, the pure direct effect can be identified from the observed data under his NPSEM-IE model because, unlike the FFRCISTG model, the NPSEM-IE model assumes untestable cross-world independencies that cannot be refuted from randomized experiments on A, M, or both. For a review, see the book by VanderWeele (2015).

The *principal stratum direct effect* of A on Y if the value of M had been set to m is the average causal effect of A on Y in the subset of the population whose value of M would have been equal to m regardless of the value of A, i.e., in the subset of the population with $M^{a=0} = M^{a=1} = m$. Then the principal stratum direct effect is defined by the contrast $\mathrm{E}\left[Y^{a=1,m}|M^{a=0} = M^{a=1} = m\right] - \mathrm{E}\left[Y^{a=0,m}|M^{a=0} = M^{a=1} = m\right]$. Interestingly, this is equal to $\mathrm{E}\left[Y^{a=1}|M^{a=0} = M^{a=1} = m\right] - \mathrm{E}\left[Y^{a=0} = 1|M^{a=0} = M^{a=1} = m\right]$. Therefore, principal stratum direct effects do not involve joint counterfactuals $Y^{a,m}$, just the counterfactuals Y^a in a subset of the population so, in that sense, they are the total (rather than direct) effect of treatment in that subset of the population. It follows that, unlike controlled or pure direct effects, principal stratum direct effects do not require that interventions on M are well-defined. Principal stratum direct effects have little policy relevance when A affects M in almost all individuals, because then they apply to the very small subset of the population with $M^{a=0} = M^{a=1}$. In practice, M is often coarsened (typically into a binary indicator) to increase the size of the principal stratum, but coarsening itself may make the principal stratum direct effect less scientifically relevant (Robins et al. 2007). Principal stratum direct effects were introduced by Robins (1986) and popularized by Rubin (2004). Frangakis and Rubin (2002) following Robins (1986), used the concept of principal stratum as a tool to handle competing events. In Chapter 23, we consider an interventionist theory of mediation (Robins and Richardson 2010) which offers yet another type of direct effect.

interested in comparing the interventions implemented during the follow-up as specified in the protocol (i.e., the per-protocol effect) and not in the effect of assignment to the interventions at baseline (i.e., the intention-to-treat effect) because nonadherence after baseline cannot possibly bias the effect of assignment at baseline.

Finally, let us consider the effect of receiving interventions other than the ones specified in the study protocol. Suppose that, while our trial is being conducted, a consensus started to emerge that strategy g_0 "receive treatment $A_k = 0$ continuously during the follow-up" is inferior to strategy g_1. Therefore some physicians began to recommend initiation of therapy when the clinical course worsened when the CD4 cell count (L_k) first dropped below 200 cells/μL. As a result, many individuals in the trial who were assigned to strategy g_0 actually followed the modified strategy g_0' "receive treatment $A_k = 0$ continuously during the follow-up but, after $L_k < 200$, switch to treatment $A_k = 1$". The contrast of outcome distributions under the interventions

- receive treatment strategy g_1 continuously between baseline $k = 0$ and end of follow-up

- receive treatment strategy g_0' continuously between baseline $k = 0$ and end of follow-up

corresponds to neither the intention-to-treat effect nor the original per-protocol effect. Rather, it is a question about the per-protocol effect in a hypothetical target trial in which individuals are randomized to either strategy g_1 or g_0'.

This example illustrates how causal effects of interest that do not correspond to the original per-protocol effect can be conceptualized as per-protocol effects in target trials that can be emulated using the randomized trial data. Interestingly, if the strategies of interest differ from those in the actual trial, it is actually disadvantageous to have all participants in the actual trial adhere to the strategies specified in the protocol. Complete adherence implies that the trial data cannot be used to emulate a target trial with a different protocol (because no individuals followed the protocol of the new target trial in the actual data). For example, a randomized trial with full adherence in which individuals with HIV are assigned to different CD4 cell count thresholds at which to initiate antiretroviral therapy is of little use to emulate a trial in which individuals are assigned to either continuous treatment or no treatment, and vice versa. It is precisely the noncompliance that allows us to use the data from a given randomized trial to emulate other randomized trials that answer different, perhaps more relevant, causal questions.

In randomized trials with sustained treatment strategies, estimating per-protocol effects raises the same issues as any comparison of sustained strategies in an observational study. As we discuss later, valid estimation of the per-protocol effect generally demands that trial investigators collect post-randomization data on adherence to the strategy and on time-varying prognostic factors associated with adherence.

See Hernán and Robins (2017) for more details about the estimation of per-protocol effects in randomized trials.

22.3 Emulating a target trial with sustained strategies

If conducting a pragmatic randomized trial is not possible, we may attempt to emulate it through the analysis of existing observational data. We then refer to the trial as the *target trial* for our observational analysis.

Specifying the protocol of the target trial is a useful device to clarify the causal question of interest that we wish our observational analysis to answer. At the very least, we need to specify the following key components of the protocol: eligibility criteria, start and end of follow-up, treatment strategies, outcomes of interest, causal contrast, and data analysis plan. Note that a precise specification of the protocol of the target trial may require some exploration of the available data. For example, only after having determined that the data included information on HIV diagnosis, can we reasonably propose to emulate a target trial of individuals with HIV.

Analogs of the causal effects described in the previous sections for randomized trials can be proposed for observational analyses that emulate a target trial.

Emulating an intention-to-treat effect is rarely possible in observational analyses of existing data because the actual assignment to a treatment strategy is unknown. In our example, the closest observational analog of the intention-to-treat effect is a comparison of initiation of the different treatment strategies. A comparison of initiators parallels the intention-to-treat analysis in target trials in which assignment and initiation of the treatment strategies always occur

If we had data on prescription (rather than dispensing) of antiretroviral therapy, a comparison of groups according to whether they did or did not receive a prescription of therapy at baseline would be somewhat more analogous to the intention-to-treat analysis in the target trial.

together at baseline, regardless of whether individuals continue on the strategies after baseline. We can define this observational analog of the intention-to-treat effect by a contrast of the outcome distribution under the hypothetical interventions

- initiate treatment $A_0 = 1$ at baseline and remain under study until the end of follow-up
- initiate treatment $A_0 = 0$ at baseline and remain under study until the end of follow-up

This observational analog of the intention-to-treat effect at time k can then be expressed as the contrast of the counterfactual risks $\Pr\left[D_k^{a_0=1,\bar{c}_k=\bar{0}} = 1\right] - \Pr\left[D_k^{a_0=0,\bar{c}_k=\bar{0}} = 1\right]$. Unlike a true intention-to-treat effect that defines the groups according to assigned strategy, this contrast defines them according to initiation of each strategy. If we were using this contrast in a randomized trial, we would be including in the same group all individuals who did not take any dose of treatment at baseline, regardless of whether they were assigned to strategy g_1 or g_0. If initiation of treatment occurs shortly after assignment to treatment, our observational analog roughly preserves a key feature of the intention-to-treat effect: the contrast is defined by interventions occurring shortly after baseline.

An observational analysis that compares initiators is equivalent to the modified intention-to-treat analysis described in Fine Point 22.2

An observational analog of the per-protocol effect, on the other hand, is defined identically as that for the target trial. In randomized trials we differentiated between the original per-protocol effect and the per-protocol effects in alternative target trials. In observational studies this difference is unnecessary because, in the absence of a pre-specified protocol, each per-protocol effect corresponds to a particular target trial. In general, we can only use observational data to emulate target trials whose intended interventions are actually followed by at least some individuals in the study. In some settings, however, investigators may be willing to use modeling, e.g., dose-response structural models, to extrapolate beyond the interventions that are actually present in the data.

Defining the causal effects in observational studies in reference to those in the target trial forces us to be explicit about the strategies that are compared. This explicit specification of the treatment strategies prevents bias because it makes it obvious that certain data analyses involve comparisons that cannot be translated into a contrast between hypothetical interventions. These data analyses should therefore be avoided when the goal of the analysis is to help decision makers choose one of several courses of action, as we discussed in Sections 3.5 and 3.6.

Another advantage of an explicit definition of the treatment strategies in observational analyses is clarity. As discussed in Fine Point 22.4, some investigators insist in classifying causal effects into either "efficacy " (loosely defined: the effect of treatment that would be observed under perfect conditions) or "effectiveness" (loosely defined: the effect of treatment that would be observed under realistic conditions). Sometimes the intention-to-treat effect in a randomized trial is interpreted as the effectiveness of treatment and the per-protocol effect in the same trial as the efficacy of treatment. Other times the intention-to-treat effect in a randomized trial is interpreted as efficacy (even under imperfect conditions such as non-adherence) whereas the per-protocol effect in the observational study that emulates it is interpreted as effectiveness (even under perfect adherence). That is, especially in settings with sustained

strategies over long periods, the labels "effectiveness" and "efficacy" are ambiguous: it is often difficult to argue that either an intention-to-treat effect in a setting with nonadherence or a per-protocol effect in a real world setting measures the causal effect of treatment under perfect conditions.

Rather than insisting on an artificial efficacy-effectiveness dichotomy, it may be more helpful to accept that all causal effects are placed somewhere along the effectiveness continuum. An explicit definition of the treatment strategies that define the causal effect of interest is then more informative because decision makers need information about the effect of well-defined interventions.

22.4 Time zero

A crucial component of target trial emulation is the determination of the start of follow-up, also referred to as baseline or time zero, in the observational analysis. Eligibility criteria need to be met at that point but not later; study outcomes begin to be counted after that point but not earlier.

In randomized experiments, the time zero for each individual is the time when they are assigned to a treatment strategy while meeting the eligibility criteria. For example, in our randomized trial of antiretroviral therapy, time zero is the time when the treatment strategies are assigned (the time of randomization), which usually occurs shortly before, or at the same time as, treatment is initiated. We do not start the follow-up, say, 2 years before or after treatment assignment. Starting before randomization would not be reasonable because the treatment strategies had yet to be assigned and the eligibility criteria have not yet been defined, much less met; starting follow-up after randomization is potentially biased as deaths during the first two years of the trial would be excluded from the analysis and any short-term effects of treatment would be missed. Even more problematic, if treatment does indeed have a short-term effect, then more susceptible individuals would have died by year 2 in the group assigned to active treatment but not in the other group. This differential proportion of susceptible individuals after two years destroys the baseline comparability achieved by randomization and opens the door to selection bias.

The same rules regarding time zero apply to observational analyses and randomized trials, and for the same reasons. Generally, the follow-up in the observational analysis should start at the time the follow-up would have started in the target trial. Otherwise the effect estimates may be hard to interpret and biased because of selection affected by treatment. Nonetheless, in observational studies for causal inference, errors in the emulation of time zero of the target trial are very frequent. These errors occur because of two common problems: 1) sometimes there is not a unique choice of time zero, and 2) sometimes the treatment strategies cannot be uniquely assigned at time zero. We now describe solutions for each of these two problems.

Example: The highly publicized discrepancy between the estimates of the effect of postmenopausal hormone therapy on heart disease in observational studies and a randomized trial was partly due to mishandling of time zero in the former (Hernán et al. 2008).

First, the problem of non-unique time zero. Consider two scenarios, according to how many times the eligibility criteria can be met throughout an individual's lifetime:

1. Eligibility criteria can be met at a single time. This is the simplest setting. Follow-up starts at the only time the eligibility criteria are met. For example, consider a study in persons with HIV to compare immediate initiation of antiretroviral therapy when the CD4 cell count first drops

below 500 cells/μL versus delayed initiation when the CD4 cell count first drops below 350 cells/μL. The follow-up of eligible individuals starts the first time their CD4 cell count drops below 500.

2. Eligibility criteria can be met at multiple times. This is the situation that often leads to confusion. For example, consider a study to compare initiation versus no initiation of hormone therapy among postmenopausal women with no history of chronic disease and no use of hormone therapy during the previous two years. If a woman meets these eligibility criteria continuously between age 51 and 65, when should her follow-up start? At age 51, 52, 53...? In the target trial a woman would be eligible to be recruited at multiple times during her lifetime, i.e., she has multiple eligible times.

In settings with multiple eligibility times, there are several alternatives to choose the time zero of each individual among her eligible times. One could choose as time zero: a) the first eligible time, b) a randomly chosen eligible time, c) every eligible time, etc. Strategy c) requires emulating multiple sequential target trials, each of them with a different start of follow-up. The number of sequential trials depends on the frequency with which data on treatment and covariates are collected:

- If fixed schedule for data collection at pre-specified times (e.g., every two years, like in many epidemiologic cohorts), then emulate a new trial starting at each pre-specified time.

- If subject-specific schedule for data collection (e.g., electronic medical records), then choose a fixed time unit (e.g., a day, week or month), and emulate a new trial starting at each time unit.

An unbiased choice of time unit can vary from study to study. For example, consider a study in which both the time-varying treatment and confounders change more then once a week for many individuals. Then choosing a week or a month as the time unit will introduce bias. This bias could be eliminated by using by the choice of a day as the unit of time. If daily data on treatment and confounders are not available, the bias could not be fully corrected.

From a statistical standpoint, the sequential emulation strategy c) can be more efficient than the previous ones because it uses more of the available data. However, because individuals may be included in multiple target trials, appropriate adjustment of the variance of the effect estimate is required. This can be achieved by bootstrapping the entire analysis.

Second, let us talk about how to tackle the impossibility of assigning a unique treatment strategy to each individual. Consider a target trial in which individuals whose CD4 cell count just dropped below 500 cells/μL are assigned to one of the following strategies: (1) start therapy immediately, (2) start therapy when CD4 cell count drops below 350 cells/μL, (3) start therapy when CD4 cell drops below 200 cells/μL. When emulating this target trial using observational data, we will find individuals who started therapy at time zero (i.e, when their CD4 cell count first dropped below 500 cells/μL) and therefore we will assign them to strategy (1). Other individuals, however, did not start therapy at time zero, which means that their data are compatible with following both strategy (2) and strategy (3) at baseline. Which strategy should we assign them to?

One possibility is to choose a single strategy at random and assign them to that strategy, but that would be statistically inefficient. Another possibility is to create two exact copies—clones—of each of these individuals in the data and assign each of the two clones to a different strategy. Clones are then censored at the time their data stop being consistent with the arm they were assigned to. For example, if the individual does not start therapy when CD4 drops to 350, then the clone assigned to "start therapy when CD4 cell count drops to

Fine Point 22.5

Grace periods. Consider a trial to compare immediate initiation of antiretroviral therapy at time zero versus delayed initiation. In the real world, antiretroviral therapy cannot be started exactly on the same day that it is assigned. Depending on the health care system, it may take weeks or months until the requisite clinical and administrative procedures are completed and patients are adequately informed. Therefore, investigators need to define a grace period (say, 3 months) after time zero during which initiation is still considered to be immediate. Otherwise the study would be estimating the effect of strategies that do not occur frequently in reality or that could not be successfully implemented in practice.

A consequence of using a grace period is that an individual's observed data is consistent with more than one strategy for the duration of the grace period. For example, in the above study, the introduction of a 3-month grace period implies that the interventions are redefined as "initiate therapy within 3 months of time zero cells/μL" versus "never initiate therapy". Therefore individuals who start therapy in month 3 after baseline have data consistent with both strategies during months 1 and 2. Had some of them died during those 2 months, to which strategy should we have assigned those deaths? As described in the text, we could randomly assign these individuals to one of the two strategies or, better, we could create two clones of each individual and assign each of the two clones to a different strategy. Clones are censored when their data are no longer compatible with their assigned strategy. For example, if the individual starts therapy in month 3, then the clone assigned to "start after 3 months" would be censored at that time. The potential bias introduced by censoring can be handled via IP weighting.

When using grace periods with cloning and censoring, the intention-to-treat effect cannot be estimated because almost everyone will contribute a clone to each of the treatment strategies. Because each individual is assigned to all strategies at baseline, a contrast based on baseline assignment (i.e., an "intention-to-treat analysis") will compare groups with essentially identical outcomes. Therefore, analyses with grace period at baseline are geared towards estimating some form of per-protocol effect and thus will generally need to incorporate adequate adjustment.

Finally, note that a well-defined initiation strategy with a grace period should specify the timing of initiation during the grace period. For details, see the Appendix in Cain et al. (2010).

For a description of the cloning + censoring + weighting procedure, see Robins et al. (2008) and Cain et al. (2010). For related work, see van der Laan and Petersen (2007).

350" would be censored at that time. The potential bias introduced by this likely informative censoring would need to be corrected by adjusting for time-varying factors via IP weighting. Importantly, if the individual had died before either clone was censored, then both clones would have died and therefore the death would have been assigned to both strategies. This double allocation of events prevents the bias that could arise if events occurring during the waiting period were systematically assigned to one of the two strategies only.

Again, because individuals may be included multiple times in the analysis via their clones, appropriate adjustment of the variance of the effect estimate is required via bootstrapping. The cloning + censoring + weighting procedure can be combined with sequential target trial emulation when the eligibility criteria can be met at multiple times. Fine Point 22.5 describes the handling of strategies that can be initiated during a grace period after time zero rather than exactly at time zero.

22.5 A unified approach to answer What If questions with data

This book describes and integrates two causal inference frameworks: counterfactuals and causal diagrams. Explicit target trial emulation recapitulates both frameworks and grounds them to actionable causal inference. By organizing causal inference around a deeply familiar scientific concept—the experiment—

the target trial framework helps investigators use their subject-matter knowledge to articulate well-defined causal inference questions. Once the causal question is stated with little ambiguity, study design and data analysis flow naturally.

The target trial framework is applicable to a wide range of causal questions across many disciplines, regardless of the terminology and methodology privileged in each field. For example, economists often refer to confounding and conditional exchangeability as *omitted variable bias* and *selection on observables*, respectively, and traditional social scientists are unlikely to use g-methods because their causal questions are not typically organized around time-varying treatments. But these disciplinary differences are superficial compared with the fundamental task that all health and social scientists interested in causal inference face: they all need to articulate their causal questions as a contrast of well-defined counterfactuals. The target trial framework facilitates that task by helping define the well-defined interventions that lead to well-defined counterfactuals.

The target trial framework also provides a common language to unify the causal analysis of randomized and observational studies. Aside from baseline randomization, there are no other necessary differences between analyses of observational data that emulate a target trial and of true randomized trials (see Fine Point 22.6). That is, a randomized trial can be viewed as a follow-up study with baseline randomization and observational longitudinal data as a follow-up study without baseline randomization.

The similarities between follow-up studies with and without baseline randomization are increasingly apparent in the health and social sciences as a growing number of randomized experiments attempt to estimate the effects of sustained treatment strategies over long periods in real world settings. These studies are a far cry from the short experiments in tightly controlled settings that put randomized trials at the top of the hierarchy of study designs in the mid-20th century. For causal questions involving treatment strategies sustained over long periods, randomized experiments with the potential for substantial deviations from protocol (e.g., imperfect adherence to the assigned strategy, loss to follow-up) are subject to confounding and selection biases that we have learned to associate exclusively with observational studies.

Time-varying confounding in observational studies is a bias with the same structure as nonrandom noncompliance in randomized trials.

In particular, when estimating a per-protocol effect, both randomized trials and observational studies may need adjustment for time-varying prognostic factors that predict drop-out (selection bias) and treatment (confounding). That is, the methodology for causal inference described in this book applies equally to the per-protocol analyses of randomized trials and observational studies. And, for the same reasons that success is not guaranteed when estimating causal effects from observational data, the per-protocol effect estimates from randomized trials may be biased too.

In view of these similarities, one might expect that randomized experiments and observational studies would be analyzed similarly, except adjustment for baseline confounders in observational analyses to estimate the analog of the intention-to-treat effect. In practice, however, the typical analyses of randomized experiments and observational studies differ radically, which is both perplexing and, as we argue below, problematic.

A natural question is whether the "intention-to-treat analysis" and the so-called "per-protocol analysis" commonly used in randomized trials validly estimate the intention-to-treat effect and per-protocol effect, respectively.

A typical intention-to-treat analysis compares the distribution of outcomes between randomized groups without any form of adjustment for confounding

Fine Point 22.6

How do the data of randomized experiments and observational studies differ? Only three things distinguish the data from randomized experiments and observational studies. In randomized experiments, (i) no baseline confounding is expected because of randomization, (ii) the randomization probabilities are known, and (iii) the assignment to a treatment strategy is known for each individual at baseline.

An observational analysis can emulate (i) if one measures and appropriately adjusts for a sufficient set of covariates, and (ii) if the model for treatment assignment given the past is correctly specified. Interestingly, (iii) is not necessary for estimating the per-protocol effect in either randomized experiments or observational studies because efficient estimators (that are functions of the sufficient statistic) do not use this information. That is, the analyst does not need to know the strategies being compared, much less who was assigned to which strategy: in a randomized trial, you can delete the randomization assignment from the dataset and still estimate a per-protocol effect if a sufficient set of confounders was measured. In a trial of dynamic strategies with perfect adherence, a sufficient set is all time-fixed and time-varying covariates used by the strategies in assigning treatment (Robins 1986).

or selection bias. Lack of adjustment for baseline confounding is justified by randomization: the randomized groups are expected to be exchangeable because they are expected to have the same risk of the outcome if both groups had been assigned to the same treatment strategy. No adjustment for post-randomization confounding (e.g., due to nonadherence) is required because, again, there cannot be post-randomization confounding for the effect of baseline assignment.

However, baseline randomization cannot ensure exchangeability between those who are and are not lost to follow-up after randomization. Because the strategies that define the intention-to-treat effect require that the individuals remain in the study until their outcome variable can be ascertained, an intention-to-treat effect estimate calculated among those who are not lost to follow-up may be affected by post-randomization selection bias if prognostic factors influence, or are associated with, differential loss to follow-up. Therefore, valid estimation of the intention-to-treat effect may require an "intention-to-treat analysis" adjusted for post-randomization (time-varying) prognostic factors to eliminate selection bias from loss to follow-up. For example, in a randomized trial of antiretroviral therapy among HIV patients, g-methods will be needed if the probability of dropping out of the study is influenced by the onset of symptoms or other risk factors for the outcome.

Fine Point 22.2 refers to an intention-to-treat analysis that does not even attempt to adjust for selection bias as a pseudo-intention-to-treat analysis.

In addition to the primary intention-to-treat analysis, many randomized trials also report the results from a so-called per-protocol analysis restricted to individuals who adhered to the instructions specified in the study protocol, as described in Fine Point 22.3 for point interventions. For sustained treatment strategies, individuals are censored at the first time they deviate from the protocol. That is, the remaining per-protocol population at each time is the set of individuals that are still adhering to the protocol. No adjustment of any kind is performed. This unadjusted analysis is questionable for three reasons.

Fine Point 22.3 refers to a per-protocol analysis that does not even attempt to adjust for confounding as a *naïve per-protocol analysis*.

First, like in an intention-to-treat analysis, there may be selection bias due to differential loss to follow-up. If so, adjustment for post-baseline (time-varying) risk factors via g-methods will be needed.

Second, the analysis partly disregards the randomized groups and therefore the subset of individuals who remain on protocol under one strategy may not be exchangeable with the subset on protocol under another strategy. That is, this "per-protocol analysis" is akin to an observational analysis and thus requires

g-methods to adjust for bias due to time-varying risk factors that affect the decision to stay on protocol. Instrumental variable estimation (Chapter 16) can sometimes be used to validly estimate per-protocol effects without explicit adjustment for any variables, but the validity of these methods depends on having a valid instrument and on strong modeling assumptions. Some forms of instrumental variable estimation are a particular case of g-estimation (see Technical Point 16.6).

For failure time outcomes, g-methods are always needed when the treatment has a causal effect on the outcome. The reason is that treatment A_k affects all variables after time k through its effect on the time-varying indicator D_{k+1}, as discussed in Technical Point 21.10.

Third, this conventional per-protocol analysis ignores that the sustained treatment strategies under comparison are dynamic strategies. A common mistake is censoring individuals who discontinue treatment as if treatment discontinuation were, by definition, a deviation from protocol—which is why this analysis is also known as on-treatment analysis. We have discussed above that individuals who stop treatment because of toxicity or a contraindication are not deviating from protocol and therefore should not be censored.

All the above considerations apply to the analysis of both randomized trials and observational data to emulate a target trial. When the goal is estimating a per-protocol effect or its observational analog, the analysis of randomized trials and observational studies should be identical. If we feel compelled to adjust for time-varying confounding and selection bias in the analysis of observational studies, we should feel equally compelled to adjust for post-randomization confounding and selection bias in the analysis of randomized trials. Adjustment for time-varying factors using g-methods will generally be necessary for per-protocol analyses of both randomized trials and observational studies. The target trial framework and g-methods make it possible to implement a unified approach to causal inference for sustained treatment strategies. Historically, randomized experiments have been considered far superior to observational studies for the purpose of making causal inferences and aiding decision-making. Unfortunately, randomized experiments are not always available because they may be expensive, infeasible, unethical, or just untimely to support an urgent decision. Therefore, as much as we value the benefits of randomization, it is a fact that many decisions will need to be made in the absence of evidence from randomized trials. When we cannot conduct the randomized experiment that would answer our causal question, we resort to attempting to emulate it using observational data. It is therefore important to use a sound approach to design and analyze observational studies. Making the target trial explicit is one step in that direction. When the goal is to assist decision making, the analysis of existing observational data need to explicitly emulate a trial and be evaluated with respect to how well they emulate their target trial.

Under some extremely rare circumstances, decisions based on quality randomized trials may be inferior to decisions based on severely confounded observational data, as described in Fine Points 22.7 and 22.8.

Fine Point 22.7

A counterintuitive comparison of a randomized trial and an observational study. An untested over-the-counter treatment A was used by many individuals with lung cancer in a country. This worried the country's drug regulator who, in response, funded a double-blind placebo-controlled randomized trial of A in a random sample of 20% of individuals diagnosed with lung cancer over the next year. All trial participants adhered to their assigned treatment. The 60-month mortality risk was 55% in the treatment arm and 45% in the placebo arm as shown in the table below:

	$A=0$	$A=1$
$Y=1$	450	550
$Y=0$	550	450
	1000	1000

As a result, the regulator banned the treatment. Later, an observational study was conducted on the 80% of lung cancer patients not selected into the trial. This study found a mortality risk of 0% in both the treated and the untreated over the same period as the trial. How can the observational and the randomized trial data be reconciled?

Let us first remember that individuals can be classified into counterfactual types: 1) "doomed" ($Y^{a=0}=Y^{a=1}=1$), 2) "hurt" ($Y^{a=0}=0, Y^{a=1}=1$), 3) "helped" ($Y^{a=0}=1, Y^{a=1}=0$), and 4) "immune" ($Y^{a=0}=Y^{a=1}=0$). By random sampling and randomization, the following three groups have the same distribution of counterfactual types: the 1000 individuals treated in the trial, the 1000 individuals untreated in the trial, and the 8000 individuals in the observational study. The key observayion is that the 0% mortality in the observational study implies (i) there are no "doomed" individuals, (ii) all individuals with $Y^{a=1}=1$ must have been of type "hurt" and received $A=0$, and (iii) all individuals with $Y^{a=0}=1$ must have been of type "helped" and received $A=1$.

We next use these observations to reconstruct the trial data by counterfactual type. As argued above, the 550 individuals with $A=1$ who died ($Y^{a=1}=1$) were of type "hurt". By randomization there must also be 550 "hurt" individuals with $A=0$ who, of course, survived. Arguing similarly, we can fill in the table for type "helped".

"Hurt"	$A=0$	$A=1$	"Helped"	$A=0$	$A=1$
$Y=1$	0	550	$Y=1$	450	0
$Y=0$	550	0	$Y=0$	0	450
	550	550		450	450

Combining the two tables, we recover the overall trial data, which implies that there are no "immune" individuals.

We conclude that the regulator should relicense treatment A and recommend that the current practice be continued, as the observational study demonstrated that somehow every individual with lung cancer had private knowledge, unavailable to the trialists, as to whether treatment was personally harmful or beneficial. We next describe an extreme scenario that could explain the source of this private knowledge.

Suppose that there was a 55/45 ethnic split in the population and that, for genetic reasons, the treatment was uniformly harmful to individuals with lung cancer in the first group (i.e., all are of type "hurt"), but was uniformly beneficial to all individuals with cancer in the second group (i.e., all are of type "helped"). Also suppose that, at some time before the trial, individuals with cancer in the first ethnic group refrained from taking the treatment after having seen several of the group's members quickly die after taking it. Conversely, suppose all individuals with cancer in the second ethnic group chose to take the treatment after having seen several of the group's members survive after taking it. In other words, suppose that there is both (i) maximal qualitative effect modification by ethnic group and (ii) maximal confounding by ethnic group in the observational data.

The extreme setting described above is of course unrealistic, but it is useful to explain an important point: The randomized trial compared the strategies "treat everybody" and "treat nobody", but the optimal strategy is "treat only individuals in ethnic group 2". When data on the effect modifier (i.e., ethnic group) are not obtained, it is not possible to assign individuals to the optimal strategy in a randomized trial. In contrast, in the observational study, all individuals followed the optimal strategy and thus had the optimal outcome of no deaths. Thus, the confounded observational study, and not the unconfounded randomized trial, revealed the correct policy.

Fine Point 22.8

Generalizing Fine Point 22.7 to realistic settings The above discussion can be generalized to more realistic settings in order to show that a design in which randomized trial and observational data are combined may be more informative than a design with randomized trial data alone, provided individuals in both the randomized and observational data are random samples of all individuals eligible for the trial.

Suppose we conduct a randomized trial for a binary treatment A and an outcome Y in a population in which a treatment is already in use (lower values of Y are preferable). Further suppose that, as may occasionally happen, the mean outcome $\mathrm{E}[Y]$ in the observational study is less than the mean outcome in both arms of the randomized trial, which implies $\mathrm{E}[Y]$ is less than both $\mathrm{E}[Y^{a=0}]$ and $\mathrm{E}\left[Y^{a=1}\right]$. Then, we might choose to leave the current community practice with respect to the treatment unchanged.

We next demonstrate that $\mathrm{E}[Y] = \mathrm{E}[Y^g]$ for a strategy g that differs from the strategies "treat everybody" $a = 1$ and "treat nobody" $a = 0$ compared in the trial. Specifically, let U be a sufficient set of unmeasured, possibly unknown, pre-treatment covariates sufficient to ensure $Y^a \amalg A|U$ and let $\Pr[A = 1|U]$ be the associated propensity score in the observational data. Then the above equality holds for the random strategy g in which treatment $A = 1$ is randomly assigned with probability $\Pr[A = 1|U]$, as this choice gives the random strategy g that generated the observational data on A, Y and the unknown U. Even though the strategy g cannot be implemented because data on U is unavailable, the above discussion could motivate the investigators to measure pre-treatment covariates V, which can be used to analyze the randomized trial data to find and then implement a deterministic dynamic strategy $g^*(x)$ such that $\mathrm{E}[Y^{g^*}]$, as estimated from the unconfounded randomized trial data, is less than the observational $\mathrm{E}[Y]$.