Chapter 21
G-METHODS FOR TIME-VARYING TREATMENTS

In the previous chapter we described a dataset with a time-varying treatment and treatment-confounder feedback. We showed that, when applied to this dataset, traditional methods for confounding adjustment could not correctly adjust for confounding. Even though the time-varying treatment had a zero causal effect on the outcome, traditional adjustment methods yielded effect estimates that were different from the null.

This chapter describes the solution to the bias of traditional methods in the presence of treatment-confounder feedback: the use of g-methods—the g-formula, IP weighting, g-estimation, and their doubly-robust generalizations. Using the same dataset as in the previous chapter, here we show that the three g-methods yield the correct (null) effect estimate. For time-fixed treatments, we described the g-formula in Chapter 13, IP weighting of marginal structural models in Chapter 12, and g-estimation of structural nested models in Chapter 15. Here we introduce each of the three g-methods for the comparison of static treatment strategies under the identifiability conditions described in Chapter 19: sequential exchangeability, positivity, and consistency.

21.1 The g-formula for time-varying treatments

Table 21.1

N	A_0	L_1	A_1	Mean Y
2400	0	0	0	84
1600	0	0	1	84
2400	0	1	0	52
9600	0	1	1	52
4800	1	0	0	76
3200	1	0	1	76
1600	1	1	0	44
6400	1	1	1	44

Consider again the data from the sequentially randomized experiment in Table 20.1 which, for convenience, we reproduce again here as Table 21.1. Suppose we are only interested in the effect of the time-fixed treatment A_1. That is, suppose we want to contrast the mean counterfactual outcomes $\mathrm{E}\left[Y^{a_1=1}\right]$ and $\mathrm{E}\left[Y^{a_1=0}\right]$. In Parts I and II we have showed that, under the identifiability conditions, each of the means $\mathrm{E}\left[Y^{a_1}\right]$ is a weighted average of the mean outcome $\mathrm{E}\left[Y|A_1=a_1,L_1=l_1\right]$ conditional on the (time-fixed) treatment and confounders. Specifically, $\mathrm{E}\left[Y^{a_1}\right]$ equals the weighted average

$$\sum_{l_1} \mathrm{E}\left[Y|A_1=a_1,L_1=l_1\right] f\left(l_1\right), \text{ where } f\left(l_1\right)=\Pr\left[L_1=l_1\right].$$

because, as shown in the previous chapter, only L_1 is needed to make the treated ($A_1=1$) and the untreated ($A_1=0$) conditionally exchangeable. This weighted average is the g-formula for $\mathrm{E}\left[Y^{a_1}\right]$: the mean outcome standardized to the distribution of the confounders (here, L_1 only) in the study population.

But, in the sequentially randomized experiment of Table 21.1, the treatment $\overline{A}=(A_0,A_1)$ is time-varying and, as we saw in the previous chapter, there is treatment-confounder feedback. That means that traditional adjustment methods cannot be relied on to unbiasedly estimate the causal effect of time-varying treatment $\overline{A}$. For example, traditional methods may not provide valid estimates of the mean outcome under "always treat" $\mathrm{E}\left[Y^{a_0=1,a_1=1}\right]$ and the mean outcome under "never treat" $\mathrm{E}\left[Y^{a_0=0,a_1=0}\right]$ even in a sequentially randomized experiment in which sequential exchangeability holds. In contrast, the g-formula can be used to calculate the counterfactual means $\mathrm{E}\left[Y^{a_0,a_1}\right]$ in a sequentially randomized experiment. To do so, the above expression of the g-formula for time-fixed treatments needs to be generalized.

The g-formula for $\mathrm{E}\left[Y^{a_0,a_1}\right]$ under the identifiability conditions (described in Chapter 19) will still be a weighted average, but now it will be a weighted average of the mean outcome $\mathrm{E}\left[Y|A_0=a_0,A_1=a_1,L_1=l_1\right]$ conditional on the time-varying treatment and confounders required to achieve sequential exchangeability. The weights are the distribution of the confounder L_1 given the past which, in this case, is the past value of treatment corresponding to the intervention. Specifically, the g-formula

$$\sum_{l_1}\mathrm{E}\left[Y|A_0=a_0,A_1=a_1,L_1=l_1\right]f\left(l_1|a_0\right)$$

equals $\mathrm{E}\left[Y^{a_0,a_1}\right]$ under (static) sequential exchangeability for Y^{a_0,a_1}. That is, for a time-varying treatment, the g-formula estimator of the counterfactual mean outcome under the identifiability conditions is the mean outcome standardized to the distribution of the confounders in the study population, with every factor in the expression conditional on past treatment and covariate history. This conditioning on prior history is not necessary in the time-fixed case in which both treatment and confounders are measured at a single time point.

The g-formula for time-varying treatments was first described by Robins (1986, 1987).

The g-formula is only computable (i.e., well-defined) if, for any value l_1 such that $f\left(l_1|a_0\right)\neq 0$, there are individuals with $\left(A_0=a_0,A_1=a_1,L_1=l_1\right)$ in the population. This is equivalent to the definition of positivity given in Technical Point 19.2 and a generalization for time-varying treatments of the discussion of positivity in Technical Point 3.1.

Let us apply the g-formula to estimate the causal effect $\mathrm{E}\left[Y^{a_0=1,a_1=1}\right]-\mathrm{E}\left[Y^{a_0=0,a_1=0}\right]$ from the sequentially randomized experiment of Table 21.1. The g-formula estimate for the mean $\mathrm{E}\left[Y^{a_0=0,a_1=0}\right]$ is $84\times 0.25+52\times 0.75=60$. The g-formula estimate for the mean $\mathrm{E}\left[Y^{a_0=1,a_1=1}\right]$ is $76\times 0.50+44\times 0.50=60$. Therefore the estimate of the causal effect $\mathrm{E}\left[Y^{a_0=1,a_1=1}\right]-\mathrm{E}\left[Y^{a_0=0,a_1=0}\right]$ is 0, as expected. The g-formula succeeds where traditional methods failed.

In a study with 2 time points, the g-formula for "never treat" is
$\mathrm{E}\left[Y|A_0=0,A_1=0,L_1=0\right]\times$
$\Pr\left[L_1=0|A_0=0\right]+$
$\mathrm{E}\left[Y|A_0=0,A_1=0,L_1=1\right]\times$
$\Pr\left[L_1=1|A_0=0\right]$

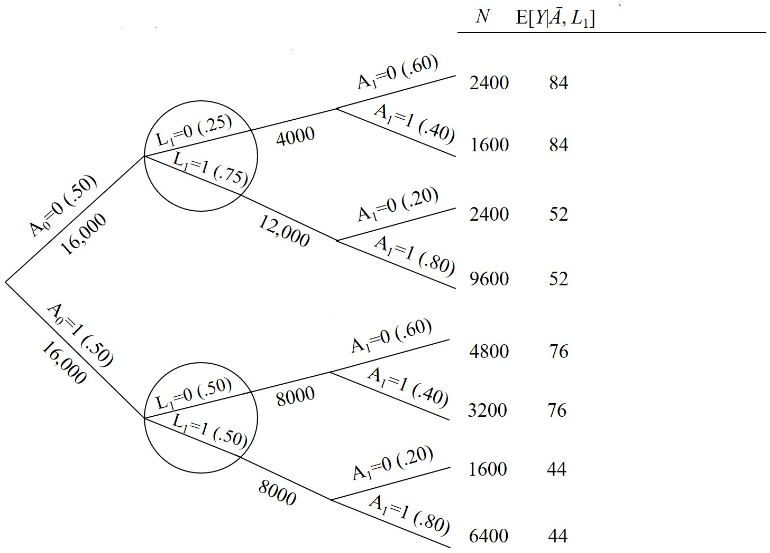


Figure 21.1

Another way to think of the g-formula is as a simulation. Under sequential exchangeability for Y and $\bar{L}$ jointly, the g-formula simulates the counterfactual outcome $Y^{\bar{a}}$ and covariate history $\bar{L}^{\bar{a}}$ that would have been observed if everybody in the study population had followed treatment strategy $\bar{a}$. In other words, the g-formula simulates (identifies) the joint distribution of the counterfactuals $\left(Y^{\bar{a}}, \bar{L}^{\bar{a}}\right)$ under strategy $\bar{a}$. To see this, first consider the causally interpreted structured tree graph in Figure 21.1, which is an alternative representation of the data in Table 21.1. Under the aforementioned identifiability condition, the g-formula can be viewed as a procedure to build a new tree in which all individuals follow strategy $\bar{a}$. For example, the causally interpreted structured tree graph in Figure 21.2 shows the counterfactual population that would have been observed if all individuals have followed the strategy "always treat" $(a_0 = 1, a_1 = 1)$.

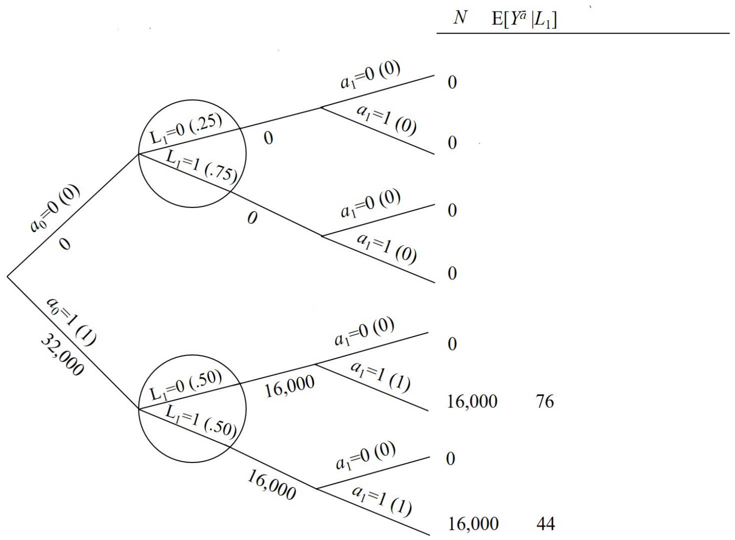


Figure 21.2

Under sequential exchangeability, $\Pr[L_1 = l_1 | A_0 = a_0] = \Pr\left[L_1^{a=0} = l_1\right]$ and $\mathrm{E}[Y | A_0 = a_0, A_1 = a_1, L_1 = l_1] = \mathrm{E}[Y^{a_0, a_1} | L_1^{a_0} = l_1]$.

Thus the g-formula is $\sum_{l_1} \mathrm{E}[Y^{a_0,a_1} | L_1^{a_0} = l_1]$ $\Pr[L_1^{a_0} = l_1]$, which equals $\mathrm{E}[Y^{a_0,a_1}]$ as required.

To simulate this counterfactual population we (i) assign probability 1 to receiving treatment $a_0 = 1$ and $a_1 = 1$ at times $k = 0$ and $k = 1$, respectively, and (ii) assign the same probability $\Pr[L_1 = l_1 | A_0 = a_0]$ and the same mean $\mathrm{E}[Y | A_0 = a_0, A_1 = a_1, L_1 = l_1]$ as in the original study population.

Two important points. First, the value of the g-formula depends on what, if anything, has been included in L. As an example, suppose we do not collect data on L_1 because we believe, incorrectly, that our study is represented by a causal diagram like the one in Figure 20.8 after removing the arrow from L_1 to A_1. Thus we believe L_1 is not a confounder and hence not necessary for identification. Then the g-formula in the absence of data on L_1 becomes $\mathrm{E}[Y | A_0 = a_0, A_1 = a_1]$ because there is no covariate history to adjust for. However, because our study is actually represented by the causal graph in Figure 20.8. (under which treatment assignment A_1 is affected by L_1), the g-formula that fails to include L_1 no longer has a causal interpretation.

Second, even when the g-formula has a causal interpretation, each of its components may lack a causal interpretation. As an example, consider the causal diagram in Figure 20.9 under which only static sequential exchangeability holds. The g-formula that includes L_1 correctly identifies the mean of Y^a. Remarkably, regardless of whether we add arrows from A_0 and A_1 to Y, the g-formula continues to have a causal interpretation as $\mathrm{E}[Y^{\bar{a}}]$, even though neither of its components—$\mathrm{E}[Y | A_0 = a_0, A_1 = a_1, L_1 = l_1]$ and $\Pr[L_1 = l_1 | A_0 = a_0]$—has any causal interpretation at all. That is, $\Pr[L_1 = l_1 | A_0 = a_0] \neq \Pr[L_1^{a_0} = l_1]$ and $\mathrm{E}[Y | A_0 = a_0, A_1 = a_1, L_1 = l_1] \neq \mathrm{E}[Y^{a_0,a_1} | L_1^{a_0} = l_1]$. The last two inequalities will be equalities in a sequential randomized trial like the one represented in Figures 20.1 and 20.2.

Fine Point 21.1

Treatment and covariate history When describing g-methods, we often refer to the treatment and covariate history that is required to achieve sequential exchangeability. For the g-formula, we say that its components are conditional on prior treatment and covariate history. For example, the factor corresponding to the probability of a discrete confounder L_2 at time $k = 2$

$$f\left(l_2|\overline{A}_1 = \bar{a}_1, \overline{L}_1 = \bar{l}_1\right) = \Pr\left[L_2 = l_2|A_0 = a_0, A_1 = a_1, L_0 = l_0, L_1 = l_1\right]$$

is conditional on treatment and confounders at prior times 0 and 1; the factor at time $k = 3$ is conditional on treatment and confounders at times 0, 1, and 2, and so on.

However, the term "history" need not be defined temporally because, as explained in Fine Point 7.4, confounders can theoretically be in the temporal future of a treatment. Conversely, as explained along with Figure 7.4, adjusting for some variables in the temporal past of treatment may introduce selection bias (referred to as M-bias). Therefore, in this book, the causally relevant "history" at time k should be understood as the set of treatments and confounders that are needed to achieve conditional exchangeability for treatment A_k. In most cases this use of history will correspond to the chronological history.

Now let us generalize the g-formula to high-dimensional settings with multiple times k. The g-formula is

$$\sum_{\bar{l}} \mathrm{E}\left[Y|\bar{A} = \bar{a}, \bar{L} = \bar{l}\right] \prod_{k=0}^{K} f\left(l_k|\bar{a}_{k-1}, \bar{l}_{k-1}\right),$$

Technical Point 21.1 shows a more general expression for the g-formula, which can be used to compute densities, not just means.

where the sum is over all possible $\bar{l}$-histories ($\bar{l}_{k-1}$ is the history through time $k-1$). The sum $\sum_{\bar{l}}$ can also be written as $\sum_{l_K} \cdots \sum_{l_1} \sum_{l_0}$. Under sequential exchangeability for $Y^{\bar{a}}$ given $\left(\bar{L}_k, \bar{A}_k\right)$ at each time k, this expression equals the counterfactual mean $\mathrm{E}\left[Y^{\bar{a}}\right]$ under treatment strategy $\bar{a}$. Fine Point 21.1 presents a more nuanced definition of the term "history".

In practice, however, the components of the g-formula cannot be directly computed if the data are high-dimensional, as is expected in observational studies with multiple confounders or time points. The quantities $\mathrm{E}\left[Y|\bar{A} = \bar{a}, \bar{L} = \bar{l}\right]$ and $f\left(l_k|\bar{a}_{k-1}, \bar{l}_{k-1}\right)$ will need to be estimated. For example, we can fit a linear regression model to estimate the conditional means $\mathrm{E}\left[Y|\bar{A} = \bar{a}, \bar{L} = \bar{l}\right]$ of the outcome variable at the end of follow-up, and logistic regression models to estimate the distribution of the discrete confounders L_k at each time $k \neq 0$ (the distribution of L_0 can be estimated without models as described in Section 13.3). The estimates from these models, $\widehat{\mathrm{E}}\left[Y|\bar{A} = \bar{a}, \bar{L} = \bar{l}\right]$ and $\widehat{f}\left(l_k|\bar{a}_{k-1}, \bar{l}_{k-1}\right)$, will then be plugged in into the g-formula. Since Chapter 13, we have referred to this estimator as the *plug-in g-formula* and, when the estimates used in the plug-in g-formula are based on parametric models, we have referred to the plug-in g-formula as the *parametric g-formula*.

For simplicity, this chapter largely focuses on the g-formula under deterministic strategies. However, under sequential exchangeability, the g-formula can be used to compute the counterfactual mean of the outcome under a random treatment strategy f^{int}. An example of a random (static) strategy is "independently at each time k, treat individuals with probability 0.3 and do not treat with probability 0.7", where $f^{int}\left(1|\bar{a}_{k-1}, \bar{l}_k\right) = 0.3$. That is, $f^{int}\left(a_k|\bar{a}_{k-1}, \bar{l}_k\right)$ is the conditional probability of treatment a_k at time k under the treatment

Technical Point 21.1

The g-formula density The g-formula density for $\left(Y,\overline{L}\right)$ evaluated at $\left(y,\bar{l}\right)$ for a deterministic static strategy $\bar{a}$ is

$$f\left(y|\bar{a}_K,\bar{l}_K\right)\prod_{k=0}^{K} f\left(l_k|\bar{a}_{k-1},\bar{l}_{k-1}\right)$$

The static g-formula density for Y is simply the marginal density of Y under the g-formula density for $\left(Y,\overline{L}\right)$:

$$\int \ldots \int f\left(y|\bar{a}_K,\bar{l}_K\right)\prod_{k=0}^{K} dF\left(l_k|\bar{a}_{k-1},\bar{l}_{k-1}\right),$$

where the integral notation $\int$ is used to accommodate settings in which some components of L_k are continuous.

The g formula density for $\left(Y,\overline{L}\right)$ and for Y for a dynamic deterministic strategy $g=(g_0,...,g_K)$, with $g_k\left(\overline{a}_{k-1},\bar{l}_k\right)$ taking values in the support of A_k, simply replaces $\overline{a}_k$ by $\overline{a}_k^g$ in the above formulae. Here, $\overline{a}_k^g$ is recursively defined for $k=0,...,K$, by $\overline{a}_k^g\equiv\overline{g}_k\left(\overline{a}_{k-1}^g,\bar{l}_k\right)\equiv\left[g_0\left(\overline{a}_{-1}^g,\bar{l}_0\right),...,g_k\left(\overline{a}_{k-1}^g,\bar{l}_k\right)\right]$ with $\overline{a}_{-1}^g$ defined to be 0. A static strategy is the special case of a dynamic strategy when each $g_k\left(\overline{a}_{k-1},\bar{l}_k\right)$ is a constant function.

In more generality, given observed data $O=\left(\bar{A},\overline{X},Y\right)$ and unobserved data $\overline{U}$, where $\overline{X}$ is the set of all measured variables other than treatment $\bar{A}$ and outcome Y, the inputs of the g-formula are (i) a deterministic treatment strategy g, (ii) a causal DAG representing the observed data (and their unmeasured common causes), (iii) a subset $\overline{L}$ of $\overline{X}$ for which we wish to adjust, and (iv) a choice of a total ordering of $\overline{L}$, $\bar{A}$, and Y consistent with the topology of the DAG, i.e., an ordering such that each variable comes after its ancestors. The vector L_k consists of all variables in L after A_{k-1} and before A_k in the ordering. The chosen ordering will usually, but not always, be temporal as discussed in Fine Point 21.1. When sequential exchangeability for Y^g and positivity holds for the chosen ordering, the g-formula density for Y equals the density $f_{Y^g}\left(y\right)$ that would have been observed in the study population if all individuals had followed strategy g. Otherwise, the g-formula can still be computed, but it lacks a causal interpretation. When positivity and exchangeability for $\left(Y^g,\overline{L}^g\right)$ hold (e.g., no arrow from any variable either in $\overline{U}$ or in $\overline{X}$ but not in $\overline{L}$ directly into any treatment variable), the g-formula density for $\left(Y,\overline{L}\right)$ equals the density $f_{Y^g,\overline{L}^g}\left(y,\bar{l}\right)$.

strategy (or *int*ervention) f^{int}. Then, the general g-formula expression is

$$\sum_{\bar{a},\bar{l}}\mathrm{E}\left[Y|\bar{A}=\bar{a},\bar{L}=\bar{l}\right]\prod_{k=0}^{K} f\left(l_k|\bar{a}_{k-1},\bar{l}_{k-1}\right)\prod_{k=0}^{K} f^{int}\left(a_k|\bar{a}_{k-1},\bar{l}_k\right).$$

Note this is the formula for the observed mean of Y if we replace $f^{int}\left(a_k|\bar{a}_{k-1},\bar{l}_k\right)$ by the observed conditional probability of treatment $f\left(a_k|\bar{a}_{k-1},\bar{l}_k\right)$.

CODE: The *gfoRmula* R package (Lin et al. 2019) is available through CRAN. The *GFORMULA* SAS macro is available through GitHub. See the book's web site.

This expression of the g-formula is general enough to accommodate both deterministic and random strategies. Under a deterministic treatment strategy, $f^{int}\left(a_k|\bar{a}_{k-1},\bar{l}_k\right)$ is always 1 for the values of a_k mandated by the strategy and 0 for the others. For example, under the strategy "never treat" or $\bar{a}=(0,0,...0)$, the probability $f^{int}\left(0|\bar{a}_{k-1},\bar{l}_k\right)=1$ at all k. Since $f^{int}\left(a_k|\bar{a}_{k-1},\bar{l}_k\right)$ equals 1 for the mandated values of treatment and 0 for all other values of treatment, it is not necessary to include the f^{int} factors, or the sum over $\bar{a}$, in the above formula. Our publicly available software implements this general expression of the g-formula and therefore can accommodate any treatment strategy.

21.2 IP weighting for time-varying treatments

Suppose we are only interested in the effect of the time-fixed treatment A_1 in Table 21.1. We then want to contrast the counterfactual mean outcomes $\mathrm{E}\left[Y^{a_1=1}\right]$ and $\mathrm{E}\left[Y^{a_1=0}\right]$. As we have seen in Chapter 12, under the identifiability conditions, each of the counterfactual means $\mathrm{E}\left[Y^{a_1}\right]$ is the mean $\mathrm{E}_{ps}\left[Y|A_1=a_1\right]$ in the pseudo-population created by the subject-specific nonstabilized weights $W^{A_1}=1/f\left(A_1|L_1\right)$ or the stabilized weights $SW^{A_1}=f\left(A_1\right)/f\left(A_1|L_1\right)$. The denominator of the IP weights is, informally, an individual's probability of receiving the treatment value that he or she received, conditional on the individual's confounder values. One can estimate $\mathrm{E}_{ps}\left[Y|A_1=a_1\right]$ from the observed study data by the average of Y among subjects with $A_1=a_1$ in the pseudo-population.

When treatment and confounders are time-varying, these IP weights for time-fixed treatments need to be generalized. For a time-varying treatment $\overline{A}=(A_0,A_1)$ and time-varying covariates $\bar{L}=(L_0,L_1)$ at two time points, the nonstabilized IP weights are

$$W^{\bar{A}}=\frac{1}{f\left(A_0|L_0\right)}\times\frac{1}{f\left(A_1|A_0,L_0,L_1\right)}=\prod_{k=0}^{1}\frac{1}{f\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)}$$

and the stabilized IP weights are

$$SW^{\bar{A}}=\frac{f\left(A_0\right)}{f\left(A_0|L_0\right)}\times\frac{f\left(A_1|A_0\right)}{f\left(A_1|A_0,L_0,L_1\right)}=\prod_{k=0}^{1}\frac{f\left(A_k|\bar{A}_{k-1}\right)}{f\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)}$$

where A_{-1} is 0 by definition. The denominator of the IP weights for a time-varying treatment is, informally, an individual's probability of receiving the treatment history that he or she received, conditional on the individual's treatment and covariate history.

Suppose we want to contrast the counterfactual means $\mathrm{E}\left[Y^{a_0=1,a_1=1}\right]$ and $\mathrm{E}\left[Y^{a_0=0,a_1=0}\right]$. Under the identifiability assumptions for static strategies, each counterfactual mean $\mathrm{E}\left[Y^{a_0,a_1}\right]$ is the mean $\mathrm{E}_{ps}\left[Y|A_0=a_0,A_1=a_1\right]$ in the pseudo-population created by the nonstabilized weights $W^{\bar{A}}$ or the stabilized weights $SW^{\bar{A}}$. That is, the IP weighted estimator of each counterfactual mean is the average of Y among individuals with $\overline{A}=(A_0,A_1)$ in the pseudo-population.

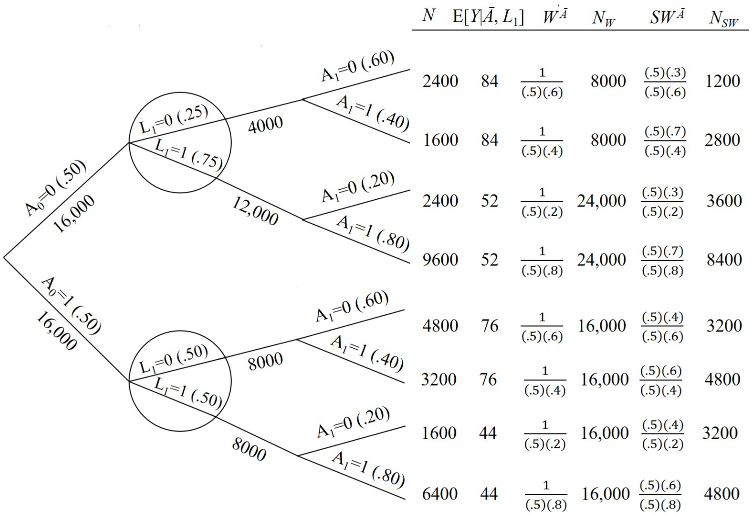
Similar to the result for time-fixed treatment in Technical Point 12.2, $\mathrm{E}_{ps}\left[Y|A_0=a_0,A_1=a_1\right]$ equals $\mathrm{E}\left[W^{\bar{A}}Y\,\mathrm{I}\left(A_0=a_0,A_1=a_1\right)\right]=\frac{\mathrm{E}\left[SW^{\bar{A}}Y\,\mathrm{I}(A_0=a_0,A_1=a_1)\right]}{\mathrm{E}\left[SW^{\bar{A}}\,\mathrm{I}(A_0=a_0,A_1=a_1)\right]}$, for both the nonstabilized and stabilized pseudo-populations, regardless of whether sequential exchangeability holds.

Let us apply IP weighting to the data from Table 21.1. The causally interpreted structured tree graph in Figure 21.3 is the tree graph in Figure 21.1 with additional columns for the nonstabilized IP weights $W^{\bar{A}}$ and the number of individuals in the corresponding pseudo-population N_W for each treatment and covariate history. The pseudo-population has a size of $128,000$, i.e., the $32,000$ individuals in the original population multiplied by 4, the number of static strategies. Because there is no L_0 in this study, the denominator of the IP weights simplifies to $f\left(A_0\right)f\left(A_1|A_0,L_1\right)$.

The IP weighted estimator for the counterfactual mean $\mathrm{E}\left[Y^{a_0=0,a_1=0}\right]$ is the mean $\mathrm{E}_{ps}\left[Y|A_0=0,A_1=0\right]$ in the pseudo-population, which we estimate as the average outcome among the $32,000$ individuals with $(A_0=0,A_1=0)$ in the pseudo-population. From the tree in Figure 21.3, the estimate is $84\times\frac{8000}{32000}+52\times\frac{24000}{32000}=60$. Similarly, the IP weighted estimate of $\mathrm{E}\left[Y^{a_0=1,a_1=1}\right]$ is also 60. Therefore the estimate of the causal effect $\mathrm{E}\left[Y^{a_0=1,a_1=1}\right]-\mathrm{E}\left[Y^{a_0=0,a_1=0}\right]$ is 0, as expected. IP weighting, like the g-formula, succeeds where traditional methods failed.

Note that our nonparametric estimates of $\mathrm{E}\left[Y^{a_0,a_1}\right]$ based on the g-formula are precisely equal to those based on IP weighting. This equality has nothing to do with causal inference. That is, even if the identifiability conditions did not hold—so neither the g-formula nor IP weighting estimates have a causal interpretation—both approaches would yield the same number.

The same estimate of 0 is obtained when using stabilized weights $SW^{\bar{A}}$ in Figure 21.3 (check for yourself). However, $\Pr_{ps}\left[A_k=1|\bar{A}_{k-1},\bar{L}_k\right]$ is $1/2$ in the nonstabilized pseudo-population and $\Pr_{ps}\left[A_k=1|\bar{A}_{k-1}\right]$ in the stabilized pseudo-population.



N	$\mathrm{E}[Y\vert\bar{A},L_1]$	$W^{\bar{A}}$	N_W	$SW^{\bar{A}}$	N_{SW}
2400	84	$\frac{1}{(.5)(.6)}$	8000	$\frac{(.5)(.3)}{(.5)(.6)}$	1200
1600	84	$\frac{1}{(.5)(.4)}$	8000	$\frac{(.5)(.7)}{(.5)(.4)}$	2800
2400	52	$\frac{1}{(.5)(.2)}$	24,000	$\frac{(.5)(.3)}{(.5)(.2)}$	3600
9600	52	$\frac{1}{(.5)(.8)}$	24,000	$\frac{(.5)(.7)}{(.5)(.8)}$	8400
4800	76	$\frac{1}{(.5)(.6)}$	16,000	$\frac{(.5)(.4)}{(.5)(.6)}$	3200
3200	76	$\frac{1}{(.5)(.4)}$	16,000	$\frac{(.5)(.6)}{(.5)(.4)}$	4800
1600	44	$\frac{1}{(.5)(.2)}$	16,000	$\frac{(.5)(.4)}{(.5)(.2)}$	3200
6400	44	$\frac{1}{(.5)(.8)}$	16,000	$\frac{(.5)(.6)}{(.5)(.8)}$	4800

Figure 21.3

Let us generalize IP weighting to high-dimensional settings with multiple times $k=0,1...K$. The general form of the nonstabilized IP weights is

$$W^{\bar{A}}=\prod_{k=0}^{K}\frac{1}{f\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)}$$

and the general form of the stabilized IP weights is

$$SW^{\bar{A}}=\prod_{k=0}^{K}\frac{f\left(A_k|\bar{A}_{k-1}\right)}{f\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)}$$

When the identifiability conditions hold, these IP weights create a pseudo-population in which (i) the mean of $Y^{\bar{a}}$ is identical to that in the actual population, but (ii) like on Figure 19.1, the randomization probabilities at each time k are the constant 1/2 (nonstabilized weights) or depend at most on past treatment history (stabilized weights). Hence the average causal effect $\mathrm{E}\left[Y^{\bar{a}}\right]-\mathrm{E}\left[Y^{\bar{a}'}\right]$ is $\mathrm{E}_{ps}\left[Y|\bar{A}=\bar{a}\right]-\mathrm{E}_{ps}\left[Y|\bar{A}=\bar{a}'\right]$ because sequential unconditional exchangeability holds in both pseudo-populations.

Our description in the text considers only static strategies. For a description of IP weighting with dynamic strategies, see Technical Point 21.2.

In a true sequentially randomized trial, the quantities $f\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)$ are known by design. Therefore we can use them to compute nonstabilized IP weights and the estimates of $\mathrm{E}\left[Y^{\bar{a}}\right]$ and $\mathrm{E}\left[Y^{\bar{a}}\right]-\mathrm{E}\left[Y^{\bar{a}'}\right]$ are guaranteed to be unbiased. In contrast, in observational studies, the quantities $f\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)$ will need to be estimated from the data. When the data are high-dimensional, we can, for example, fit a logistic regression model to estimate the conditional probability of a dichotomous treatment $\Pr\left[A_k=1|\bar{A}_{k-1},\bar{L}_k\right]$ at each time k. The estimates $\widehat{f}\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)$ from these models will then replace $f\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)$ in $W^{\bar{A}}$. If the estimates $\widehat{f}\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)$ are based on a misspecified logistic model for the $\Pr\left[A_k=1|\bar{A}_{k-1},\bar{L}_k\right]$, the resulting estimates of

Technical Point 21.2

IP weighting for dynamic treatment strategies. Consider the deterministic dynamic strategy $g = (g_0, ..., g_K)$ with $g_k \equiv g_k\left(\overline{a}_{k-1}, \bar{l}_k\right)$. The g-formula for an outcome Y under g equals $\mathrm{E}\left[Y\,\mathrm{I}(\bar{A}_K = \overline{A}^g_K)W^{\bar{A}}\right]$, where $\overline{a}^g_K$ was defined in Technical Point 21.1. Further, $\mathrm{E}\left[Y\,\mathrm{I}(\bar{A}_K = \overline{A}^g_K)W^{\bar{A}}\right] = \mathrm{E}_{ps}[Y|\bar{A}_K = \overline{A}^g_K]$ where $\mathrm{E}_{ps}[Y|\bar{A}_K = \overline{A}^g_K]$ is the mean of Y among the members of the pseudo-population who follow strategy g. Unlike for static strategies, $\mathrm{E}\left[Y\,\mathrm{I}(\bar{A}_K = \overline{A}^g_K)SW^{\bar{A}}\right] / \mathrm{E}\left[\mathrm{I}(\bar{A}_K = \overline{A}^g_K)SW^{\bar{A}}\right]$ does not equal the g-formula because the numerator of $SW^{\bar{A}}$ depends on A. Hence stabilized weights cannot be used with dynamic strategies. For a random dynamic strategy f^{int}, the g-formula is equal to $\mathrm{E}\left[Y\prod\limits_{k=0}^{K} f^{int}\left(A_k|\bar{A}_{k-1}, \bar{L}_k\right)W^{\bar{A}}\right] = \mathrm{E}\left[Y\prod\limits_{k=0}^{K}\frac{f^{int}\left(A_k|\bar{A}_{k-1}, \bar{L}_k\right)}{f\left(A_k|\bar{A}_{k-1}, \bar{L}_k\right)}\right]$.

$\mathrm{E}\left[Y^{\bar{a}}\right]$ and $\mathrm{E}\left[Y^{\bar{a}}\right] - \mathrm{E}\left[Y^{\bar{a}'}\right]$ will be biased. For stabilized weights $SW^{\bar{A}}$ we must also obtain an estimate of $\widehat{f}\left(A_k|\bar{A}_{k-1}\right)$ for the numerator. Even if this estimate is based on a misspecified model, the estimates of $\mathrm{E}\left[Y^{\bar{a}}\right]$ and $\mathrm{E}\left[Y^{\bar{a}}\right] - \mathrm{E}\left[Y^{\bar{a}'}\right]$ remain unbiased, although $\widehat{f}\left(a_k|\overline{a}_{k-1}\right)$ in the stabilized pseudo-population will no longer be consistent for the observed data density $f\left(a_k|\overline{a}_{k-1}\right)$.

In practice, a common approach is to fit a single model for $\Pr\left[A_k = 1|\bar{A}_{k-1}, \bar{L}_k\right]$ rather than a separate model at each time k. The model includes functions of time k—a time-varying intercept—as covariates, and possibly product terms with other covariates.

Suppose that we obtain two estimates of $\mathrm{E}\left[Y^{\bar{a}}\right]$, one using the parametric g-formula and another one using IP weights estimated via parametric models, and that the two estimates differ by more than can be reasonably explained by sampling variability (the sampling variability of the difference of the estimates can be quantified by bootstrapping). We can then conclude that the parametric models used for the g-formula or the parametric models used for IP weighting (or both) are misspecified. This conclusion is always true, regardless of whether the identifiability assumptions hold. An implication is that one should always estimate $\mathrm{E}\left[Y^{\bar{a}}\right]$ using both methods and, if the estimates differ substantially (according to some prespecified criterion), reexamine all the models and modify them where necessary. In the next section, we describe how doubly-robust estimators can help deal with model misspecification.

There is no logical guarantee of no model misspecification even when the estimates from both parametric approaches are similar, as they may both be biased in the same direction.

Also, as we discussed in the previous section, it is not infrequent that the number of unknown quantities $\mathrm{E}\left[Y^{\bar{a}}\right]$ far exceeds the sample size. Thus we need to specify a model that combines information from many strategies to help estimate a given $\mathrm{E}\left[Y^{\bar{a}}\right]$. For example, we can hypothesize that the effect of treatment history $\bar{a}$ on the mean outcome increases linearly as a function of the cumulative treatment $cum\left(\bar{a}\right) = \sum\limits_{k=0}^{K} a_k$ under strategy $\bar{a}$. This hypothesis is encoded in the *marginal structural mean model*

This marginal structural model is unsaturated. Remember, saturated models have an equal number of unknowns on both sides of the equation.

$$\mathrm{E}\left[Y^{\bar{a}}\right] = \beta_0 + \beta_1 cum\left(\bar{a}\right)$$

for all $\bar{a}$, which is a more general version of the marginal structural mean model for time-fixed treatments discussed in Chapter 12. There are 2^K different unknown quantities on the left hand side of model, one for each of the 2^K different strategies $\bar{a}$, but only 2 unknown parameters β_0 and β_1 on the right hand side. The parameter β_1 measures the average causal effect of the time-varying treatment $\bar{A}$. The average causal effect $\mathrm{E}\left[Y^{\bar{a}}\right] - \mathrm{E}\left[Y^{\bar{a}=\bar{0}}\right]$ is equal to $\beta_1 \times cum\left(\bar{a}\right)$.

As discussed in Chapter 12, to estimate the parameters of the marginal structural model, we can fit the linear regression model

$$\mathrm{E}\left[Y|\overline{A}\right] = \theta_0 + \theta_1 cum\left(\overline{A}\right)$$

by ordinary least squares in either the stabilized or nonstabilized pseudo-population. This is mathematically equivalent to fitting the same linear model by weighted least squares in the original study population, with weights $SW^{\bar{A}}$ or $W^{\bar{A}}$, respectively (in an actual data analysis, these weights are replaced by their estimates). Under the identifiability conditions, the weighted least squares estimate of θ_1 is consistent for the causal parameter β_1 rather than for the associational parameter θ_1.

In statistics courses, it is often proven that, under a correctly specified model for $\mathrm{E}\left[Y|\overline{A}\right]$, both ordinary and weighted least squares estimates are consistent for the associational parameter θ_1. This proof assumes that the weights only depend on $\overline{A}$. When, as in our case, the weights depend on covariates $\overline{L}$ that are correlated with Y given $\overline{A}$, the weighted regression is no longer consistent for θ_1.

As also discussed in Chapter 12, the variance of $\widehat{\beta}_1$—and thus of the contrast $\mathrm{E}\left[Y^{\bar{a}}\right] - \mathrm{E}\left[Y^{\bar{a}=\bar{0}}\right]$—can be estimated by the nonparametric bootstrap or by computing its analytic variance (which requires additional statistical analysis and programming). We can also construct a conservative 95% confidence interval by using the *robust variance estimator* of $\widehat{\beta}_1$, which is directly outputted by most statistical software packages. For a non-saturated marginal structural model the width of the intervals will typically be narrower when the model is fit with the weights $SW^{\bar{A}}$ than with the weights $W^{\bar{A}}$, so the $SW^{\bar{A}}$ weights are preferred.

Of course, the estimates of $\mathrm{E}\left[Y^{\bar{a}}\right]$ will be incorrect if the marginal structural mean model is misspecified, that is, if the mean counterfactual outcome depends on the treatment strategy through a function of the time-varying treatment other than cumulative treatment $cum\left(\bar{a}\right)$ (say, cumulative treatment only in the final 5 months $\sum_{k=K-5}^{K} a_k$) or depends nonlinearly (say, quadratically) on cumulative treatment. However, if we fit the model

$$\mathrm{E}\left[Y|\overline{A}\right] = \theta_0 + \theta_1 cum\left(\overline{A}\right) + \theta_2 cum_{-5}\left(\overline{A}\right) + \theta_3 cum\left(\overline{A}\right)^2$$

with weights $SW^{\bar{A}}$ or $W^{\bar{A}}$, a Wald test on two degrees of freedom of the joint hypothesis $\theta_2 = \theta_3 = 0$ is a test of the null hypothesis that our marginal structural model is correctly specified. That is, IP weighting of marginal structural models is not subject to the g-null paradox described in Technical Point 21.3. In practice, one might choose to use a marginal structural model that includes different summaries of treatment history $\overline{A}$ as covariates, and that uses flexible functions like, say, cubic splines.

This test will generally have good statistical power against the particular directions of misspecification mentioned above, especially when using the weights $SW^{\bar{A}}$ and the bootstrap to estimate the variance.

Finally, as we discussed in Section 12.5, we can use a marginal structural model to explore effect modification by a subset V of the covariates in L_0. For example, for a dichotomous baseline variable V, we would elaborate our marginal structural mean model as

$$\mathrm{E}\left[Y^{\bar{a}}|V\right] = \beta_0 + \beta_1 cum\left(\bar{a}\right) + \beta_2 V + \beta_3 cum\left(\bar{a}\right) V$$

The parameters of this model can be estimated by fitting the ordinary linear regression model $\mathrm{E}\left[Y|\overline{A}, V\right] = \theta_0 + \theta_1 cum\left(\overline{A}\right) + \theta_2 V + \theta_3 V cum\left(\overline{A}\right)$ by weighted least squares with IP weights $W^{\bar{A}}$ or, better, $SW^{\bar{A}}\left(V\right) = \prod_{k=0}^{K} \frac{f\left(A_k|\bar{A}_{k-1}, V\right)}{f\left(A_k|\bar{A}_{k-1}, \bar{L}_k\right)}$. In the presence of treatment-confounder feedback, V can only include baseline variables. If V had components of L_k for $k > 0$ then the parameters $\theta_1$and θ_3 could be different from 0 even if treatment had no effect on the mean outcome at any time.

We now describe a doubly robust estimator of the counterfactual mean $\mathrm{E}\left[Y^g\right]$ for any strategy g.

Technical Point 21.3

The g-null paradox. When using the parametric g-formula, model misspecification will result in biased estimates of $\mathrm{E}\left[Y^{\bar{a}}\right]$, even if the identifiability conditions hold. Suppose there is treatment-confounder feedback and the sharp null hypothesis of no effect of treatment on Y is true, i.e.,

$$Y^{\bar{a}} - Y^{\bar{a}'} = 0 \text{ with probability } 1 \text{ for all } \bar{a}' \text{ and } \bar{a}.$$

Then the value of the g-formula for $\mathrm{E}\left[Y^{\bar{a}}\right]$ is the same for any strategy $\bar{a}$, even though $\mathrm{E}\left[Y|\bar{A}=\bar{a},\bar{L}=\bar{l}\right]$ and $f\left(l_k|\bar{a}_{k-1},\bar{l}_{k-1}\right)$ will both depend on $\bar{a}$ as discussed in Chapter 20. Now suppose we use standard non-saturated parametric models $\mathrm{E}\left[Y|\bar{A}=\bar{a},\bar{L}=\bar{l};\theta\right]$ and $f\left(l_k|\bar{a}_{k-1},\bar{l}_{k-1};\varphi\right)$ based on distinct (i.e., variation-independent) parameters θ and φ to estimate the components of the g-formula. Then Robins and Wasserman (1997) showed that, when L_k has any discrete components, these models cannot all be correctly specified because the estimated value of the g-formula for $\mathrm{E}\left[Y^{\bar{a}}\right]$ will generally depend on $\bar{a}$. As a consequence, inference based on the estimated g-formula might theoretically result in the sharp null hypothesis being falsely rejected, even in a sequentially randomized experiment. This phenomenon is referred to as the null paradox of the estimated g-formula for time-varying treatments. For additional discussion, see Cox and Wermuth (1999) and McGrath et al. (2022). Fortunately, the g-null paradox has not prevented null parametric g-formula effect estimates in practice, presumably because the bias induced by the paradox is small compared with typical random variability.

In contrast, as described in Chapters 12 and 14, neither IP weighting of marginal structural mean models nor g-estimation of structural nested mean models suffer from the null paradox. These models are correctly specified under the sharp null no matter what functional form we choose for treatment. For example, the marginal structural mean model $\mathrm{E}\left[Y^{\bar{a}}\right] = \beta_0 + \beta_1 cum\left(\bar{a}\right)$ is correctly specified under the null because, in that case, $\beta_1 = 0$ and $\mathrm{E}\left[Y^{\bar{a}}\right]$ would not depend on the function of $\bar{a}$. Also, as defined in Section 21.4, any structural nested mean model $\gamma_k\left(\bar{a}_{k-1},\bar{l}_k,\beta\right)$ is correctly specified under the sharp null with $\beta = 0$ being the true parameter value and $\gamma_k\left(\bar{a}_{k-1},\bar{l}_k,\beta\right) = 0$, regardless of the function of past treatment and covariate history.

21.3 A doubly robust estimator for time-varying treatments

Part II briefly mentioned doubly robust methods that combine IP weighting and the g-formula. As we know, IP weighting requires a correct model for treatment A conditional on the confounders L, and the g-formula requires a correct model for the outcome Y conditional on treatment A and the confounders L. Doubly robust methods require a correct model for *either* treatment A *or* outcome Y. If at least one of the two models is correct (and one need not know which of the two models is correct), a doubly robust estimator consistently estimates the causal effect. Fine Point 13.2 described a doubly robust plug-in estimator for the average causal effect of a time-fixed treatment A on an outcome Y. In this section, we first review a slightly different doubly robust plug-in estimator for time-fixed treatments and then extend it to time-varying treatments.

Doubly robust estimators give us two chances to get it right when, as in most observational studies, there are many confounders and non-saturated models are required.

Suppose we want to construct a doubly robust estimator of the average causal effect $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$ of a time-fixed binary treatment A on a binary outcome Y under exchangeability, positivity, and consistency in a setting with many confounders L. We will construct doubly robust estimators of $\mathrm{E}\left[Y^{a}\right]$ as previously discussed in Technical Points 13.2 and 13.3. The difference between doubly robust estimators for $\mathrm{E}\left[Y^{a=1}\right]$ and for $\mathrm{E}\left[Y^{a=0}\right]$ is a doubly robust estimator of the average causal effect. Our doubly robust procedure for $\mathrm{E}\left[Y^{a}\right]$ will use estimates of an outcome model for $\mathrm{E}[Y|A=a,L=l]$ and a model for $\Pr[A=1|L]$ and then combine them appropriately. Our procedure has three steps.

The first step is to compute the predicted values $\widehat{f}(a|L) \equiv \widehat{\Pr}[A=a|L]$ from the treatment model. The second step is to compute the predicted values $\widehat{\mathrm{E}}[Y|A=a,L] = b\left(a,L;\widehat{\theta}\right)$ from the maximum likelihood fit *restricted to individuals with* $A=a$ of the linear logistic model $b(a,L;\theta)$ that includes $\hat{W}^a = 1/\widehat{f}(a|L)$ as a covariate, such as $b(a,L;\theta) = \mathrm{expit}\left(\theta_{a,0} + \theta_{a,1}L + \theta_{a,2}\hat{W}^a\right)$. The third step is to estimate $\mathrm{E}\left[Y^{a=1}\right]$ and $\mathrm{E}\left[Y^{a=0}\right]$ as the standardized means $\widehat{\mathrm{E}}\left[b\left(1,L;\widehat{\theta}\right)\right]$ and $\widehat{\mathrm{E}}\left[b\left(0,L;\widehat{\theta}\right)\right]$, where $\widehat{\mathrm{E}}$ denotes the sample average over all individuals, both treated and untreated. The difference $\widehat{\mathrm{E}}\left[b\left(1,L;\widehat{\theta}\right)\right] - \widehat{\mathrm{E}}\left[b\left(0,L;\widehat{\theta}\right)\right]$ is a doubly robust estimator of the causal effect $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$. That is, under the identifiability conditions, this estimator consistently estimates the average causal effect if either the model for the treatment is correct or the models for the outcome are correct. It is important to realize that treated and untreated individuals with the same value of L also have the same value of $b\left(1,L;\widehat{\theta}\right) = \mathrm{expit}\left(\widehat{\theta}_{1,0} + \widehat{\theta}_{1,1}L + \widehat{\theta}_{1,2}/\widehat{f}(a=1|L)\right)$. They also have the same value of $b\left(0,L;\widehat{\theta}\right) = \mathrm{expit}\left(\widehat{\theta}_{0,0} + \widehat{\theta}_{0,1}L + \widehat{\theta}_{0,2}/\widehat{f}(a=0|L)\right)$.

Let us now extend this doubly robust estimator to settings with time-varying treatments in which we are interested in comparing the counterfactual means $\mathrm{E}\left[Y^{\bar{a}}\right]$ and $\mathrm{E}\left[Y^{\bar{a}'}\right]$ under two treatment strategies $\bar{a}$ and $\bar{a}'$. The doubly robust procedure to estimate $\mathrm{E}\left[Y^{\bar{a}}\right]$ for a time-varying treatment follows the same 3 steps as the procedure to estimate $\mathrm{E}\left[Y^{a}\right]$ for a time-fixed treatment. However, as we will see, the second step is a bit more involved because it requires the fitting of sequential regression models. To simplify notation, we show how to obtain a doubly robust estimator of $\mathrm{E}\left[Y^{\bar{a}}\right]$ under the treatment strategy "always treated", i.e., $\bar{a} = \bar{1}$ where $\bar{1} = \bar{1}_K$ is the vector of $K+1$ 1's.

This doubly robust estimator is due to Bang and Robins (2005) and is closely related to an earlier estimator (Robins, 2000). The estimator is a *targeted minimum loss-based estimator* (TMLE), also known as a targeted maximum likelihood estimator, in the nomenclature later introduced by van der Laan and Rubin (2006) and van der Laan and Gruber (2012).

The first step requires fitting a regression model $\pi_k\left(\bar{L}_k;\alpha\right)$ for $\pi_k\left(\bar{L}_k\right) = \Pr\left[A_k = 1|\bar{A}_{k-1} = \bar{1}_{k-1}, \bar{L}_k\right]$ pooled over all persons and times k. An individual contributes to the fit of the model at time k only if the individual has been treated (continuously) through $k-1$, i.e., $\bar{A}_{k-1} = \bar{1}_{k-1}$. We then use predicted values $\pi_k\left(\bar{L}_k;\widehat{\alpha}\right)$ from this model to estimate for those individuals treated through m ($\bar{A}_m = \bar{1}_m$), the time-varying IP weights $W^{\bar{A}_m} = \prod_{k=0}^{m}\dfrac{1}{f\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)}$ which equals $W^{\bar{1}_m} = \prod_{k=0}^{m}\frac{1}{\pi_k\left(\bar{L}_k\right)}$. That is, for an always-treated individual with $\bar{A}_K = \bar{1}_K$, we assign a different weight $W^{\bar{1}_m}$ for each time point m rather than just the single weight $W^{\bar{1}_K}$ at the end of follow-up as we did in the previous section. For example, if we fit the parametric model $\pi_k\left(\bar{L}_k;\alpha\right) = \mathrm{expit}\left(\alpha_{0,k} + \alpha_2 L_k\right)$ for $\Pr\left[A_k = 1|\bar{A}_{k-1} = 1, \bar{L}_k\right]$, then, in our example of Table 21.1 with two time points ($K=1$), the predicted values $\widehat{\Pr}\left[A_1 = 1|A_0 = 1, \bar{L}_1\right]$ and $\widehat{\Pr}\left[A_0 = 1|L_0\right]$ are $\widehat{\pi}_1 = \mathrm{expit}\left(\hat{\alpha}_{0,1} + \hat{\alpha}_2 L_1\right)$ and $\widehat{\pi}_0 = \mathrm{expit}\left(\hat{\alpha}_{0,0} + \hat{\alpha}_2 L_0\right)$ (because $A_{-1} \equiv 0$). Here, we used the abbreviation $\widehat{\pi}_k$ for $\pi_k\left(\bar{L}_k;\widehat{\alpha}\right)$. We then compute the time-varying IP weight estimates $\hat{W}^{\bar{1}_m} = \prod_{k=0}^{m}\frac{1}{\widehat{\pi}_k}$ for individuals treated through m. We have reached the end of Step 1.

The second step requires fitting a separate outcome model $b_m\left(\bar{L}_m;\beta_m\right)$ at each time m, starting from the last time K and ending at $m=0$. The time m regression model is only fit to individuals treated through m and includes $\hat{W}^{\bar{1}_m} = \hat{W}^{\bar{A}_m}$ as a covariate. The time K model has dependent variable Y. The time m model for $m < K$ has as dependent variable the predicted outcomes from the fit of the time $m+1$ model, i.e., $\widehat{B}_{m+1} = \widehat{b}_{m+1}\left(\bar{L}_{m+1};\beta_{m+1}\right)$. When

Technical Point 21.4

A K+2 robust augmented IP weighted estimator. We consider the case $K = 1$ as the argument generalizes to arbitrary K. The ICE plug-in estimator of the g-formula ψ is $\widehat{\psi}_{gfor} = P_n[\widehat{b}_0(L_0)]$, where P_n denotes a sample average, $\widehat{b}_0(L_0) = \widehat{\mathrm{E}}[\widehat{b}_1(L_0, L_1)|A_0 = 1, L_0]$, and $\widehat{b}_1(L_0, L_1) = \widehat{\mathrm{E}}[Y|L_0, A_0 = 1, L_1, A_1 = 1]$. The IP weighted estimator $\widehat{\psi}_{IPW}$ of ψ is $P_n[A_0A_1Y/(\widehat{\pi}_0\widehat{\pi}_1))]$ where $\widehat{\pi}_0$ and $\widehat{\pi}_1$ are estimates of $\pi_0 = \Pr(A_0 = 1|\, L_0)$ and $\pi_1 = \Pr(A_1 = 1\,|\, L_0, L_1, A_0 = 1)$. Robins et al. (1994) derived an augmented IP weighted estimator $\widehat{\psi}_{TR} = P_n[\widehat{U}_{TR}]$ of ψ where

$$\widehat{U}_{TR} = A_0A_1Y/(\widehat{\pi}_0\widehat{\pi}_1) - \frac{A_0}{\widehat{\pi}_0}\{\frac{A_1}{\widehat{\pi}_1} - 1\}\widehat{b}_1(L_0, L_1) - \{\frac{A_0}{\widehat{\pi}_0} - 1\}\widehat{b}_0(L_0)$$

We now show that $\widehat{\psi}_{TR}$ is triply (i.e., $K + 2$) robust. First, $\widehat{\psi}_{TR}$ is consistent (singly robust) for ψ if $\widehat{\pi}_0$ and $\widehat{\pi}_1$ are consistent since the sample averages of the last 2 terms of $\widehat{U}_{TR}$ are then consistent for 0 and the sample average of the first term is precisely $\widehat{\psi}_{IPW}$. Second, $\widehat{\psi}_{TR}$ is doubly robust because $\widehat{\psi}_{TR}$ is consistent when $\widehat{\mathrm{E}}[Y\,|\,L_0, A_0 = 1, L_1, A_1 = 1]$ and $\widehat{\mathrm{E}}[b_1(L_0, L_1)\,|\,A_0 = 1, L_0]$ are consistent for $\mathrm{E}(Y\,|\,L_0, A_0 = 1, L_1, A_1 = 1)$ and $\mathrm{E}[b_1(L_0, L_1)\,|\,A_0 = 1, L_0]$. Here $\widehat{\mathrm{E}}[b_1(L_0, L_1)\,|\,A_0 = 1, L_0]$ applies the same regression algorithm to the true $b_1(L_0, L_1)$ as was applied to $\widehat{b}_1(L_0, L_1)$ to obtain $\widehat{b}_0(L_0)$. To see this, we arrange terms to obtain $\widehat{U}_{TR} = \widehat{b}_0(L_0) + \frac{A_0A_1}{\widehat{\pi}_0\widehat{\pi}_1}(Y - \widehat{b}_1(L_0, L_1)) + \frac{A_0}{\widehat{\pi}_0}(\widehat{b}_1(L_0, L_1) - \widehat{b}_0(L_0))$. The sample averages of the last 2 terms are consistent for 0 and the sample average of the first term is $\widehat{\psi}_{gfor}$. Third, $\widehat{\psi}_{TR}$ is triply robust because it is consistent if both $\widehat{b}_1(L_0, L_1)$ and $\widehat{\pi}_0$ are consistent (Molina et al. 2017). This follows because $\widehat{U}_{TR}$ can be rewritten as $A_0\widehat{b}_1(L_0, L_1)/\widehat{\pi}_0 + \frac{A_0A_1}{\widehat{\pi}_0\widehat{\pi}_1}(Y - \widehat{b}_1(L_0, L_1)) - \left(\frac{A_0}{\widehat{\pi}_0} - 1\right)\widehat{b}_0(L_0)$. Hence, the sample average of the last 2 terms converges to zero and the sample average of the first converges to $\mathrm{E}[b_0(L_0\ )]$. However it is not consistent when only $\widehat{\pi}_1$ and $\widehat{\mathrm{E}}[b_1(L_0, L_1)\,|\,A_0 = a_0, L_0]$ are consistent.

By modifying our estimator $\widehat{\psi}_{TR}$ we can construct a quadruply robust (i.e., 2^{K+1}) estimator $\widehat{\psi}_{QR}$ that is consistent when only $\widehat{\pi}_1$ and $\widehat{\mathrm{E}}[b_1(L_0, L_1)\,|\,A_0 = a_0, L_0]$ are consistent (Tchetgen Tchetgen 2009). Let

$$\widetilde{b}_0(L_0) = \widehat{\mathrm{E}}\left[\frac{A_1Y}{\widehat{\pi}_1} - \left(\frac{A_1}{\widehat{\pi}_1} - 1\right)\ \widehat{b}_1(L_0, L_1)\,|\,A_0 = 1, L_0\right]$$

Then $\widehat{\psi}_{QR} = P_n[\widehat{U}_{QR}]$, where $\widehat{U}_{QR}$ is $\widehat{U}_{TR}$ except with $\widehat{b}_0(L_0)$ replaced by $\widetilde{b}_0(L_0)$. The advantage of $\widetilde{b}_0(L_0)$ over $\widehat{b}_0(L_0)$ is that $\widetilde{b}_0(L_0)$ is itself doubly robust in the sense that it is consistent for $b_0(L_0) = \mathrm{E}[b_1(L_0, L_1)\,|\,A_0 = 1, L_0]$ if $\widehat{\mathrm{E}}[b_1(L_0, L_1)\,|\,A_0 = 1, L_0]$ is consistent for $b_0(L_0)$ and either $\widehat{\pi}_1$ or $\widehat{b}_1(L_0, L_1) = \widehat{\mathrm{E}}[Y\,|\,L_0, A_0 = 1, L_1, A_1 = 1]$ are consistent, which implies that $\widehat{\psi}_{QR}$ is quadruply robust.

we reach the predicted value $\widehat{B}_0 = b_0\left(\bar{L}_0; \widehat{\beta}\right)$ we have completed step 2.

In step 3 we compute our estimate of $\widehat{\mathrm{E}}\left[Y^{\bar{a}=\bar{1}}\right]$ as the sample average over all individuals of $\widehat{B}_0$. If (i) the outcome models $b_m\left(\bar{L}_m; \beta_m\right)$ are correctly specified for all m, or (ii) the treatment models $\pi_k\left(\bar{L}_k; \alpha\right)$ are correctly specified for all m, then $\widehat{\mathrm{E}}\left[Y^{\bar{a}=\bar{1}}\right]$ will be (asymptotically) unbiased for $\mathrm{E}\left[Y^{\bar{a}=\bar{1}}\right]$. In that case, $\widehat{\mathrm{E}}\left[Y^{\bar{a}=\bar{1}}\right]$ is said to be doubly robust. However, $\widehat{\mathrm{E}}\left[Y^{\bar{a}=\bar{1}}\right]$ is actually multiply robust since it is also (asymptotically) unbiased for $\mathrm{E}\left[Y^{\bar{a}=\bar{1}}\right]$ if, for any $m \in \{0, 1, .., K-1\}$, the treatment model is correct for times 0 to m and the outcome model is correct for times $m+1$ to K. We refer to this property of the estimator as $K+2$ robustness. In Technical Points 21.4 and 21.5, we explain why these robustness properties are true and we show there exist estimators with even better robustness properties than $\widehat{\mathrm{E}}\left[Y^{\bar{a}=\bar{1}}\right]$. In fact, we

For a binary Y, we could fit a logistic model $b_m\left(\bar{L}_m; \beta_m\right) = \text{expit}\left[\gamma_m X_m + \varsigma_m \hat{W}^{\bar{1}_m}\right]$; X_m is a vector function of covariates $\bar{L}_m$ and $\beta_m = (\gamma_m, \varsigma_m)$. Even though $\widehat{B}_K$ is not a whole number, it is guaranteed to be in [0,1] and thus can be used as the outcome variable in a logistic model. For a continuous Y, we could fit a linear regression model $\gamma_m X_m + \varsigma_m \hat{W}^{\bar{1}_m}$.

Technical Point 21.5

A plug-in K+2 robust estimator. A potential drawback of the estimator $\widehat{\psi}_{TR}$ of Technical Point 21.12 was that, for binary Y, $\widehat{\psi}_{TR}$ could lie outside the support $[0,1]$ of ψ in a given sample. In contrast, $\widehat{\psi}_{gfor} = P_n\left[\widehat{b}_0(L_0)\right]$ is a plug-in estimator of ψ and always lies within $[0,1]$ if one estimates $\mathrm{E}[Y \mid L_0, A_0 = a_0, L_1, A_1 = a_1]$ and $b_0(L_0) = \mathrm{E}[b_1(L_0, L_1) \mid A_0 = a_0, L_0]$ using (parametric or nonparametric) logistic regression models. One obtains a plug-in estimator $\widehat{\psi}_{TR,plug} = P_n\left[\widehat{b}_0(L_0)\right]$ that is also triply robust if, for

$$\widehat{U}_{TR} = \widehat{b}_0(L_0) + \frac{A_0 A_1}{\widehat{\pi}_0 \widehat{\pi}_1}(Y - \widehat{b}_1(L_0, L_1)) + \frac{A_0}{\widehat{\pi}_0}(\widehat{b}_1(L_0, L_1) - \widehat{b}_0(L_0))$$

it can be guaranteed that $P_n\left[\frac{A_0 A_1}{\widehat{\pi}_0 \widehat{\pi}_1}(Y - \widehat{b}_1(L_0, L_1))\right]$ and $P_n\left[\frac{A_0}{\widehat{\pi}_0}(\widehat{b}_1(L_0, L_1) - \widehat{b}_0(L_0))\right]$ are both zero in every sample. For example, one achieves $P_n\left[\frac{A_0 A_1}{\widehat{\pi}_0 \widehat{\pi}_1}(Y - \widehat{b}_1(L_0, L_1))\right] = 0$ by including a univariate term $\theta_1\left\{\frac{A_0 A_1}{\widehat{\pi}_0 \widehat{\pi}_1}\right\}$ in a linear logistic model for $b_1(L_0, L_1) = \mathrm{E}\left[Y \mid L_0, A_0 = 1, L_1, A_1 = 1\right]$ with dependent variable Y fit by maximum likelihood to individuals with $A_0 = A_1 = 1$. One next achieves $P_n\left[\frac{A_0}{\widehat{\pi}_0}(\widehat{b}_1(L_0, L_1) - \widehat{b}_0(L_0))\right] = 0$ by including a term $\theta_0 \frac{A_0}{\widehat{\pi}_0}$ in a logistic model for $b_0(L_0) \equiv \mathrm{E}[b_1(L_0, L_1) \mid A_0 = a_0, L_0]$ with dependent variable $\widehat{b}_1(L_0, L_1)$ fit by maximizing a logistic likelihood to individuals with $A_0 = 1$. The estimator $\widehat{\mathrm{E}}[Y^{\bar{a}=\bar{1}}]$ given in the main text is an instance of $\widehat{\psi}_{TR,plug}$.

exhibit an estimator of $\mathrm{E}\left[Y^{\bar{a}=\bar{1}}\right]$ that is 2^{K+1} robust.

Molina et al. (2017) noted that this estimator was actually $K+2$ robust. Rotnitzky et al. (2017) studied the asymptotic bias of this and other multiply robust estimator when using nonparametric and machine learning estimators of the treatment and outcome regression functions.

To estimate the counterfactual mean $\mathrm{E}\left[Y^{\bar{a}=\bar{0}}\right]$ under the treatment strategy "never treated", repeat the above steps using $\bar{a} = \bar{0}$ where $\bar{a} = \bar{0}_K$ is the vector of $K+1$ 0's. The difference of means estimated under each strategy is a multiply robust estimator of the average causal effect $\mathrm{E}\left[Y^{\bar{a}=\bar{1}}\right] - \mathrm{E}\left[Y^{\bar{a}=\bar{0}}\right]$.

The multiply robust estimator described here can only be used to estimate the counterfactual mean $\mathrm{E}\left[Y^{\bar{a}}\right]$ under a static treatment strategy $\bar{a}$. Technical Point 21.6 describes a multiply robust estimator for the counterfactual mean $\mathrm{E}\left[Y^{g}\right]$ under a treatment strategy g that can be either static or dynamic and either deterministic or random. This estimator is sometimes referred to as a targeted minimum loss-based estimator (TMLE).

The implementation of multiply robust estimators has been historically hampered by computational constraints and lack of user-friendly software, especially for hazards-based survival analysis. We anticipate that, in the near future, software will become available and these multiply robust estimators (fit using machine learning and sample splitting) will become more common when studying the effect of complex treatment strategies on failure time outcomes. See Fine Point 21.2 for a description of the different representations of the g-formula and their connections to the above estimator.

21.4 G-estimation for time-varying treatments

If we were only interested in the effect of the time-fixed treatment A_1 in Table 21.1, we might have recourse to structural nested mean models for the conditional causal effect of a time-fixed treatment within levels of the covariates, as described in Chapter 14. Those models had a single equation because there was

Technical Point 21.6

A multiply robust estimator. Let $f^g\left(a_m|\overline{a}_{m-1},\overline{l}_m\right)$ denote the treatment density at time m under strategy g. For a static $\overline{a}^*$, $f^g\left(a_m|\overline{a}_{m-1},\overline{l}_m\right) = \mathrm{I}\left(a_m = a_m^*\right)$; for a deterministic dynamic g, $f^g\left(a_m|\overline{a}_{m-1},\overline{l}_m\right) = \mathrm{I}\left(a_m = g_m\left(\overline{a}_{m-1},\overline{l}_m\right)\right)$; and for a random dynamic f^{int}, $f^g\left(a_m|\overline{a}_{m-1},\overline{l}_m\right) = f^{int}\left(a_m|\overline{a}_{m-1},\overline{l}_m\right)$. Let $C_k^g = \mathrm{I}\left(\prod_{m=0}^k f^g\left(A_m|\overline{A}_{m-1},\overline{L}_m\right) = 0\right)$ equal 0 if an individual's observed treatment history $\overline{A}_k$ is compatible with g and 1 otherwise. The following algorithm computes a multiply robust plug-in estimator $\widehat{\psi}_{dr,plug}$ of $\psi = \mathrm{E}\left[Y^g\right]$ based on one proposed by Rotnitzky et al. (2017), which is closely related to estimators by Robins (2000), Bang and Robins (2005), van der Laan and Gruber (2012), and Petersen et al. (2014).

1. Fit models $f_m\left(a_m|\overline{a}_{m-1},\overline{l}_m;\alpha_m\right)$ for $f\left(a_m|\overline{a}_{m-1},\overline{l}_m\right)$ for $m = 0, 1...K.$. Obtain the MLE $\hat{\alpha}_m$ of the vector parameter α_m. For each time m, compute the weight $\hat{W}^{g,m} = \prod_{k=0}^m \frac{f^g\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)}{f_k\left(A_k|\bar{A}_{k-1},\bar{L}_k;\hat{\alpha}_k\right)}$

2. Set $\hat{T}_{K+1} = Y$.

3. Recursively, for $m = K, K-1, ..., 0$.

    (a) Fit a generalized linear model $b_m\left(\bar{A}_m,\bar{L}_m;\gamma_m,\varsigma_m\right) = \phi\left[\gamma_m d_m\left(\bar{A}_m,\bar{L}_m\right) + \varsigma_m \hat{W}^{g,m}\right]$, with ϕ an inverse canonical link, for the conditional expectation $\mathrm{E}\left[\hat{T}_{m+1}|\bar{A}_m,\bar{L}_m,C_m^g = 0\right]$ by iteratively reweighted least squares (IRLS) among individuals with $C_m^g = 0$; then $\left(\widehat{\gamma}_m,\widehat{\varsigma}_m\right)$ satisfies $\widehat{\mathrm{E}}\left\{\mathrm{I}\left(C_m^g = 0\right)\left({d_m\left(\bar{A}_m,\bar{L}_m\right) \atop \hat{W}^{g,m}}\right)\left(\hat{T}_{m+1} - b_m\left(\bar{A}_m,\bar{L}_m;\widehat{\gamma}_m,\widehat{\varsigma}_m\right)\right)\right\} = 0$

    (b) set $\hat{T}_m = \sum_{a_m} b_m\left(a_m,\bar{A}_{m-1},\bar{L}_m;\widehat{\gamma}_m,\widehat{\varsigma}_m\right) f^g\left(a_m|\bar{A}_{m-1},\bar{L}_m\right)$

4. $\widehat{\psi}_{dr,plug} = \widehat{\mathrm{E}}\left[\hat{T}_0\right]$

As pointed out by Molina et al. (2017), $\widehat{\psi}_{dr,plug}$ is $K+2$ robust because, in addition to being doubly robust, it is also (asymptotically) unbiased for ψ when, for any $p \in \{1, ..., K\}$, the models $b_m\left(\bar{A}_m,\bar{L}_m;\gamma_m,\varsigma_m\right)$ are correctly specified for $m \in \{K, ..., p\}$ and the models $f_m\left(a_m|\overline{a}_{m-1},\overline{l}_m;\alpha_m\right)$ are correctly specified for $m \in \{p-1, ..., 0\}$.

When $\hat{W}^{g,m}$ is not used as a covariate, the above algorithm computes the iterative conditional expectation (ICE) estimator of the g-formula for $\mathrm{E}[Y^g]$ (Fine Point 21.2), which is a non-doubly robust estimator of the g-formula.

a single time point $k = 0$. The extension to time-varying treatments requires that the model specifies as many equations as time points in the data. For the time-varying treatment $\overline{A} = (A_0, A_1)$ at two time points in Table 21.1, we specify a (saturated) *additive structural nested mean model* with two equations

$$\text{For time } k = 0\text{: } \mathrm{E}\left[Y^{a_0,a_1=0} - Y^{a_0=0,a_1=0}|A_0 = a_0\right] = \beta_0 a_0$$

$$\text{For time } k = 1\text{: } \mathrm{E}\left[Y^{a_0,a_1} - Y^{a_0,a_1=0}|L_1^{a_0} = l_1, A_0 = a_0, A_1^{a_0} = a_1\right] =$$

$$= a_1\left(\beta_{11} + \beta_{12}l_1 + \beta_{13}a_0 + \beta_{14}a_0 l_1\right)$$

By consistency, the conditional expectation for time $k = 1$ can be written as $\mathrm{E}\left[Y^{a_0,a_1} - Y^{a_0,a_1=0}|L_1 = l_1, A_0 = a_0, A_1 = a_1\right]$. Since we assume sequential exchangeability for Y, we can and will replace (i) the conditional expectation for $k = 0$ by $\mathrm{E}\left[Y^{a_0,a_1=0} - Y^{a_0=0,a_1=0}\right]$ since $A_0 = a_0$ can be removed from the conditioning event, and (ii) the conditional expectation for $k = 1$ by $\mathrm{E}\left[Y^{a_0,a_1} - Y^{a_0,a_1=0}|L_1^{a_0} = l_1, A_0 = a_0\right]$ since $A_1^{a_0} = a_1$ can be removed from the conditioning event.

Fine Point 21.2

Representations of the g-formula. The g-formula can be mathematically represented in several ways. These different representations of the g-formula are nonparametrically equivalent but lead to different estimators in practice. Throughout this book we have emphasized a representation of the g-formula that is the generalized version of standardization (in the epidemiologic jargon). That is, the g-formula for a mean outcome is $\sum_l \mathrm{E}\left[Y|A=a, L=l\right] f(l)$ for a time-fixed treatment and, as described in this chapter, $\sum_{\bar{l}} \mathrm{E}\left[Y|\bar{A}=\bar{a}, \bar{L}=\bar{l}\right] \prod_{k=0}^{K} f\left(l_k|\bar{a}_{k-1}, \bar{l}_{k-1}\right)$ for a time-varying treatment. Because a plug-in estimator based on this representation of the g-formula requires estimates of the joint density of the confounders $\prod_{k=0}^{K} f\left(l_k|\bar{a}_{k-1}, \bar{l}_{k-1}\right)$ over time, we refer to it as a joint density modeling estimator of the g-formula.

An alternative representation of the g-formula is as iterated conditional expectations. For a time-fixed treatment, we implicitly used this g-formula representation $\mathrm{E}\left[\mathrm{E}\left[Y|A=a, L=l\right]\right]$ in Section 13.3. For a time-varying treatment, the representation is an *iterated conditional expectation* (ICE) that can be recursively defined (Robins 1986). A plug-in estimator based on the ICE representation of the g-formula requires the fitting of sequential predictive algorithms (e.g., regression models). The ICE estimator is described in Section 21.3 and Technical Point 21.4, where we combine it with the estimation of IP weights to construct doubly (actually $K+2$) robust estimators.

Another representation of the g-formula is IP weighting. In fact, as shown in Technical Point 2.3 for time-fixed treatments, the standardized mean and the IP weighted mean are equal under positivity. The same is true for time-varying treatments (Robins and Rotnitzky, 1992; Robins, 1993; Young et al., 2014). As described in this chapter, an estimator based on the IP weighting representation of the g-formula requires the estimation of the conditional density of treatment over time given past treatment and covariate history. We refer to these estimators as IP weighted estimators rather than as g-formula estimators.

Hence the equation at time $k=1$ models the effect of treatment at time $k=1$ within each of the 4 treatment and covariate histories defined by (A_0, L_1). This component of the model is saturated because the 4 parameters β_1 in the second equation parameterize the effect of a_1 on Y within the 4 possible levels of past treatment and covariate history. The first equation models the effect of treatment at time $k=0$ when treatment at time $k=1$ is set to zero. This component of the model is also saturated because it has one parameter β_0 to estimate the effect within the only possible history (there is no prior treatment or covariates, so everybody has the same history). The two equations of the structural nested model are the reason why the model is referred to as *nested*. The first equation models the effect of receiving treatment at time 0 and never again after time 0, the second equation models the effect of receiving treatment at time 1 and never again after time 1, and so on if we had more time points.

Effect of a_1 is:

- β_{11} in individuals with $A_0 = 0, L_1^{a_0=0} = 0$
- $\beta_{11}+\beta_{12}$ in those with $A_0 = 0, L_1^{a_0=0} = 1$
- $\beta_{11}+\beta_{13}$ in those with $A_0 = 1, L_1^{a_0=1} = 0$
- $\beta_{11}+\beta_{12}+\beta_{13}+\beta_{14}$ in those with $A_0 = 1, L_1^{a_0=1} = 1$

By consistency, $L_1^{a_0} = L_1$ when $A_0 = a_0$.

Let us use g-estimation to estimate the parameters of our structural nested model with $K=1$. We follow the same approach as in Chapter 14. We start by considering the additive rank-preserving structural nested model for each individual i

$$Y_i^{a_0,0} = Y_i^{0,0} + \psi_0 a_0$$
$$Y_i^{a_0,a_1} = Y_i^{a_0,0} + \psi_{11}a_1 + \psi_{12}a_1 L_{1,i}^{a_0} + \psi_{13}a_1 a_0 + \psi_{14}a_1 a_0 L_{1,i}^{a_0},$$

where the second equation is restricted to individuals with $A_0 = a_0$. That is, the second equation is actually two equations, one for individuals with $A_0 = 1$ and one for individuals with $A_0 = 0$. This allows us to replace, by consistency, $L_{1,i}^{a_0}$ by $L_{1,i}$, which will be needed for identification of the model parameters from the observed data when, as in Figure 19.6, we do not have sequential exchangeability for L_1. We represent $Y_i^{a_0=0,a_1=0}$ by $Y_i^{0,0}$ to simplify the notation.

The first equation is a rank-preserving model because the effect ψ_0 is exactly the same for every individual. Thus if $Y_i^{0,0}$ for subject i exceeds $Y_j^{0,0}$ for subject j, the same ranking of i and j will hold for $Y^{1,0}$—the model preserves ranks across strategies. Also, under equation 2, if $Y_i^{1,0}$ for subject i exceeds $Y_j^{1,0}$ for subject j, we can only be certain that $Y_i^{1,1}$ for individual i also exceeds $Y_j^{1,1}$ for individual j if both have the same values a_0 of $A_{0,i}$ and l_1 of $L_{1,i} = L_{1,i}^{a_0}$. Because the preservation of the ranking is conditional on local factors (i.e., the value $L_1^{a_0=1}$), we refer to the second equation as a conditionally, or locally, rank-preserving model.

As discussed in Chapter 14, rank preservation is biologically implausible because of individual heterogeneity in unmeasured genetic and environmental risks. That is why our primary interest is in the structural nested mean model, which is totally agnostic as to whether or not there is additional effect heterogeneity across individuals due to unmeasured factors. However, given sequential exchangeability for Y, a class of g-estimators (described below) of ψ for the rank-preserving model are consistent for the parameters β of the mean model, even if the rank-preserving model is misspecified.

The proof can be found in Robins (1994). Note that to fit an unsaturated structural nested mean model by g-estimation, positivity is not required.

The first step in g-estimation is linking the model to the observed data, as we did in Chapter 14 for a time-fixed treatment. To do so, note that, by consistency, the counterfactual outcome Y^{a_0,a_1} is equal to the observed outcome Y for individuals who happen to be treated with treatment values a_0 and a_1. Formally, $Y^{a_0,a_1} = Y^{A_0,A_1} = Y$ for individuals with $(A_0 = a_0, A_1 = a_1)$. Similarly $Y^{a_0,0} = Y^{A_0,0}$ for individuals with $(A_0 = a_0, A_1 = 0)$, and $L_1^{a_0} = L_1$ for individuals with $A_0 = a_0$. Now we can rewrite the structural nested model in terms of the observed data as

$$\begin{aligned} Y^{A_0,0} &= Y - (\psi_{11}A_1 + \psi_{12}A_1L_1 + \psi_{13}A_1A_0 + \psi_{14}A_1A_0L_1) \\ Y^{0,0} &= Y^{A_0,0} - \psi_0 A_0 \end{aligned}$$

(we are omitting the individual index i to simplify the notation).

The second step in g-estimation is to use the observed data to compute the candidate counterfactuals $H_1\left(\psi^\dagger\right)$ and $H_0\left(\psi^\dagger\right)$. To do so, we use the structural nested model with the true value ψ of the parameter replaced by some value $\psi^\dagger$:

$$\begin{aligned} H_1\left(\psi^\dagger\right) &= Y - \left(\psi_{11}^\dagger A_1 + \psi_{12}^\dagger A_1L_1 + \psi_{13}^\dagger A_1A_0 + \psi_{14}^\dagger A_1A_0L_1\right) \\ H_0\left(\psi^\dagger\right) &= H_1\left(\psi^\dagger\right) - \psi_0^\dagger A_0 \end{aligned}$$

As in Chapter 14, the goal is to find the value $\psi^\dagger$ of the parameters that is equal to the true value ψ. When $\psi^\dagger = \psi$ and $\overline{A}_{k-1} = \overline{a}_{k-1}$, the candidate counterfactual $H_k\left(\psi^\dagger\right)$ equals the true counterfactual $Y^{\overline{a}_{k-1},\underline{0}_k}$ under treatment $\overline{a}_{k-1}$ through time $k-1$ and treatment 0 afterwards. We can now use sequential exchangeability to conduct g-estimation at each time point. Fine Point 21.3 describes how to g-estimate the parameters ψ of our saturated structural nested model. It turns out that all parameters of the structural nested model are 0, which implies that all counterfactual means $\mathrm{E}\left[Y^g\right]$ under any static or dynamic strategy g are equal to 60. This result agrees with those obtained by the g-formula and by IP weighting. G-estimation, like the g-formula and IP weighting, succeeds where traditional methods failed.

In practice, however, we will encounter observational studies with multiple times k and multiple covariates L_k at each time. In general, a structural

Fine Point 21.3

G-estimation with a saturated structural nested model. Sequential exchangeability at $k=1$ implies that, within any of the four joint strata of (A_0, L_1), the mean of $Y^{A_0,0}$ among individuals with $A_1 = 1$ is equal to the mean among individuals with $A_1 = 0$. Therefore, the means of $H_1\left(\psi^\dagger\right)$ must also be equal when $\psi^\dagger = \psi$.

Consider first the stratum $(A_0, L_1) = (0,0)$. From data rows 1 and 2 in Table 21.2, we find that the mean of $H_1(\psi)$ is 84 when $A_1 = 0$ and $84 - \psi_{11}$ when $A_1 = 1$. Hence $\psi_{11} = 0$. Next we equate the means of $H_1(\psi)$ in data rows 3 and 4 corresponding to stratum $(A_0, L_1) = (0,1)$ to obtain $52 = 52 - \psi_{11} - \psi_{12}$. Since $\psi_{11} = 0$, we conclude $\psi_{12} = 0$. Continuing we equate the means of $H_1(\psi)$ in data rows 5 and 6 to obtain $76 = 76 - \psi_{11} - \psi_{13}$. Since $\psi_{11} = \psi_{12} = 0$, we conclude $\psi_{13} = 0$. Finally, equating the means of $H_1(\psi)$ in data rows 7 and 8, we obtain $44 = 44 - \psi_{11} - \psi_{12} - \psi_{13} - \psi_{14}$ so $\psi_{14} = 0$ as well.

To estimate ψ_0, we first substitute the values ψ_{11}, ψ_{12}, ψ_{13}, and ψ_{14} into the expression for the mean of $H_0(\psi)$ in Table 21.2. In this example, all parameters were equal to 0, so the mean of $H_0(\psi)$ was equal to the mean of the observed outcome Y. We then use the first equation of the structural equation model to compute the mean of $H_0(\psi)$ for each data row in the table by subtracting $\psi_0 A_0$ from the mean of $H_1(\psi)$, as shown in Table 21.3. Sequential exchangeability $Y^{0,0} \perp\!\!\!\perp A_0$ at time $k=0$ implies that the means of $H_0(\psi)$ among the $16,000$ subjects with $A_0 = 1$ and the $16,000$ subjects with $A_0 = 0$ are identical. The mean of $H_0(\psi)$ is $84 \times 0.25 + 52 \times 0.75 = 60$ among individuals with $A_0 = 0$, $(76 - \psi_0) \times 0.5 + (44 - \psi_0) \times 0.5 = 60 - \psi_0$ among individuals with $A_0 = 1$. Hence $\psi_0 = 0$. We have completed g-estimation.

Table 21.2

A_0	L_1	A_1	Mean $H_1(\psi)$
0	0	0	84
0	0	1	$84 - \psi_{1,1}$
0	1	0	52
0	1	1	$52 - \psi_{11} - \psi_{12}$
1	0	0	76
1	0	1	$76 - \psi_{11} - \psi_{13}$
1	1	0	44
1	1	1	$44 - \psi_{11} - \psi_{12}$ $-\psi_{13} - \psi_{14}$

Table 21.3

A_0	L_1	A_1	Mean $H_0(\psi)$
0	0	0	84
0	0	1	84
0	1	0	52
0	1	1	52
1	0	0	$76 - \psi_0$
1	0	1	$76 - \psi_0$
1	1	0	$44 - \psi_0$
1	1	1	$44 - \psi_0$

nested mean model has as many equations as time points $k = 0, 1...K$. The most general form of structural nested mean models that we discuss in the main text is the following (even more general structural nested mean models are discussed in Technical Point 21.13). For each time $k = 0, 1...K$,

$$\begin{aligned} \mathrm{E}\left[Y^{\overline{a}_{k-1}, a_k, \underline{0}_{k+1}} - Y^{\overline{a}_{k-1}, \underline{0}_k} | \overline{L}_k^{\overline{a}_{k-1}} = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1}, A_k = a_k\right] \\ = a_k \gamma_k\left(\overline{a}_{k-1}, \overline{l}_k, \beta\right) \end{aligned}$$

where $\left(\overline{a}_{k-1}, a_k, \underline{0}_{k+1}\right)$ is a static strategy that assigns treatment $\overline{a}_{k-1}$ between times 0 and $k-1$, treatment a_k at time k, and treatment 0 from time $k=1$ until the end of follow-up K. The strategies $\left(\overline{a}_{k-1}, a_k, \underline{0}_{k+1}\right)$ and $\left(\overline{a}_{k-1}, \underline{0}_k\right)$ differ only in that the former has treatment a_k at k while the latter has treatment 0 at time k. Here each $\gamma_k\left(\overline{a}_{k-1}, \overline{l}_k, \psi^\dagger\right)$ is a known function of a parameter vector $\psi^\dagger$ such that $\gamma_k\left(\overline{a}_{k-1}, \overline{l}_k, \psi^\dagger = 0\right) = 0$ and β is the true value of $\psi^\dagger$. Again, under sequential exchangeability for Y, we can drop $A_k = a_k$ from the above conditioning event. In our example with $K = 1$, $\gamma_0\left(\overline{a}_{-1}, \overline{l}_0, \beta\right)$ is just β_0 ($\overline{l}_0$ and $\overline{a}_{-1}$ can both be taken to be identically 0) and $\gamma_1\left(\overline{a}_0, \overline{l}_1, \beta\right)$ is $\beta_{11} + \beta_{12} l_1 + \beta_{13} a_0 + \beta_{14} a_0 l_1$.

Thus, a structural nested mean model is a model for the effect on the mean of Y of a last blip of treatment of magnitude a_k at k, as a function $\gamma_k\left(\overline{a}_{k-1}, \overline{l}_k, \beta\right)$ of past treatment and covariate history $\left(\overline{a}_{k-1}, \overline{l}_k\right)$. See Technical Point 21.7 for the relationship between structural nested models and marginal structural models.

We are now ready to discuss estimation of the parameters of a general structural nested mean model with blip funcion $\gamma_k\left(\overline{a}_{k-1}, \overline{l}_k, \beta\right)$. To motivate our estimation procedure, we will use the fact that a correctly specified locally rank preserving model with true parameter ψ is also a correctly specified structural nested mean model with true parameter $\beta = \psi$ (though the converse is

This blip function satisfies $\gamma_k\left(\overline{a}_{k-1}, \overline{l}_k, 0\right) = 0$ so $\beta = 0$ under the null hypothesis of no effect of treatment.

Technical Point 21.7

Marginal structural models and structural nested models. A structural nested mean model is a semiparametric marginal structural mean model if and only if, for all $\left(\overline{a}_{k-1}, \overline{l}_k, \beta\right)$,

$$\gamma_k\left(\overline{a}_{k-1}, \overline{l}_k, \beta\right) = \gamma_k\left(\overline{a}_{k-1}, \beta\right)$$

does not depend on $\overline{l}_k$. Specifically, it is a semiparametric marginal structural mean model with the functional form

$$\mathrm{E}\left[Y^{\overline{a}}\right] = \alpha_0 + \sum_{k=0}^{K} a_k \gamma_k\left(\overline{a}_{k-1}, \beta\right),$$

where $a_0 = \mathrm{E}\left[Y^{\overline{0}_K}\right]$ is an unknown constant. However, such a structural nested mean model is not simply a marginal structural mean model, because it also imposes the additional strong assumption that effect modification by past covariate history is absent. In contrast, a marginal structural model is agnostic as to whether there is effect modification by time-varying covariates.

If we specify a structural nested mean model $\gamma_k\left(\overline{a}_{k-1}, \beta\right)$, then we can estimate β either by g-estimation or IP weighting. However the most efficient g-estimator will be more efficient than the most efficient IP weighted estimator when the structural nested mean model (and thus the marginal structural mean model) is correctly specified, because g-estimation uses the additional assumption of no effect modification by past covariates to increase efficiency.

In contrast, suppose the marginal structural mean model is correct but the structural nested mean model is incorrect because $\gamma_k\left(\overline{a}_{k-1}, \overline{l}_k, \beta\right) \neq \gamma_k\left(\overline{a}_{k-1}, \beta\right)$. Then the g-estimates of β and $\mathrm{E}\left[Y^{\overline{a}}\right]$ will be biased, while the IP weighted estimates remain unbiased. Thus we have a classic variance-bias trade off. Given the marginal structural model, g-estimation can increase efficiency if $\gamma_k\left(\overline{a}_{k-1}, \overline{l}_k, \beta\right) = \gamma_k\left(\overline{a}_{k-1}, \beta\right)$, but introduces bias otherwise.

not true). Given a structural nested mean model, we can define

$$H_k\left(\psi^{\dagger}\right) = Y - \sum_{j=k}^{K} A_j \gamma_j\left(\overline{A}_{j-1}, \overline{L}_j, \psi^{\dagger}\right)$$

A correctly specified locally rank preserving model with true parameter vector ψ is equivalent to the statement that $H_k(\psi)$ is exactly equal to the counterfactual $Y^{\overline{A}_{k-1}, \underline{0}_k}$ in which the effects of the treatments from time j through K have been removed. In particular, $H_0(\psi)$ is the value of $Y^{\overline{0}}$ under no treatment.

However, if the assumption of local rank preservation is incorrect (as will essentially always be the case if there is a treatment effect) but the structural nested mean model is correct, we still have that $\mathrm{E}\left[H_k(\beta) | \overline{A}_k, \overline{L}_k\right]$ equals $\mathrm{E}\left[Y^{\overline{A}_{k-1}, \underline{0}_k} | \overline{A}_k, \overline{L}_k\right]$ and that $\mathrm{E}\left[H_0(\beta)\right]$ equals $\mathrm{E}\left[Y^{\overline{0}}\right]$. Thus, $\mathrm{E}\left[Y^{\overline{0}}\right]$ can be consistently estimated as the sample average of $H_0\left(\widehat{\beta}\right)$ if we obtain a consistent estimator of $\widehat{\beta}$. This is what g-estimation provides.

With multiple time points or covariates, we will need to fit an unsaturated structural nested mean model. For example, we might hypothesize that the function $\gamma_k\left(\overline{a}_{k-1}, \overline{l}_k, \beta\right)$ is the same for all k. The simplest model would be $\gamma_k\left(\overline{a}_{k-1}, \overline{l}_k, \beta\right) = \beta_1$, which assumes that the effect of a last blip of treatment is the same for all past histories and all times k. Other options are $\beta_1 + \beta_2 k$, which assumes that the effect varies linearly with the time k of treatment, and $\beta_1 + \beta_2 k + \beta_3 a_{k-1} + \beta_4 l_k + \beta_5 l_k a_{k-1}$, which allows the effect of treatment at k to be modified by the most recent treatment and covariate values.

To describe g-estimation for structural nested mean models with multiple time points, suppose the nonsaturated model is $\gamma_k\left(\overline{a}_{k-1}, \bar{l}_k, \beta\right) = \beta_1$. The corresponding rank-preserving model entails $H_k\left(\psi^\dagger\right) = Y - \sum_{j=k}^{K} A_j \psi^\dagger$, which can be computed from the observed data for any value $\psi^\dagger$. We will then choose values ψ_{low} and ψ_{up} that are much smaller and larger, respectively, than any substantively plausible value of ψ, and will compute (for each individual and time) the value of $H_k\left(\psi^\dagger\right)$ for each $\psi^\dagger$ on a grid from ψ_{low} to ψ_{up}, say $\psi_{low}, \psi_{low}+0.1, \psi_{low}+0.2, ..., \psi_{up}$.

Then, for each value of $\psi^\dagger$, we will fit a pooled (over time) logistic regression model

$$\text{logit}\Pr\left[A_k = 1 | H_k\left(\psi^\dagger\right), \overline{L}_k, \overline{A}_{k-1}\right] = \alpha_0 + \alpha_1 H_k\left(\psi^\dagger\right) + \alpha_2 W_k$$

for the probability of treatment at time k for $k = 0, ..., K$. Here $W_k = w_k\left(\overline{L}_k, \overline{A}_{k-1}\right)$ is a vector of covariates calculated from an individual's covariate and treatment data $\left(\overline{L}_k, \overline{A}_{k-1}\right)$, α_2 is a row vector of unknown parameters, and each person contributes $K+1$ observations. The g-estimate of β is the grid value of $\psi^\dagger$ for which the estimate of α_1 is closest to 0. We can eliminate the need to search over the grid by defining the estimate $\widehat{\beta}$ to be the value of $\psi^\dagger$ such that the p-value of the score test of $\alpha_1 = 0$ is equal to 1. That is $\widehat{\beta}$ is the value of $\psi^\dagger$ that solves

The limits of the 95% confidence interval for ψ are the limits of the set of values $\psi^\dagger$ that result in a P-value > 0.05 when testing for $\alpha_1 = 0$.

$$\sum_{i=1,k=0}^{i=N,k=K} \left\{A_i - \text{expit}\left(\widehat{\alpha}_0 + \widehat{\alpha}_2 W_{i,k}\right)\right\} H_{i,k}\left(\psi^\dagger\right) = 0$$

where $\widehat{\alpha}_0$ and $\widehat{\alpha}_2$ are obtained by fitting the above logistic model with the term α_1 set to 0. Standard equation solvers can be used. The estimator $\widehat{\beta}$ will be consistent if (i) the structural nested mean model is correct, (ii) sequential exchageability for Y holds, (iii) the model $\text{logit}\Pr\left[A_k = 1 | \overline{L}_k, \overline{A}_{k-1}\right] = \alpha_0 + \alpha_0 W_k$ is correct, and (iv) $H_k\left(\psi^\dagger\right)$ enters the above logistic model linearly (i.e., as $H_k\left(\psi^\dagger\right)$) rather than as $\left\{H_k\left(\psi^\dagger\right)\right\}^2$ or any other non-linear function (see Technical Point 14.2).

The procedure described above is the generalization to time-varying treatments of the g-estimation procedure described in Chapter 14. For simplicity, we considered a structural nested model with a single parameter β_1, which implies that the effect does not vary over time k or by treatment and covariate history. Suppose now that the parameter β is a vector. To be concrete suppose we consider the model with $\gamma_k\left(\overline{a}_{k-1}, \bar{l}_k, \beta\right) = \beta_0 + \beta_1 k + \beta_2 a_{k-1} + \beta_3 l_k + \beta_4 l_k a_{k-1}$ so β is 5-dimensional and l_m is 1-dimensional. Now to estimate 5 parameters one requires 5 additional covariates in the treatment model. For example, we could fit the model $\text{logit}\Pr\left[A_k = 1 | H_k\left(\psi^\dagger\right), \overline{L}_k, \overline{A}_{k-1}\right] =$

$$\alpha_0 + H_k\left(\psi^\dagger\right)\left(\alpha_1 + \alpha_2 k + \alpha_3 A_{k-1} + \alpha_4 L_k + \alpha_5 L_k A_{k-1}\right) + \alpha_6 W_k$$

The particular choice of covariates does not affect the consistency of the point estimate of β, but it determines the width of its confidence interval.

The earlier g-estimation procedure then requires a search over a 5-dimensional grid, one dimension for each component β_j of β. So if we had 20 grid points for each component we would have 20^5 different values of β on our 5 dimensional grid. However, when the dimension of β is greater than 2, finding the g-estimate $\widehat{\beta}$ by a grid search may be computationally difficult. In that case we can eliminate the need to search over the grid by defining the g-estimate $\widehat{\beta}$ to be the

A 95% joint confidence interval for β_j are the set of values for which the 5 degree-of-freedom score test does not reject at the 5% level. A less computationally demanding approach is the univariate 95% Wald confidence interval$\widehat{\beta}_j \pm 1.96$ times its standard error.

Technical Point 21.8

A closed form estimator for linear structural nested mean models. When, as in all the examples we have discussed, $\gamma_k\left(\overline{A}_{k-1}, \overline{L}_k, \beta\right) = \beta^T R_k$ is linear in β with $R_k = r_k\left(\bar{L}_k, \bar{A}_{k-1}\right)$ being a vector of known functions, then, given the model $\operatorname{logit}\Pr\left[A_k = 1|\overline{L}_k, \overline{A}_{k-1}\right] = \alpha^T W_k$, there is an explicit closed form expression for $\widehat{\beta}$ given by

$$\widehat{\beta} = \left\{\sum_{i=1,k=0}^{i=N,k=K} A_{i,k} X_{i,k}\left(\widehat{\alpha}\right) Q_{i,k} S_{i,k}^T\right\}^{-1} \left\{\sum_{i=1,k=0}^{i=N,k=K} Y_i X_{i,k}\left(\widehat{\alpha}\right) Q_{i,k}\right\}$$

with $X_{i,k}\left(\widehat{\alpha}\right) = \left[A_{i,k} - \operatorname{expit}\left(\widehat{\alpha}^T W_{i,k}\right)\right]$, $S_{i,k} = \sum_{i=1,j=k}^{i=N,j=K} R_{i,j}$, and the choice of dimension-β functions $Q_{i,k} = q_k\left(\bar{L}_{i,k}, \bar{A}_{i,k-1}\right)$ affects efficiency but not consistency. See Robins (1994) for the optimal choice of Q_k.

In fact, when $\gamma_k\left(\overline{a}_{k-1}, \bar{l}_k, \beta\right)$ is linear in β, we can obtain a closed-form 2^{K+1} multiply robust estimator $\widetilde{\beta}$ of β by specifying a working model $\varsigma^T D_k = \varsigma^T d_k\left(\bar{L}_k, \bar{A}_{k-1}\right)$ for $\operatorname{E}\left[H_k\left(\beta\right)|\bar{L}_k, \bar{A}_{k-1}\right] = \operatorname{E}\left[Y^{\overline{A}_{k-1}, \underline{0}_k}|\bar{L}_k, \bar{A}_{k-1}\right]$ and defining

$$\begin{pmatrix} \widetilde{\beta} \\ \widetilde{\varsigma} \end{pmatrix} = \left\{\sum_{i=1,k=0}^{i=N,k=K} \begin{pmatrix} A_{i,k} X_{i,k}\left(\widehat{\alpha}\right) Q_{i,k} \\ D_{i,k} \end{pmatrix} \left(S_{i,k}^T, D_{i,k}^T\right)\right\}^{-1} \left\{\sum_{i=1,k=0}^{i=N,k=K} Y_i \begin{pmatrix} X_{i,k}\left(\widehat{\alpha}\right) Q_{i,k} \\ D_{i,k} \end{pmatrix}\right\}$$

Specifically, $\widetilde{\beta}$ will be a consistently asymptotically normal estimator of ψ if, for each k, either the model $\varsigma^T D_k$ for $\operatorname{E}\left[Y^{\overline{A}_{k-1}, \underline{0}_k}|\bar{L}_k, \bar{A}_{k-1}\right]$ is correct or the model for $\operatorname{logit}\Pr\left[A_k = 1|\overline{L}_k, \overline{A}_{k-1}\right]$ is correct.

value of $\psi^\dagger$ such that the p-value of the score test of $\alpha_{1-5} = \left(\alpha_1, ..., \alpha_5\right)^T = 0$ is equal to 1. That is $\widehat{\beta}$ is the value of $\psi^\dagger$ that solves the 5 dimensional estimating equation

$$\sum_{i=1,k=0}^{i=N,k=K} \left\{A_i - \operatorname{expit}\left(\widehat{\alpha}_0 + \widehat{\alpha}_6^T W_{i,k}\right)\right\} H_{i,k}\left(\psi^\dagger\right)\left(1, k, A_{i,k-1}, L_{i,k}, L_{i,k}A_{i,k-1}\right)^T = 0$$

where $\widehat{\alpha}_0$ and $\widehat{\alpha}_6$ are obtained by fitting the above logistic model with α_{1-5} set to zero. Standard equation solvers can be used. Indeed, the solution $\widehat{\beta}$ to this last equation exists in closed form when, as in all examples discussed in this section, the structural nested mean model is linear in β. See Technical Point 21.8, which also describes a multiply robust form of the estimator.

Given a consistent g-estimator $\widehat{\beta}$ of the parameters of the structural nested mean model, the last step is the estimation of the counterfactual mean $\operatorname{E}\left[Y^g\right]$ under the strategies g of interest. As discussed earlier, $\operatorname{E}\left[Y^{\overline{0}}\right]$ can be consistently estimated by the sample average $\widehat{\operatorname{E}}\left[H_0\left(\widehat{\beta}\right)\right]$. If there is no effect modification by past covariate history, i.e., $\gamma_k\left(\overline{a}_{k-1}, \bar{l}_k, \beta\right) = \gamma_k\left(\overline{a}_{k-1}, \beta\right)$ then $\operatorname{E}\left[Y^{\overline{a}}\right]$ under a static strategy $\overline{a}$ is estimated as

$$\widehat{\operatorname{E}}\left[Y^{\overline{a}}\right] = \widehat{\operatorname{E}}\left[Y^{\overline{0}_K}\right] + \sum_{k=0}^{K} a_k \gamma_k\left(\overline{a}_{k-1}, \widetilde{\beta}\right)$$

On the other hand, if the structural nested mean model depends on L_k or we want to estimate $\operatorname{E}\left[Y^g\right]$ under a dynamic strategy g, then we need to simulate the L_k using the algorithm described in Technical Point 21.9.

Technical Point 21.9

Estimation of $\mathrm{E}\left[Y^{g}\right]$ after g-estimation of a structural nested mean model. Suppose the identifiability assumptions hold, one has obtained a doubly robust g-estimate $\widetilde{\beta}$ of a structural nested mean model $\gamma_k\left(\bar{a}_{k-1}, \bar{l}_k, \beta\right)$ and one wishes to estimate $\mathrm{E}\left[Y^{g}\right]$ under a dynamic strategy g. To do so, one can use the following steps of a Monte Carlo algorithm:

1. Estimate the mean response $\mathrm{E}\left[Y^{\bar{0}_K}\right]$ had treatment always been withheld by the sample average of $H_0\left(\widetilde{\beta}\right)$ over the N study subjects. Call the estimate $\widehat{\mathrm{E}}\left[Y^{\bar{0}_K}\right]$.
2. Fit a parametric model for $f\left(l_k|\bar{a}_{k-1}, \bar{l}_{k-1}\right)$ to the data, pooled over persons and times, and let $\widehat{f}\left(l_k|\bar{a}_{k-1}, \bar{l}_{k-1}\right)$ denote the estimate of $f\left(l_k|\bar{a}_{k-1}, \bar{l}_{k-1}\right)$ under the model.
3. Do for $v = 1, ..., V$,
   (a) Draw $l_{v,0}$ from $\widehat{f}\left(l_0\right)$.
   (b) Recursively for $k = 1, ..., K$ draw $l_{v,k}$ from $\widehat{f}\left(l_k|\bar{a}_{v,k-1}, \bar{l}_{v,k-1}\right)$ with $\bar{a}_{v,k-1} = \bar{g}_{k-1}\left(\bar{l}_{v,k-1}\right)$, the treatment history corresponding to the strategy g.
   (c) Let $\widehat{\Delta}_{g,v} = \sum_{j=0}^{j=K} a_{v,j}\gamma_j\left(\bar{a}_{v,j-1}, \bar{l}_{v,j}, \widetilde{\beta}\right)$ be the v^{th} Monte Carlo estimate of $Y^{g} - Y^{\bar{0}_K}$, where $a_{v,j} = g_j\left(\bar{l}_{v,j-1}\right)$.
4. Let $\widehat{\mathrm{E}}\left[Y^{g}\right] = \widehat{\mathrm{E}}\left[Y^{\bar{0}_K}\right] + \sum_{v=1}^{v=V}\widehat{\Delta}_{g,v}/V$ be the estimate of $\widehat{\mathrm{E}}\left[Y^{g}\right]$.

If the model for $f\left(l_k|\bar{a}_{k-1}, \bar{l}_{k-1}\right)$, the structural nested mean model $\gamma_k\left(\bar{a}_{k-1}, \bar{l}_k, \beta\right)$, and either the treatment model $\Pr\left[A_k = 1|\bar{L}_k, \bar{A}_{k-1}\right]$ or the outcome model $\mathrm{E}\left[Y^{\bar{A}_{k-1}, \underline{0}_k}|\bar{L}_k, \bar{A}_{k-1}\right]$ are correctly specified, then $\widehat{\mathrm{E}}\left[Y^{g}\right]$ is consistent for $\mathrm{E}\left[Y^{g}\right]$. Confidence intervals can be obtained using the nonparametric bootstrap.

Note that $\gamma_k\left(\bar{a}_{k-1}, \bar{l}_k, \widetilde{\beta}\right)$ will converge to 0 if the estimate $\widetilde{\beta}$ is consistent for $\beta = 0$. Thus $\widehat{\Delta}_{g,v}$ will converge to zero and $\widehat{\mathrm{E}}\left[Y^{g}\right]$ to $\widehat{\mathrm{E}}\left[Y^{\bar{0}_K}\right]$ even if the model for $f\left(l_k|\bar{a}_{k-1}, \bar{l}_{k-1}\right)$ is incorrect. That is, the structural nested mean model preserves the null if the identifiability conditions hold and we either know (as in a sequentially randomized experiment) $\Pr\left[A_k = 1|\bar{L}_k, \bar{A}_{k-1}\right]$ or have a correct model for either $\Pr\left[A_k = 1|\bar{L}_k, \bar{A}_{k-1}\right]$ or $\mathrm{E}\left[Y^{\bar{A}_{k-1}, \underline{0}_k}|\bar{L}_k, \bar{A}_{k-1}\right]$ for each k.

21.5 Censoring is a time-varying treatment

You may want to re-read Section 12.6 for a refresher on censoring.

Throughout this chapter we have used an example in which there is no censoring: the outcomes of all individuals in Table 21.1 are known. In practice, however, we will often encounter situations in which some individuals are lost to follow-up and therefore their outcome values are unknown or (right-)censored. We have discussed censoring and methods to handle it in Part II of the book. In Chapter 8, we showed that censoring may introduce selection bias, even under the null. In Chapter 12, we discussed how we are generally interested in the causal effect if nobody in the study population had been censored.

However, in Part II we only considered a greatly simplified version of censoring under which we did not specify *when* individuals were censored during the follow-up. That is, we considered censoring C as a time-fixed variable. A more realistic view of censoring is as a time-varying variable $C_1, C_2, ...C_{K+1}$, Conditioning on being uncensored ($C = 0$) induces selection bias under the null when C is either a collider on a pathway between treatment A and the outcome Y, or the descendant of one such collider.

where C_m is an indicator that takes value 0 if the individual remains uncensored at time m and takes value 1 otherwise. Censoring is a monotonic type of missing data, i.e., if an individual's $C_m = 0$ then all previous censoring indicators are also zero ($C_1 = 0, C_2 = 0....C_{m-1} = 0$). Also, by definition, $C_0 = 0$ for all individuals in a study; otherwise they would have not been included in the study.

If an individual is censored at time m, i.e., when $C_m = 1$, then treatments, confounders, and outcomes measured after time m are unobserved. Therefore, the analysis becomes necessarily restricted to uncensored person-times, i.e., those with $C_m = 0$. For example, the g-formula for the counterfactual mean outcome $\mathrm{E}\left[Y^{\bar{a}}\right]$ from section 21.1 needs to be rewritten as

$$\sum_{\bar{l}} \mathrm{E}\left[Y|\bar{C} = \bar{0}, \bar{A} = \bar{a}, \bar{L} = \bar{l}\right] \prod_{k=0}^{K} f\left(l_k|c_k = 0, \bar{a}_{k-1}, \bar{l}_{k-1}\right),$$

with all the terms being conditional on remaining uncensored.

Suppose the identifiability conditions hold with treatment A_m replaced by (A_m, C_{m+1}) at all times m. Then it is easy to show that the above expression corresponds to the g-formula for the counterfactual mean outcome $\mathrm{E}\left[Y^{\bar{a},\bar{c}=\bar{0}}\right]$ under the joint treatment $(\bar{a}, \bar{c} = \bar{0})$, i.e., the mean outcome that would have been observed if all individuals have received treatment strategy $\bar{a}$ and no individual had been lost to follow-up.

The use of the superscript $\bar{c} = \bar{0}$ makes it explicit the causal contrast that many have in mind when they refer to the causal effect of treatment $\bar{A}$, even if they choose not to use the superscript $\bar{c} = \bar{0}$.

The counterfactual mean $\mathrm{E}\left[Y^{\bar{a},\bar{c}=\bar{0}}\right]$ can also be estimated via IP weighting of a structural mean model when the identifiability conditions hold for the joint treatment $\left(\bar{A}, \bar{C}\right)$. To estimate this mean, we might fit, e.g., the outcome regression model

$$\mathrm{E}\left[Y|\overline{A}, \bar{C} = \bar{0}\right] = \theta_0 + \theta_1 cum\left(\overline{A}\right)$$

to the pseudo-population created by the nonstabilized IP weights $W^{\bar{A}} \times W^{\bar{C}}$ where

$$W^{\bar{C}} = \prod_{k=1}^{K+1} \frac{1}{\Pr\left(C_k = 0|C_{k-1} = 0, \bar{A}_{k-1}, \bar{L}_{k-1}\right)}$$

We estimate the denominator of the weights by fitting a logistic regression model for $\Pr\left(C_k = 0|C_{k-1} = 0, \bar{A}_{k-1}, \bar{L}_{k-1}\right)$. Technical Point 21.10 shows the extension to survival analysis with a failure time outcome.

In the pseudo-population created by the nonstabilized IP weights, the censored individuals are replaced by copies of uncensored individuals with the same values of treatment and covariate history. Therefore the pseudo-population has the same size as the original study population *before* censoring, that is, before any losses to follow-up occur. The nonstabilized IP weights abolish censoring in the pseudo-population.

Or we can use the pseudo-population created by the stabilized IP weights $SW^{\bar{A}} \times SW^{\bar{C}}$, where

$$SW^{\bar{C}} = \prod_{k=1}^{K+1} \frac{\Pr\left(C_k = 0|C_{k-1} = 0, \bar{A}_{k-1}\right)}{\Pr\left(C_k = 0|C_{k-1} = 0, \bar{A}_{k-1}, \bar{L}_{k-1}\right)}$$

Remember:
The estimated IP weights $SW^{\bar{C}}$ have mean 1 when the model for $\Pr\left(C_k = 0|\bar{A}_{k-1}, C_{k-1} = 0, \bar{L}_k\right)$ is correctly specified.

We estimate the denominator and numerator of the IP weights via two separate models for $\Pr\left(C_k = 0|C_{k-1} = 0, \bar{A}_{k-1}, \bar{L}_{k-1}\right)$ and $\Pr\left(C_k = 0|C_{k-1} = 0, \bar{A}_{k-1}\right)$, respectively.

The pseudo-population created by the stabilized IP weights is of the same size as the original study population *after* censoring, i.e., the proportion of

Technical Point 21.10

Survival analysis with time-varying treatments. Chapter 17 describes g-methods to estimate the effect of point interventions on failure time outcomes. This chapter describes g-methods to estimate the effect of sustained strategies on non-failure time outcomes. In practice, we often use g-methods to estimate the effect of sustained strategies on failure time outcomes by combining the methods described in Chapter 17 with those in this chapter. Below we sketch two approaches, based on the g-formula and on IP weighting, to estimate the counterfactual risk $\Pr\left[D_{k+1}^{\bar{a},\bar{c}=\bar{0}}=1\right]$ under treatment strategy $\bar{a}$ if sequential exchangeability, positivity, and consistency hold. The causal diagram in Figure 21.4 depicts such setting with two time points and the failure time outcome represented by time-varying indicators as in Chapter 17. From each indicator D_k there should be arrows into all future variables on the graph, but we omitted these arrows to reduce clutter. For simplicity, we also omitted the time-varying indicators for censoring.

The risk $\Pr\left[D_{k+1}^{\bar{a},\bar{c}=\bar{0}}=1\right]$ is identified by 1 minus the g-formula for $\Pr\left[D_{k+1}^{\bar{a},\bar{c}=\bar{0}}=0\right]$:

$$\sum_{\bar{l}_k}\Pr\left[D_{k+1}=0|\bar{A}_k=\bar{a}_k,\bar{L}_k=\bar{l}_k,D_k=C_{k+1}=0\right]\times$$

$$\prod_{m=0}^{k} f\left(l_m|\bar{a}_{m-1},\bar{l}_{m-1},D_m=C_m=0\right)\Pr\left[D_m=0|\bar{A}_{m-1}=\bar{a}_{m-1},\bar{L}_{m-1}=\bar{l}_{m-1},D_{m-1}=C_m=0\right].$$

A plug-in g-formula estimate can then be obtained by fitting models for the discrete-time hazards $\Pr\left[D_{k+1}=1|\bar{A}_k=\bar{a}_k,\bar{L}_k=\bar{l}_k,D_k=C_{k+1}=0\right]$ and for the conditional density $f\left(l_k|\bar{a}_{k-1},\bar{l}_{k-1},D_k=C_k=0\right)$ of the confounders L over time. As described in Chapter 17, a pooled logistic model can be used to approximate the hazards. See Young et al. (2011) for details and an application. Wen et al. (2021) describe ICE g-formula estimators.

An alternative is to fit a pooled logistic model for the hazards $\Pr\left[D_{k+1}=1|\bar{A}_k=\bar{a}_k,\bar{L}_k=\bar{l}_k,D_k=C_{k+1}=0\right]$ in which each individual at time k receives the time-varying nonstabilized IP weight $W_k^{\bar{A}}\times W_k^{\bar{C}}$, where

$$W_k^{\bar{A}}=\prod_{m=0}^{k}\frac{1}{f\left(A_m|\bar{A}_{m-1},D_m=C_m=0,\bar{L}_m\right)},\quad W_k^{\bar{C}}=\prod_{m=1}^{k}\frac{1}{\Pr\left(C_m=0|\bar{A}_{m-1},D_{m-1}=C_{m-1}=0,\bar{L}_{m-1}\right)},$$

or its corresponding stabilized IP weight at each time k. The parameters of that model estimate the parameters of a marginal structural pooled logistic model for $\Pr\left[D_{k+1}^{\bar{a},\bar{c}=\bar{0}}=1|D_k^{\bar{a},\bar{c}=\bar{0}}=0\right]$ (Robins 1998a). For details and an application, see Hernán et al. (2001). Wen et al. (2022) review multiply robust estimators for survival analysis with time-varying treatments.

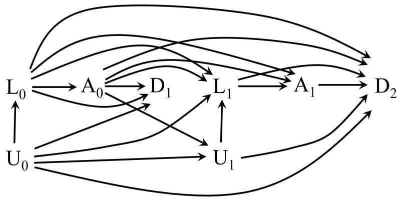


Figure 21.4

censored individuals in the pseudo-population is identical to that in the study population at each time k. The stabilized weights do not eliminate censoring in the pseudo-population, they make censoring occur at random at each time k with respect to the measured covariate history $\bar{L}_k$. That is, there is selection but no selection bias. Regardless of the type of IP weights used, in the pseudo-population there are no arrows from L_k and A_k into future C_m for $m>k$. Importantly, under the exchangeability conditions for the joint treatment $\left(\bar{A},\bar{C}\right)$, IP weighting can unbiasedly estimate the joint effect of $\left(\bar{A},\bar{C}\right)$ even when some components of $\bar{L}$ are affected by prior treatment.

Finally, when using g-estimation of structural nested models, we first need to adjust for selection bias due to censoring by IP weighting. In practice, this means that we first estimate nonstabilized IP weights $W^{\bar{C}}$for censoring to create a pseudo-population in which nobody is censored, and then apply g-estimation to the pseudo-population.

21.6 The big g-formula

This chapter and the previous two chapters privilege methods that rely on sequential exchangeability given the measured covariates $\overline{L}$ and identification by the g-formula. The reason is that, in practice, few causal analyses of complex longitudinal data have relied on other identifying conditions and formulas. For example, there are few realistic applications based on the identifying conditions under which the front door formula is the identifying formula. However, regardless of substantive plausibility and practical applications, different identifying conditions and their formulas are mathematically linked to sequential exchangeability and the g-formula based on all variables—both measured and unmeasured—as we now explain.

When sequential exchangeability holds given the measured covariates $\overline{L}$, we have discussed how the g-formula based on the measured time-varying covariates $\overline{L}$ identifies causal effects of a time-varying treatment $\overline{A}$ on an outcome Y. Now suppose we have a causal DAG with both observed variables $\left(\overline{A}, \overline{L}, Y\right)$ and unobserved variables $\overline{U}$, and that the measured variables $\overline{L}$ are insufficient to achieve sequential exchangeability.

For any causal DAG, the combination of measured and unmeasured variables $\overline{X} = (\overline{L}, \overline{U})$ ensures (joint) sequential exchangeability as any parent of a treatment variable is contained in either $\overline{A}$ or $\overline{X}$. Therefore, if every variable on a causal diagram were measured and positivity held, the g-formula based on $\overline{X}$ would identify the counterfactual mean $\mathrm{E}\left[Y^g\right]$ under any treatment strategy g. We refer to the g-formula with $\overline{L}$ replaced by $\overline{X}$ as the *big g-formula* because it is not based solely on the observed data.

We refer to $\left(\overline{A}, \overline{L}, Y, \overline{U}\right)$ as factuals to distinguish them from counterfactuals. Factuals are variables that exist in the actual world. In contrast to the observed variables, some factuals, such as $\overline{U}$, are not available for data analysis, often because they were not measured.

Given a causal DAG, treatment $\overline{A}$ and outcome Y, a treatment strategy g, and factuals $\left(\overline{A}, \overline{L}, Y, U\right)$, we can explicitly write down the big g-formula for the distribution (density) of Y^g. The big g-formula depends only on the distribution of the factuals $\left(\overline{A}, \overline{L}, Y, U\right)$.

The big g-formula is the right formula to identify the counterfactual density under any treatment strategy, but the big g-formula cannot be used in practice because it includes unmeasured variables. An interesting math question is then: can the big g-formula be reduced to a functional of the joint distribution of the observed data $\left(\overline{A}, \overline{L}, Y\right)$? If it can, then we will have a new formula that is not expressed as a g-formula but that (i) reproduces the results of the big g-formula (and therefore is a correct formula) and (ii) is written in terms of the distribution of the observed variables only (and therefore is a formula that can be used in data analyses).

For example, under the identifying conditions referred to as the front door criterion, the big g-formula for $\mathrm{E}[Y^a]$ reduces to a formula that only includes observed variables—the front door formula (see the proof in Technical Point 21.11). Therefore, the front door formula is a valid formula for the mean of $\mathrm{E}\left[Y^a\right]$ under the front door assumptions embedded in the causal diagram of Figure 7.14.

These questions were completely settled by the work of Tian and Pearl (2002), Shpitser and Pearl (2006), and Huang and Valtorta (2006).

More generally, we would like to be able to answer the following two questions. First, can we always determine whether the big g-formula can be rewritten as a formula that depends only on the distribution of the observed variables $\left(\overline{A}, \overline{L}, Y\right)$, while making no assumptions other than the joint distribution of $\left(\overline{A}, \overline{L}, Y, U\right)$ obeys the d-separation relations implied by the causal DAG? Second, when the answer to the previous question is yes, can we explicitly display such an identifying formula? Both of these questions have been answered in the affirmative.

Importantly, these are purely mathematical questions about properties of

Technical Point 21.11

A big g-formula proof of the front door formula. In Technical Point 7.4, we provided a proof of the front door formula for the counterfactual probability $\Pr\left[Y^{a}=y\right]$ under the causal diagram of Figure 7.14. Here we provide another proof using the big g-formula. This second proof relies on the conditional independencies implied by Figure 7.14, but it does not require that the counterfactuals Y^{m} exist.

The big g-formula for $\Pr\left[Y^{a}=y\right]$ under Figure 7.14 is

$$\sum_{m}\sum_{u}\Pr[Y=y|M=m,A=a,U=u]\Pr[M=m|A=a,U=u]\Pr[U=u].$$

Since data on U are not available, $\Pr\left[Y^{a}=y\right]$ is identified if and only if the big g-formula depends exclusively on the distribution of the observed data (Y,M,A). We now show that is indeed the case because, under the above assumptions, the g-formula reduces to the front door formula.

Using d-separation, we can rewrite the big g-formula as

$\sum_{m}\Pr[M=m|A=a]\sum_{u}\Pr[Y=y|M=m,U=u]\left\{\sum_{a'}\Pr[U=u|A=a']\Pr[A=a']\right\}$

by $U\perp\!\!\!\perp M|A$ and $A\perp\!\!\!\perp Y|M,U$

$=\sum_{m}\Pr[M=m|A=a]\sum_{a'}\left\{\sum_{u}\Pr[Y=y|M=m,A=a',U=u]\Pr[U=u|M=m,A=a']\right\}\Pr[A=a']$

by $U\perp\!\!\!\perp M|A$ and $A\perp\!\!\!\perp Y|M,U$

$=\sum_{m}\Pr[M=m|A=a]\sum_{a'}\Pr[Y=y|M=m,A=a']\Pr[A=a']$, which is the front door formula.

We now provide yet another proof of the front door formula that also does not require that the counterfactuals Y^{m} exist. After establishing that $\Pr\left[Y^{a}=y\right]$ is a function of the distribution of (Y,M,A,U) given by the big g-formula, we can apply a coupling argument. Suppose all agree on substantive grounds that a well-defined Y^{m} does not exist. Yet any factual data distribution that is Markov with respect to Figure 7.14 is compatible with an underlying FFRCISTG model "as detailed as the data"(Robins and Richardson, 2010) which, by definition, formally includes a variable Y^{m}. The proof in Technical Point 7.4 demonstrated that, under this model, the big g-formula equals the front door formula. It follows that there cannot exist a factual distribution Markov with respect to Figure 7.14 where this equality fails; for if it failed, that factual distribution would not be compatible with an FFRCISTG model "as detailed as the data".

Technical Point 21.12 presents an alternative proof of the front door formula based on a SWIG property.

distributions over $\left(\overline{A},\overline{L},Y,U\right)$ known to obey certain independence relations characterized by d-separation on the DAG. That is, these questions make no reference to either counterfactuals or to causality. The only connection to causality is the claim that the DAG is a causal DAG. If so, the big g-formula will have a causal interpretation. If not, the affirmative answers, though still true, will have no causal meaning. Of course, in observational analyses, we can never know with certainty that a graph that we conjecture to be a causal diagram is indeed a causal diagram.

Technical Point 21.12

A front door formula proof using d-separation of treatment nodes on SWIGs. Here we provide another proof of the front door formula using an important property of SWIGs that we have yet to discuss.

Given a causal diagram G, let $G^{\overline{a}}$ be the associated SWIG for strategy $\overline{a}$, and $B^{\overline{a}}$ and $C^{\overline{a}}$ two disjoint subsets of the observed non-treatment nodes $\left(Y^{\overline{a}}, \overline{L}^{\overline{a}}\right)$. We assume only treatment counterfactuals are well-defined. The SWIG $G^{\overline{a}}$ satisfies the following property (Shpitser et al., 2022): If the fixed node a_m is d-separated from $B^{\overline{a}}$ conditional on $C^{\overline{a}}$, then $\Pr\left(B^{\overline{a}} = b | C^{\overline{a}} = c\right)$ does not depend on a_m. This property does not conflict with the previously discussed fact that any path that contains a treatment a_m as a non-endpoint is blocked. To make clear what the new property means, consider the SWIG G^a implied by the front door diagram in Figure 7.14. On SWIG G^a, define $B^a = Y^a$ and $C^a = \left(M^{a'}, A\right)$. Then a is d-separated from B^a given C^a as the only path from a to Y^a goes through the non-collider $M^{a'}$ in C^a. Thus, according to our property $\mathrm{E}\left[Y^a | M^a, A\right] = \mathrm{E}\left[Y^{a'} | M^{a'}, A\right]$ for any a and a'. Note the property is not cross-world; rather, it specifies a relationship between different single-world counterfactual distributions.

We now use this SWIG property to prove the front door formula when well-defined counterfactuals Y^m do not exist. We continue to assume that (Y^a, M^a, A) factor according to the SWIG G^a and $\left(Y^{a'}, M^{a'}, A\right)$ factor according to the SWIG $G^{a'}$. We follow the proof in Technical Point 7.4 until we come to the point where we must prove

$$\mathrm{E}\left[Y^a | M^a\right] = \sum_{a'} \mathrm{E}\left[Y | M, A = a'\right] \Pr\left(A = a'\right).$$

We now have $\mathrm{E}\left[Y^a | M^a\right] = \sum_{a'} \mathrm{E}\left[Y^a | M^a, A = a'\right] \Pr\left(A = a' | M^a\right) = \sum_{a'} \mathrm{E}\left[Y^a | M^a, A = a'\right] \Pr\left(A = a'\right)$ by $M(a)$ d-separated from A. Our new SWIG property implies that $\mathrm{E}\left[Y^a | M^a, A = a'\right] = \mathrm{E}\left[Y^{a'} | M^{a'}, A = a'\right] = \mathrm{E}\left[Y | M, A = a'\right]$ where the last equality is by consistency. Thus, $\mathrm{E}\left[Y^a | M^a\right] = \sum_{a'} \mathrm{E}\left[Y | M, A = a'\right] \Pr\left(A = a'\right)$ as required. Interestingly, it follows that, although $\mathrm{E}\left[Y^a | M^a\right] = \mathrm{E}\left[Y^{a'} | M^{a'}\right]$ for all a, a', nonetheless $\mathrm{E}\left[Y^a | M^a\right] \neq \mathrm{E}\left[Y | M\right]$ because $\mathrm{E}\left[Y | M\right] = \sum_{a'} \mathrm{E}\left[Y | M, A = a'\right] \Pr\left(A = a' | M\right)$ and, unlike the counterfactual M^a, the observed factual $M = M^A$ is not independent of A.

Technical Point 21.13

Formal definition of a general structural nested mean model. Robins (2004) noted there is nothing special about $\bar{0}$ as the strategy that is followed after a final blip of treatment in a structural nested mean model (SNMM). We can instead define the blip functions relative to an arbitrary strategy g as follows. Given $g = (g_0, g_1, ..., g_K)$, an additive SNMM is a model for the causal effect on Y (conditional on treatment and covariate history through time t) of a blip a_t of treatment at t and then following g from time $t+1$ onward versus following g from time t onward. That is, an additive SNMM models the counterfactual contrast

$$\gamma_t^g\left(\overline{a}_t, \bar{l}_t\right) = \mathrm{E}[Y^{\overline{a}_{t-1}, a_t, \underline{g}_{t+1}} - Y^{\overline{a}_{t-1}, g_t, \underline{g}_{t+1}} \,|\, \overline{A}_{t-1} = \overline{a}_{t-1}, A_t = a_t, \overline{L}_t = \bar{l}_t]$$

for $t = 0, ..., K$ with $\overline{a} = (a_0, a_1, ..., a_K)$, $\underline{g}_{t+1} = (g_{t+1}, ..., g_K)$. We write $\gamma_t^g\left(\overline{a}_t, \bar{l}_t\right)$ as $\gamma_t^g\left(\overline{a}_{t-1}, a_t, \overline{l_t}\right)$ and $Y^{\overline{a}_{t-1}, \underline{g}_t}$ as $Y^{\overline{a}_{t-1}, g_t, \underline{g}_{t+1}}$ when we want to emphasize the unique role of a_t and g_t. Note that $\gamma_t^g\left(\overline{a}_{t-1}, a_t, \overline{l_t}\right) \equiv 0$ when $a_t = g_t\left(\overline{a}_{t-1}, \overline{l_t}\right)$. If, as in the main text, we assume sequential exchangeability, then $A_t = a_t$ can be dropped from the conditioning event in the definition of $\gamma_t^g\left(\overline{a}_t, \bar{l}_t\right)$.

An SNMM assumes $\gamma_t^g\left(\overline{a}_t, \bar{l}_t\right) = \gamma_t^g\left(\overline{a}_t, \bar{l}_t; \beta\right)$ where $\gamma_t^g\left(\overline{a}_t, \bar{l}_t; \beta^\dagger\right)$ is a known function taking the value 0 if the finite-dimensional parameter vector $\beta^\dagger$ equals 0 or $a_t = g_t\left(\overline{a}_{t-1}, \overline{l_t}\right)$. If we define

$$H_k\left(\gamma^g\right) = Y - \sum_{t=k}^{K} \gamma_t^g\left(\overline{A}_t, \overline{L}_t\right),$$

it follows from consistency alone (Robins 2004) that $\mathrm{E}[H_k\left(\gamma^g\right)|\overline{L}_k, \overline{A}_k] = \mathrm{E}[Y^{\overline{A}_{k-1}, \underline{g}_k}|\overline{L}_k, \overline{A}_k]$ for $k = 0, ..., K$ and $\mathrm{E}[H_0\left(\gamma^g\right)] = \mathrm{E}[Y^g]$. Therefore, if we can identify the $\gamma_t^g\left(\overline{a}_t, \bar{l}_t\right)$, we can identify $\mathrm{E}[Y^{\overline{A}_{k-1}, \underline{g}_k}|\overline{L}_k, \overline{A}_k]$ and $\mathrm{E}[Y^g]$. Under positivity and sequential exchangeability, the last set-off equation implies $\mathrm{E}[H_k\left(\gamma^g\right)|\overline{L}_k, \overline{A}_k] = \mathrm{E}[H_k\left(\gamma^g\right)|\overline{L}_k, \overline{A}_{k-1}]$ which implies the $\gamma_t^g\left(\overline{a}_t, \bar{l}_t\right)$ are nonparametrically identified. Robins (2004) also defined an optimal regime structural nested model (opt-SNMM) and showed how, under positivity and sequential exchangeability, one can use the opt-SNMM to estimate the optimal treatment strategy $g_{opt} = \arg\max_g[\mathrm{E}\left(Y_g\right)]$.

But sequential exchangeability is not the only possible identifying assumption. For example, Zahn et al. (2022) showed that the $\gamma_t^g\left(\overline{a}_t, \bar{l}_t\right)$ are identified under a time-varying parallel trends assumption that generalizes the identifying assumption typically made for difference-in-differences estimation with time-varying treatments and covariates.

In Technical Point 21.9, we took g in the SNMM to be the strategy "never treat", i.e., $g = \bar{0}$, and we described an algorithm to identify $\mathrm{E}[Y^g]$ for every strategy g under the assumption of sequential exchangeability. When sequential exchangeability does not hold, we can use other assumptions (e.g., time-varying parallel trends) that suffice to identify $\mathrm{E}[Y^g]$ for the g used to define the SNMM, but not to identify $\mathrm{E}[Y^{g'}]$ for any other strategy g'. To do so, we need additional assumptions. For example Shahn et al. (2022) showed that if, in addition to assuming time-varying parallel trends, one assumes that, conditional on past treatment and measured covariate history, there is no additive effect modification by unmeasured confounders U, then $\mathrm{E}[Y^{g'}]$ is identified for all g'. This implies that the optimal strategy $g_{opt} = \mathrm{argmax}_{g'}\,\mathrm{E}[Y^{g'}]$ is identified. Shahn et al. (2022) show how one can use structural nested mean models to estimate g_{opt}.