"Tutorial For Causal Inference" - Chapter 20 of Handbook of Big Data

Laura Balzer, University of California, Berkeley
Maya Petersen, University of California, Berkeley
Mark J. van der Laan, University of California, Berkeley

Available at: https://works.bepress.com/laura_balzer/40/
20

Tutorial for Causal Inference

Laura Balzer, Maya Petersen, and Mark van der Laan

CONTENTS

20.1 Why Bother with Causal Inference? ........................................... 361
20.2 The Scientific Question .......................................................... 363
20.3 The Causal Model ................................................................. 363
20.4 The Target Causal Quantity .................................................. 366
20.5 The Observed Data and Their Link to the Causal Model ............ 368
20.6 Assessment of Identifiability .................................................. 369
20.7 Estimation and Inference ....................................................... 373
20.8 Interpretation of the Results .................................................. 376
20.9 Conclusion ............................................................................ 377
Appendix: Extensions to Multiple Time Point Interventions ........... 378
Acknowledgments ........................................................................ 381
References .................................................................................... 381

20.1 Why Bother with Causal Inference?

This book has mostly been dedicated to large-scale computing and machine learning algorithms. These tools help us describe the relationships between variables in vast, complex datasets. This chapter goes one step further by introducing methods, as well as their limitations, to learn causal relationships from these data. Consider, for example, the following questions:

1. What proportion of patients taking drug $X$ suffered adverse side effects?
2. Which patients taking drug $X$ are more likely to suffer adverse side effects?
3. Would the risk of adverse effects be lower if all patients took drug $X$ instead of drug $Y$?

The first question is purely descriptive; the second can be characterized as a prediction problem, whereas the last is causal. Causal inference is distinct from statistical inference in that it seeks to make conclusions about the world under changed conditions [1]. In the third example, our goal is to make inferences about how the distribution of patient outcomes would differ if all patients had taken drug $X$ versus if the same patients, over the same time frame and under the same conditions, had taken drug $Y$. Purely statistical analyses are sometimes endowed with causal interpretations. Furthermore, many of our noncausal questions have causal elements. For example, Geng et al. [2] sought to assess whether sex was
an independent predictor of mortality among patients initiating drug therapy (i.e., describe a noncausal association) but in the absence of loss to follow up (i.e., a change to the existing conditions).

In this chapter, we review a formal framework for causal inference to (1) state the scientific question; (2) express our causal knowledge and limits of that knowledge; (3) specify the causal parameter; (4) specify the observed data and their link to the causal model; (5) assess identifiability of our causal parameter as some function of the observed data distribution; (6) estimate the corresponding statistical parameter, incorporating methods discussed in this book; and (7) interpret our results [3–5]. Access to millions of data points does not obviate the need for this framework. Analyses of big data are not immune to the problems of small data. Instead, one might argue that analyses of big data exacerbate many of the problems of small data. As illustrated in Figure 20.1, there are many sources of association between two variables, including direct effects, indirect effects, measured confounding, unmeasured confounding, and selection bias [6]. Methods to delineate causation from correlation are perhaps more pressing now than ever [7,8].

**FIGURE 20.1**
Some of the sources of dependence between an exposure $A$ and an outcome $Y$: (a) the exposure $A$ directly affects the outcome $Y$; (b) the exposure $A$ directly affects the outcome $Y$ as well as indirectly affects it through the mediator $Z$; (c) the exposure $A$ has no effect on the outcome $Y$, but an association is induced by a measured common cause $W$; (d) the exposure $A$ has no effect on the outcome $Y$, but an association is induced by an unmeasured common cause $U$; (e) the exposure $A$ has no effect on the outcome $Y$, but an association is induced by only examining data among those not censored $C$; (f) all these sources of dependence are present. Please note this is not an exhaustive list.
20.2 The Scientific Question

The first step in the causal roadmap is to specify the scientific objective. As a running example, we will consider the timing of antiretroviral therapy (ART) initiation and its impact on outcomes among HIV+ individuals. Early ART initiation has been shown to improve patient outcomes as well as reduce transmission between discordant couples [9,10]. Suppose we want to learn the effect of immediate ART initiation (i.e., irrespective of CD4+ T-cell count) on mortality. Large consortiums, such as the International Epidemiologic Databases to Evaluate AIDS and Sustainable East Africa Research in Community Health, are providing unprecedented quantities of data to answer this and other questions [12,13].

To sharply frame our scientific aim, we need to further specify the system, including the target population (e.g., patients and context), the exposure (e.g., criteria and timing), and the outcome. As a second try, consider our goal as learning the impact of initiating ART within 1 month of diagnosis on 5-year all-cause mortality among adults, recently diagnosed with HIV in sub-Saharan Africa. This might seem like an insurmountable task, and it may seem safer to frame our question in terms of an association. Indeed, there seems to be a tendency to shy away from causal language when stating the scientific objective. However, we are not fundamentally interested in the correlation between early ART initiation and mortality among HIV+ adults. Instead, we want to isolate the effect of interest from the spurious sources of dependence (e.g., confounding, selection bias, informative censoring) as shown in Figure 20.1. The framework, discussed in this chapter, provides a pathway from our scientific aim to estimation of a statistical parameter that best approximates our causal effect, while keeping any assumptions transparent.

20.3 The Causal Model

The second step of the roadmap is to specify our causal model. Causal inference is distinct from statistics in that it requires something more than a sample from the observed data distribution. In particular, causal inference requires specification of background knowledge, and causal models provide a rigorous language for expressing this knowledge and its limits. In this chapter, we focus on structural causal models [14] to formally represent which variables potentially affect one another, the roles of unmeasured factors, and the functional form of those relationships. Structural causal models unify causal graphs [15], structural equations [16,17], and counterfactuals. We also briefly introduce the Neyman–Rubin potential outcomes framework [18–20] and discuss its relation to the structural causal model.

Consider again our running example. Let $W$ denote the set of baseline covariates, including sociodemographics, clinical measurements, and social constructs. The exposure $A$ is an indicator, equalling 1 if the patient initiated ART within 1 month of diagnosis and equalling 0 otherwise (i.e., initiation took longer than 1 month). Finally, the outcome $Y$ is an indicator that the patient did not survive 5 years of follow-up. These factors have scientific meaning to the question and comprise the set of endogenous variables: $X = \{W, A, Y\}$. They can be measurable (e.g., age and sex) or unmeasurable and are affected by other variables in the model.

Each endogenous variable is associated with a set of background factors $U = (U_W, U_A, U_Y)$ with some joint distribution $P_U$. These represent all the unmeasured factors, affecting other variables in the model but not included in $X$. For example, $U_A$ could include unknown clinic-level factors, influencing whether or not a patient initiates early ART.
Likewise, $U_Y$ may include a patient’s genetic risk profile. Furthermore, there might be shared unmeasured causes between the endogenous variables. For example, socioeconomic status may impact whether a patient initiates early ART as well as his/her 5-year mortality.

Each endogenous variable is also associated with a structural equation. These functions help encode our causal knowledge. Suppose, for example, we believe that the set of baseline covariates possibly impact whether a patient initiates early ART, and that both the covariates and the exposure may affect subsequent morality. Then we write each endogenous variable as a deterministic function of its parents, variables that may impact its value:

$$W = f_W(U_W)$$
$$A = f_A(W, U_A)$$
$$Y = f_Y(W, A, U_Y).$$

(20.1)

These functions $F = \{f_W, f_A, f_Y\}$ are left unspecified (nonparametric). For example, the third equation $f_Y$ encodes that the covariates $W$ and the exposure $A$ may have influenced the value taken by the outcome $Y$. We have not, however, restricted their relationships: $A$ and any member of $W$ may interact on an additive (or any other) scale to affect $Y$ and the impacts of $A$ and $W$ on $Y$ may be nonlinear.

The structural causal model, denoted $M^F$, is defined by all possible distributions of $P_U$ and all possible sets of functions $F$, which are compatible with our assumptions (if any). For the above example, there is some true joint distribution $P_{U,0}$ of health care access, personal preferences for ART use, socioeconomic factors, etc. Randomly sampling a patient from the population corresponds to drawing a particular realization $u$ from $P_{U,0}$. Likewise, there are some true structural equations $F_0$ that would deterministically generate the endogenous variables $X = x$ if given input $U = u$. For a given distribution $P_U$ and set of functions $F$, the structural causal model $M^F$ describes the following data generating process for $(U, X)$:

1. Drawing the background factors $U$ from some joint probability distribution $P_U$
2. Generating the baseline covariates $W$ as some deterministic function $f_W$ of $U_W$
3. Generating the exposure $A$ as some deterministic function $f_A$ of covariates $W$ and $U_A$
4. Generating the outcome $Y$ as some deterministic function $f_Y$ of covariates $W$, the exposure $A$, and $U_Y$

Thus, the model $M^F$ is the collection of all possible probability distributions $P_{U,X}$ for the exogenous and endogenous variables $(U, X)$. The true joint distribution is an element of the causal model: $P_{U,X,0} \in M^F$. The structural causal model is also sometimes also called a nonparametric structural equation model [14,21].

In other settings, we may have more in-depth knowledge about the data generating process. This knowledge is generally encoded in two ways. First, excluding a variable from the parent set of $X_i$ encodes that this variable does not directly impact the value $X_i$ takes. These assumptions are known as exclusion restrictions. Second, restricting the set of allowed distributions for $P_U$ encodes that some variables do not have any unmeasured common causes. These assumptions are known as independence assumptions. Suppose, for example, that patients were randomized $R$ to early ART initiation, but adherence $A$ was imperfect. Then the treatment assignment $R$ would only be determined by chance (e.g., a coin flip) and not influenced by baseline covariates $W$. The unmeasured factors determining treatment assignment would be independent from all other unmeasured factors:

$$U_{R \perp \!\!\!\!\!\!\!\!\!\perp} (U_W, U_A, U_Y).$$
This is an independence assumption that restricts the allowed distribution of background factors $P_U$. Furthermore, suppose that randomization $R$ only affects the mortality $Y$ through its effect on adherence $A$. The resulting structural equations are then

$$
\begin{align*}
W &= f_W(U_W) \\
R &= f_R(U_R) \\
A &= f_A(W, R, U_A) \\
Y &= f_Y(W, A, U_Y).
\end{align*}
$$

(20.2)

We have made two exclusion restrictions: (1) the baseline covariates $W$ do not influence randomization $R$ and (2) randomization $R$ has no direct effect on the outcome $Y$. The structural causal model is then defined by all probability distributions for $U$ that are compatible with our independence assumptions and all sets of functions $F = (f_W, f_R, f_A, f_Y)$ that are compatible with our exclusion restrictions.

A causal graph can be drawn from the structural causal model [14]. Each endogenous variable (node) is connected to its parents and background error term with a directed arrow. The potential dependence between the background factors is encoded by the inclusion of a node representing any unmeasured common cause. Exclusion restrictions are encoded by absence of a directed arrow. Likewise, independence assumptions are encoded with the absence of a node representing an unmeasured common cause. The corresponding causal graphs for the two examples are given in Figure 20.2.

**FIGURE 20.2**
Directed acyclic graphs representing the structural causal model for our study (Equation 20.1) and for the hypothetical randomized trial (Equation 20.2). (a) This graph only encodes the time ordering between baseline covariates $W$, the exposure $A$, and the outcome $Y$. A single node $U$ represents the unmeasured common causes of the endogenous variables. (b) This graph encodes the randomization $R$ of some treatment with incomplete adherence $A$. There are two exclusion restrictions: The baseline covariates $W$ do not impact the randomization $R$, and the randomization $R$ has no direct effect on the outcome $Y$. There is also an independence assumption: the unmeasured factors contributing to randomization are independent of the unmeasured factors, contributing to the other variables.
Common Pitfall: Oversimplifying the Causal Model

Structural causal models and their corresponding graphs are powerful precisely because they do not impose unsubstantiated assumptions on the data generating process. There is a tendency, however, to present oversimplified models and graphs. It is crucial to remember that these are formal models, and every exclusion restriction or independence assumption (or equivalently, every arrow omitted) represents a real assumption about the true data generating system. Often, our knowledge is limited to the causal ordering of the variables in our system. Sometimes, we might not even have this information, forcing us to represent our knowledge using more than one possible model and graph.

20.4 The Target Causal Quantity

The structural causal model $\mathcal{M}^F$ describes the system not only as it currently exists but also as it would exist under changed conditions. The structural equations are autonomous: an intervention on one equation does not affect the remaining ones. Therefore, we can modify a function and see how changes are transmitted through the system. For example, modifying the treatment decision does not change the effect of the treatment on the outcome. Therefore, we can make a targeted modification to represent our intervention of interest. In our running example (Equation 20.1 and Figure 20.2a), a self-selected group of patients initiated early ART. To answer our scientific question, we need to modify how this exposure variable was generated. Specifically, we can intervene to start all patients on ART within 1 month of testing HIV+ (i.e., deterministically set $A = 1$), and we can intervene to delay all patients from starting ART until 1 month after testing HIV+ (i.e., deterministically set $A = 0$):

$$
W = f_W(U_W) \\
A = 1 \\
Y_1 = f_Y(W, 1, U_Y)
$$

$$
W = f_W(U_W) \\
A = 0 \\
Y_0 = f_Y(W, 0, U_Y)
$$

Alternative exposure mechanisms include dynamic interventions [22–25], which are responsive to patient characteristics, and stochastic interventions* [26], which are nondeterministic.

The counterfactual outcome $Y_a$ is then the outcome a patient would have had, if possibly contrary to fact, he or she had received exposure level $A = a$. More formally, $Y_a = Y_a(u)$ is defined as the solution to the equation $f_y$ under an intervention to set $A = a$ (with input $U = u$). Therefore, $Y_a(U)$ is a postintervention random variable, whose probability distribution is induced by the set of structural equations $F$ and the joint distribution of the background factors $P_Y$. In other words, the structural causal model $\mathcal{M}^F$ is also a model on the distribution of counterfactuals. In the Neyman–Rubin causal framework, these quantities are known as potential outcomes [18,19,27]. They are assumed to exist for all units under

*For simplicity, we have been considering the time-scale to be in months. Depending on our scientific question and the data resolution, we might be interested in shorter or longer intervals. If our time interval were days, then an intervention to start by day 30 (i.e., within 1 month) is a stochastic intervention. Alternatively, we could consider an intervention to initiate therapy on each day or not. For further discussion of longitudinal treatment regimes, see Appendix.
all treatment levels of interest. For this example, the *full data* would consist of baseline covariates and the outcomes under all possible exposures: $X^F = (W, (Y_a : a \in \{0, 1\}))$. The structural causal model $M^F$ also serves as a model for the set of possible full data distributions, each corresponding to a different intervention on the endogenous variables.

The distribution of these counterfactuals (potential outcomes) can then be used to define the target causal parameter. Consider, for example, the average treatment effect:

$$
\Psi^F(P_{U,X}) = E_{U,X}(Y_1) - E_{U,X}(Y_0),
$$

where the subscript $(U, X)$ denotes the expectation over the distribution $P_{U,X}$ (which implies the distribution of the counterfactual random variables $(Y_1, Y_0)$). In words, $\Psi^F(P_{U,X})$ is the difference in the expected counterfactual outcome if everyone in the population were exposed and the expected counterfactual outcome if everyone in the population were not exposed. Formally, $\Psi^F$ is a mapping from a distribution $P_{U,X}$ in the causal model $M^F$ to the real number line. For our example, $\Psi^F(P_{U,X})$ is the difference in the counterfactual risk of mortality if all patients immediately initiated ART and if all patients delayed ART initiation. For a binary outcome, this causal quantity corresponds the causal risk difference.

We could also specify this contrast on the relative scale, within a certain stratum of the population (e.g., those with baseline CD4 counts above 350 cells/mm$^3$), for the actual study units (i.e., the sample average treatment effect [18]) or for some other population (i.e., transportability [28–30]).

*Marginal structural models* provide an alternative way to define our target parameter [31]. They are a summary measure of how the counterfactual outcome changes as a function of the exposure and possibly pretreatment covariates. Consider, for example, the impact of reducing the time (in months) between HIV diagnosis and treatment initiation. The intervention variable $A$ would then be continuous. (An alternative approach would be to treat the exposure as a time-dependent binary variable as discussed in Appendix.)

To generate the relevant counterfactual outcomes, we would repeatedly intervene on the structural causal model to set $A = a$ for all levels of $a$ in the exposure set of interest $A = \{1, 2, 3, \ldots\}$. If we knew the true shape of the relationship between the expected counterfactual outcome $E_{U,X}(Y_a)$ and the treatment level $a$, we could summarize it with a parametric model [31], such as the following:

$$
\logit[E_{U,X}(Y_a)] = m(a|\beta) \quad \text{with } m(a|\beta) = \beta_0 + \beta_1 a.
$$

This model assumes that the counterfactual mortality risk is a function linear on the logistic scale of time to treatment initiation $a$. This marginal structural model restricts the set of possible counterfactual distributions and therefore places assumption on our causal model $M^F$.

In many cases, we do not have sufficient information to confidently specify a parametric model for this dose–response curve. Instead, we can use a *working marginal structural model* as a summary of the causal relationship of interest [32]. The target causal parameter is then the projection of the true causal curve onto a working model. Consider, for example,

$$
\beta(P_{U,X}|m) = \arg\min_{\beta} E_{U,X} \left[ \sum_{a \in A} - \log [m(a|\beta)^{Y_a} (1 - m(a|\beta))^{(1-Y_a)}] \right].
$$

*Under the Neyman–Rubin framework, we would assume the existence of the potential outcomes $Y_a$ for all exposures $a \in A$.*
where our projection is the negative log-likelihood loss. Intuitively, we can think of this projection as summarizing the full data (i.e., all counterfactuals) with a parametric regression curve. As usual, the quality of the summary depends on the underlying causal curve and the question of interest.

20.5 The Observed Data and Their Link to the Causal Model

Thus far, we have not specified the data that will be or have been collected in our study. Instead, we have discussed endogenous variables $X$ (observable and possibly unobservable), background factors $U$ (unobservable), and set of counterfactuals $(Y_a : a \in \mathcal{A})$. In this step, we specify the observed data, their link to the causal model and the resulting statistical model.

Suppose we have a simple random sample of $n$ patients from our target population. On each patient, we measure some baseline covariates $W$, including sex, age, and CD4 count, the exposure $A$ (whether or not the patient initiates ART within 1 month of diagnosis), and the outcome $Y$ as the patient’s 5-year mortality. Then the observed data for a given patient are $O = (W, A, Y)$, which has some true, but unknown distribution $P_0$.

We assume that the observed data are generated by sampling $n$ times from a distribution compatible with (contained in) the structural causal model. Recall the structural causal model provides a description of the data generating system under existing conditions as well as under specific interventions. The distribution of the background factors $P_U$ and the structural equations $F$ identify the distribution of the endogenous variables $X$ as well as the distribution of the observed data $O$. The observed data $O$ are a subset of $(U, X)$. Suppose, for example, we observe all the endogenous nodes (i.e., if $O = X$). Then we have

$$P(O = o) = \sum_u P_{U,X}(X = x | U = u)P_U(U = u) = \sum_u I(X(u) = x)P_U(U = u),$$

where the summation generalizes to an integral for continuous valued variables. This framework naturally accommodates more complicated links, such as case–control sampling and matched sampling [33,34].

Thereby, the structural causal model $\mathcal{M}_F$, which is the set of possible distributions for $(U, X)$, implies our statistical model $\mathcal{M}$, which is the set of possible distributions for the observed data $O$. The true distribution of the observed data $P_0$ is implied by the true distribution $P_{U,X,0}$ of $(U, X)$ and is an element of the statistical model: $P_0 \in \mathcal{M}$. The causal model may, but often does not, place any restrictions on the statistical model. For example, the causal model, describing the data generating process for our observational study (Figure 20.2a), implies a nonparametric statistical model. There are no restrictions on the possible observed data distributions. In contrast, the causal model, corresponding to the randomized trial (Figure 20.2b), will only generate distributions, where the randomization $R$ is independent of the baseline covariates $W$. This is a testable assumption and implies a semiparametric statistical model. We refer the reader to Pearl [14,15] for further discussion of a graphical criterion to evaluate independence between two variables as implied by a structural causal model or its corresponding directed acyclic graph.

Suppose that instead of specifying a structural causal model, we chose to follow the Neyman–Rubin framework. Specifically, we assumed the existence of the potential outcomes
Tutorial for Causal Inference

$Y_a : a \in A$ in Step 3. To relate these potential outcomes to the observed data, we need the stable unit treatment value assumption \cite{35}. First, the potential outcomes for one unit must not be impacted by the treatment assignment of another unit (i.e., no interference)*. Second, there must not be multiple versions of the treatment $A = a$. With this assumption, we can map the potential outcomes to the observed outcomes:

$$Y_i = A_i Y_{1,i} + (1 - A_i) Y_{0,i}.$$  

For unit $i$, we only get to see the outcome $Y_i$, corresponding to the unit’s observed exposure $A_i$. As a result, causal inference can be treated as a missing data problem.

Common Pitfall: Specifying a Statistical Model Based on Convenience

Sometimes researchers specify a parametric multivariable model to relate the conditional mean of the observed outcome to the observed exposure and baseline covariates. We could, for example, assume that a main terms logistic regression describes the relationship between observed mortality risk, early ART initiation, and the measured covariates. Although these parametric models are often recognized as being misspecified, estimation and inference proceed as if they were true. Formally, the statistical model is the set of possible distributions for the observed data and should reflect real knowledge, however limited. Structural causal models make explicit the implications for background knowledge on the observed data distribution. In many cases, background knowledge is not sufficient to place any restrictions on the distribution of the observed data. Thereby, use of a formal causal model highlights that in many practical data applications, a nonparametric statistical model is appropriate.

20.6 Assessment of Identifiability

In Step 3 (Section 20.4), we specified our scientific question as a causal parameter $\Psi^\mathcal{F}(P_{U,X})$, a function of the distribution of counterfactuals (potential outcomes). In Step 4 (Section 20.5), we specified the observed data $O$ and the statistical model $\mathcal{M}$. In this step, we establish whether our causal parameter can be written as some function of the observed data distribution. More formally, for each $P_{U,X}$ compatible with the structural causal model $\mathcal{M}^\mathcal{F}$, we want to establish the equivalence between the causal parameter $\Psi^\mathcal{F}(P_{U,X})$ and the statistical parameter $\Psi(P)$. If so, we state that the causal parameter is identified. If not, we explicitly state the additional assumptions needed to make inferences about the causal parameter using the observed data distribution. We keep these convenience-based assumptions separate from our knowledge-based assumptions, reflected in the structural causal model $\mathcal{M}^\mathcal{F}$.

Consider a simplified example, where we want to learn the 5-year mortality risk if, possibly contrary to the fact, all HIV+ adults initiated ART within 1 month of diagnosis:

*The structural causal model, given in Equation 20.1, implicitly assumes independence between study units. Recent work relaxing this assumption and considering a network of interacting units is given in the work by van der Laan [36].
Suppose we have not collected any baseline covariates; therefore, the observed data are simply \( O = (A, Y) \). Then the causal parameter will only equal the observed mortality risk among exposed if the only source of association is due to the effect of interest:

\[
P(Y = y | A = 1) = P_{U,X}(Y_1 = y | A = 1) \]

The first equality is by the definition of counterfactuals and then second holds if the counterfactual outcome \( Y_a \) is independent of the exposure \( A \). In the absence of baseline covariates, the outcomes is only a function of the exposure and the background error: \( Y = f_Y(A, U_Y) \). Once we intervene to set \( A = a \), the counterfactual outcome is only a function of its error: \( Y_a(U) = f_Y(a, U_Y) \). If the unmeasured factors contributing to the outcome \( U_Y \) are independent of those contributing the exposure \( U_A \), then the randomization assumption holds \( Y_a \perp A \), and the counterfactual risk \( P_{U,X}(Y_1 = 1) \) is identified as the observed risk among those exposed \( P(Y = 1 | A = a) \). The randomization assumption is equivalent to stating that there are no unmeasured confounders of the exposure–outcome relation. Intuitively, this assumption holds by design a randomized trial.

In most observational settings, the assumption of no common (measured or unmeasured) causes of the exposure and outcome will not hold. We can weaken the randomization assumption by conditioning on a set of measured baseline covariates: \( Y_a \perp A | W \). The adjustment set \( W \) needs to block all spurious sources of association without creating any new sources of dependence or blocking any of the effect of \( A \) on \( Y \). As illustrated in Figure 20.3, the back-door criterion can aid the evaluation of the randomization assumption [14]. A set of variables \( W \) satisfies the back-door criterion for the relationship of \((A, Y)\) if (1) no node in \( W \) is a descendant of \( A \) and (2) \( W \) blocks all back-door paths from \( A \) to \( Y \), where back door refers to a path with an arrow into \( A \). The rationale for condition 1 is to avoid blocking the path of interest or introducing spurious

![Figure 20.3](image-url)

**FIGURE 20.3**

Considering the back-door criterion for the basic structure. For all the graphs, the exposure \( A \) and the outcome \( Y \) do not share an unmeasured common cause. (a) The covariates \( W \) are not sufficient to block all back-door paths. Conditioning covariates \( W \) blocks the path \( Y \rightarrow W \rightarrow A \). However, conditioning on \( W \) (a collider of \( U \) and \( U^* \)) opens a new path: \( Y \rightarrow U^* \rightarrow U \rightarrow A \). (b) The covariates \( W \) and the outcome \( Y \) also do not share an unmeasured common cause. The covariates \( W \) are sufficient to block all back-door paths. (c) The exposure \( A \) and the covariates \( W \) also do not share an unmeasured common cause. The covariates \( W \) are sufficient to block all back-door paths. (d) All the unmeasured background factors are independent. The covariates \( W \) are sufficient to block all back-door paths.
associations (i.e., conditioning on a collider). The rationale for condition 2 is to block any remaining spurious sources of association. For the basic structure (Figure 20.3), the randomization assumption will hold if the following independence assumptions are true:

\[ U_A \perp \!
\perp U_Y \text{ and } U_A \perp \!
\perp U_W \text{ or } U_Y \perp \!
\perp U_W. \]

There must not be any unmeasured common causes of the exposure and the outcome, and of the exposure and covariates or of the outcome and covariates. As illustrated in Figure 20.4, this graphical criteria can aid in the selection of an appropriate adjustment set.

When the randomization assumption holds, we can identify the distribution of counterfactuals within strata of covariates. Specifically, we have that for each \( P_{U,X} \in \mathcal{M}^{F} \)

\[ P_{U,X}(Y_a = y|W = w) = P_{U,X}(Y_a = y|A = a, W = w) = P(Y = y|A = a, W = w), \]

where the distribution \( P \) of the observed data is implied by \( P_{U,X} \). This gives us the G-computation identifiability result [27] for the true distributions \( P_{U,X,0} \) and \( P_0 \):

\[ E_{U,X,0}(Y_a) = \sum_{w} E_0(Y|A = a, W = w)P_0(W = w), \]

where the summation generalizes to an integral for continuous covariates. Likewise, we can identify the difference in the expected counterfactual outcomes (i.e., the average treatment effect) in terms of the difference in the conditional mean outcomes, averaged with respect to the covariate distribution:

\[ E_{U,X,0}(Y_1 - Y_0) = \sum_{w} \left[ E_0(Y|A = 1, W = w) - E_0(Y|A = 0, W = w) \right] P_0(W = w). \]

Identifiability also relies on having sufficient support in the data. The G-computation formula requires that the conditional mean \( E_0(Y|A = a, W = w) \) is well defined for all possible values of \( w \) and levels of \( a \) of interest. In a nonparametric statistical model, each exposure of interest must occur with some positive probability for each possible covariate stratum:

\[ \min_{a \in A} P_0(A = a|W = w) > 0, \text{ for all } w \text{ for which } P_0(W = w) > 0. \]

This condition is known as the positivity assumption and as the experimental treatment assignment assumption.

Suppose, for example, that the randomization assumption holds conditionally on a single binary baseline covariate. Then our statistical estimand could be rewritten as

\[ \Psi(P_0) = \left[ E_0(Y|A = 1, W = 1) - E_0(Y|A = 0, W = 1) \right] P_0(W = 1) + \left[ E_0(Y|A = 1, W = 0) - E_0(Y|A = 0, W = 0) \right] P_0(W = 0). \]

As an extreme, suppose that in the population, there are zero exposed patients with this covariate: \( P_0(A = 1|W = 1) = 0 \). Then there would be no information about outcomes under the exposure for this subpopulation. To identify the treatment effect, we could consider a different target parameter (e.g., the effect among those with \( W = 0 \)) or consider additional modeling assumptions (e.g., the effect is the same among those with \( W = 1 \) and \( W = 0 \)).
Both options are a bit dissatisfying and other approaches may be taken [37]. The risk of violating the positivity assumption is exacerbated with higher dimensional data (i.e., as the number of covariates or their levels grow).

In many cases, our initial assumptions, encoded in the structural causal model $\mathcal{M}^F$, are not sufficient to identify the causal effect $\Psi^F(\mathcal{P}_{U,X})$. Indeed, for our running example (Figure 20.2a), the set of baseline covariates is not sufficient to block the back-door paths from the outcome to the exposure. The question then becomes how to proceed? Possible options include giving up, gathering more data, or continuing to estimation while clearly acknowledging the lack of identifiability during the interpretation step. To facilitate the third option, we can use $\mathcal{M}^{F*}$ to denote the structural causal model, augmented with additional convenience-based assumptions needed for identifiability. This gives us a way to proceed, while separating our real knowledge $\mathcal{M}^F$ from our wished identifiability assumptions $\mathcal{M}^{F*}$.

Overall, identifiability assumptions and resulting estimands are specific to the causal parameter $\Psi^F(\mathcal{P}_{U,X})$. We are focusing on a point treatment effect (i.e., distribution of counterfactuals under interventions on a single node or variable). Different identifiability results are needed for interventions on more than one node (e.g., longitudinal treatment effects and direct effects) and interventions responding to patient characteristics (e.g., dynamic regimes). Furthermore, a given causal parameter may have more than one identifiability result (e.g., instrumental variables and the front-door criterion). See, for example, Pearl [14].

**Common Pitfall: Stating vs. Evaluating the Identifiability Assumptions**

There is a temptation to simply state the identifiability assumptions and proceed to the analysis. The identifiability assumptions require careful consideration. Directed

**FIGURE 20.4**

Considering the back-door criterion. (a) The set of covariates $W^2$ is sufficient to block the back-door path from $Y \rightarrow W^2 \rightarrow A$. Therefore, the randomization assumption will hold conditionally on $W^2$. Further adjustment for $W^1$ is unnecessary and potentially harmful. (b) The randomization assumption holds conditionally on $\emptyset$. Adjusting for $W$ (i.e., conditioning on a collider of $U$ and $U^*$) opens a back-door path and induces a spurious association between $A$ and $Y$. (c) The randomization assumption holds conditionally on $(W,L)$. The covariates $L$ are needed to block the back-door path from $Y \rightarrow L \rightarrow U \rightarrow A$, even though $L$ occurs temporally after the exposure $A$. 
acyclic graphs facilitate the evaluation of assumptions by subject-matter experts without extensive statistical training. When interpreting the analysis, any convenience-based causal assumptions should be transparently stated and explained.

20.7 Estimation and Inference

In the previous step, we defined the parameter of interest as a mapping from the statistical model to the parameter space: $\Psi : \mathcal{M} \to \mathbb{R}$. In other words, the statistical parameter is a function, whose input is any distribution $P$ compatible with the statistical model and whose output is a real number. The parameter mapping applied to the true observed data distribution $P_0$ is called the estimand and denoted $\Psi(P_0)$. Recall we have $n$ independent, identically distributed (i.i.d.) copies of the random variable $O = (W, A, Y)$. The empirical distribution $P_n$ corresponds to putting a weight $1/n$ on each copy of $O_i$. An estimator is a function, whose input is the observed data (a realization of $P_n$) and output a value in the parameter space.

In this chapter, we consider substitution estimators based on the G-computation identifiability result [27]:

$$\Psi(P_0) = E_0\left[ E_0(Y|A=1,W) - E_0(Y|A=0,W) \right].$$

(20.3)

A simple substitution estimator for $\Psi(P_0)$ can be implemented as follows:

1. Estimate the conditional expectation of the outcome, given the exposure and covariates, denoted $\hat{E}(Y|A,W)$.
2. Use this estimate to generate the predicted outcomes for each unit, setting $A = 1$ and $A = 0$.
3. Take the sample average of the difference in these predicted outcomes:

$$\hat{\Psi}(P_n) = \frac{1}{n} \sum_{i=1}^{n} \hat{E}(Y_i|A_i = 1, W_i) - \hat{E}(Y_i|A_i = 0, W_i).$$

The last step corresponds to estimating the marginal covariate distribution $P_0(W)$ with the sample proportion: $\frac{1}{n} \sum_i I(W_i = w)$.

There are many options available for estimating the conditional expectation $E_0(Y|A,W)$. Often, parametric models are used to relate the conditional mean outcome to the possible predictor variables and the exposure. Suppose, for example, we knew that the conditional expectation of a continuous outcome could be described by the following parametric model:

$$E_0(Y|A,W) = \beta_0 + \beta_1 A + \beta_2 W_1 + \beta_3 W_2 + \beta_4 A^* W_1 + \beta_5 A^* W_2,$$

where $W = \{W_1, W_2\}$ denotes the set of covariates, needed for identifiability. Then this knowledge should have been encoded in our structural causal model $\mathcal{M}_F$ with implied restrictions on our statistical model $\mathcal{M}$. (In other words, we avoid introducing new assumptions during the analysis.) The coefficients in this regression model could be estimated with maximum likelihood or with ordinary least squares regression. The estimate $\hat{\beta}_1$ does not, however, provide an estimate of the G-computation identifiability result. The
exact interpretation of $\hat{\beta}_1$ depends on which variables and which interactions are included in the parametric model. To obtain an estimate of $\Psi(P_0)$, we need to average the predicted outcomes with respect to the distribution of covariates:

$$
\Psi(P_n) = \frac{1}{n} \sum_{i=1}^{n} \hat{E}(Y_i|A_i = 1, W_i) - \hat{E}(Y_i|A_i = 0, W_i)
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} (\hat{\beta}_1 + \hat{\beta}_4 W_{1,i} + \hat{\beta}_5 W_{2,i}).
$$

As a second example, suppose we knew that the conditional risk of a binary outcome could be described by the following parametric model:

$$
\logit[E_0(Y|A, W)] = \beta_0 + \beta_1 A + \beta_2 W_1 + \cdots + \beta_{11} W_{10},
$$

where $W = \{W_1, \ldots, W_{10}\}$ denotes the set of covariates, needed for identifiability. Then the estimate $\hat{\beta}_1$ would provide an estimate of the logarithm of the conditional odds ratio. An estimate of the G-computation identifiability result is given by averaging the expected outcomes under the exposure $A = 1$ and control $A = 0$:

$$
\Psi(P_n) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{1 + \exp(- (\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_4 W_{1,i} + \cdots + \hat{\beta}_{11} W_{10,i}))} - \frac{1}{1 + \exp(- (\hat{\beta}_0 + \hat{\beta}_2 W_{1,i} + \cdots + \hat{\beta}_{11} W_{10,i}))} \right).
$$

In most cases, our background knowledge is inadequate to describe the conditional expectation $E_0(Y|A, W)$ with such parametric models. Indeed, with high dimensional data, the sheer number of potential covariates will likely make it impossible to correctly specify the functional form. If the assumed parametric model is incorrect, the point estimates will often be biased and inference misleading. In other words, the structural causal model $M_F$, representing our knowledge of the underlying data generating process, often implies a nonparametric statistical model $M$. Our estimation approach should respect the statistical model.

To avoid unsubstantiated assumptions about functional form, it is sometimes possible to estimate $E_0(Y|A, W)$ with the empirical mean in each exposure–covariate stratum. Unfortunately, even when all covariates are discrete valued, nonparametric maximum likelihood estimators quickly become ill-defined due to the curse of dimensionality; the number of possible exposure–covariate combinations far exceeds the number of observations. Again, this problem becomes exacerbated with big data, where, for example, there are hundreds of potential covariates under consideration.

Various model selection routines can help alleviate these problems. For example, stepwise regression will add and subtract variables in hopes of minimizing the Akaike information criterion or the Bayesian information criterion. Other data-adaptive methods, based on cross-validation, involve splitting the data into training and validation sets. Each possible algorithm (e.g., various parametric models or semiparametric methods) is then fit on the training set and its performance assessed on the validation set. The measure of performance can be defined by a loss function, such as the L2-squared error or the negative log likelihood. Super learner, for example, uses cross-validation to select the candidate algorithm with the best performance or to build the optimal (convex) combination of estimates from candidate algorithms [38,39]. (For further details, see Chapter 19.) A point estimate could then be obtained by averaging the difference in predicted outcomes for each unit under the exposure and under the control.

Although these data-adaptive methods avoid betting on one a priori specified parametric regression model and are amenable to semiparametric algorithms, there is no reliable
way to obtain statistical inference for parameters, such as the G-computation estimand \( \Psi(P_0) \). Treating the final algorithm as if it were prespecified ignores the selection process. Furthermore, the selected algorithm was tailored to maximize/minimize some criterion with regard to the conditional expectation \( E_0(Y|A,W) \) and will, in general, not provide the best bias–variance trade-off for estimating the statistical parameter \( \Psi(P_0) \). Indeed, estimating the conditional mean outcome \( Y \) in every stratum of \( (A,W) \) is a much more ambitious task than estimating one number (the difference in conditional means, averaged with respect to the covariate distribution). Thus, without an additional step, the resulting estimator will be overly biased relative to its standard error, preventing accurate inference.

Targeted maximum likelihood estimation (TMLE) provides a way forward \([3,40]\). TMLE is a general algorithm for the construction of double robust, semiparametric, efficient substitution estimators. TMLE allows for data-adaptive estimation while obtaining valid statistical inference. The algorithm is detailed in Chapter 22. Although TMLE is a general algorithm for a wide range of parameters, we focus on its implementation for the G-computation estimand. Briefly, the TMLE algorithm uses information in the estimated exposure mechanism \( \hat{P}(A|W) \) to update the initial estimator of the conditional mean \( E_0(Y|A,W) \). The targeted estimates are then substituted into the parameter mapping. The updating step achieves a targeted bias reduction for the parameter of interest \( \Psi(P_0) \) and serves to solve the efficient score equation. As a result, TMLE is a double robust estimator; it will be consistent for \( \Psi(P_0) \) is either the conditional expectation \( E_0(Y|A,W) \) or the exposure mechanism \( P_0(A|W) \) is estimated consistently. When both functions are consistently estimated at a fast enough rate, the TMLE will be efficient in that it achieves the lowest asymptotic variance among a large class of estimators. These asymptotic properties typically translate into lower bias and variance in finite samples. The advantages of TMLE have been repeatedly demonstrated in both simulation studies and applied analyses \([37,41–43]\). The procedure is available with standard software such as the \texttt{tmle} and \texttt{ltmle} packages in R \([44–46]\).

Thus far, we have discussed obtaining a point estimate from a simple or targeted substitution estimator. To create confidence intervals and test hypotheses, we also need to quantify uncertainty. A simple substitution estimator based a correctly specified parametric model is asymptotically linear, and its variance can be approximated by the variance of its influence curve, divided by sample size \( n \). It is worth emphasizing that our estimand \( \Psi(P_0) \) often does not correspond to a single coefficient, and therefore we usually cannot read off the reported standard error from common software. Under reasonable conditions, the TMLE is also asymptotically linear and inference can be based on an estimate of its influence curve.

Overall, this chapter focused on substitution estimators (simple and targeted) of the G-computation identifiability result \([27]\). The simple substitution estimator only requires an estimate of the marginal distribution of baseline covariates \( P_0(W) \) and the conditional expectation of the outcome, given the exposure and covariates \( E_0(Y|A,W) \). TMLE also requires an estimate of the exposure mechanism \( P_0(A|W) \). There are many other algorithms available for estimation of \( \Psi(P_0) \). A popular class of estimators relies only on estimation of the exposure mechanism \([47–49]\). Inverse probability of treatment weighting (IPTW) estimators, for example, control for measured confounders by up-weighting exposure–covariate groups that are underrepresented and down-weighting exposure–covariate groups that are overrepresented (relative to what would be seen were the exposure randomized). Its double robust counterpart, augmented-IPTW, shares many of the same properties as TMLE \([50,51]\). A key distinction is that IPTW and augmented-IPTW are solutions to estimating equations and therefore respond differently in the face of challenges due to strong confounding and rare outcomes \([37,52]\). Throughout, we maintain that estimators should...
respect knowledge encoded in the statistical model and not introduce new assumptions. An estimator should be selected for analysis based on its performance (e.g., bias, variance, robustness) as opposed to convenience or habit.

Common Pitfall: Confusing Estimation Methods with the Causal Parameters

Causal models and causal parameters help to specify a statistical estimation problem (i.e., the observed data, statistical model, and estimand) that is optimally informed by background knowledge and aims to answer the underlying scientific or policy question. However, there is nothing causal about the estimation step. A given estimand can be estimated in many different ways, and alternative algorithms can be compared simply based on their statistical properties, such as bias and variance. For example, (working) marginal structural models are often used to define a target counterfactual parameter equal, under needed causal assumptions, to a specific estimand. This estimand can be estimated with inverse probability weights [31,53], regression of the outcome on exposure and confounders, or double robust efficient methods [3,54]. There is nothing more or less causal about these estimators.

20.8 Interpretation of the Results

The last step of the roadmap is interpreting the results. In our running example, the identifiability assumptions did not hold. Nonetheless, the statistical estimand (Equation 20.3) always has a statistical interpretation as the difference in the expected outcome, given the exposure and covariates in the adjustment set, and the expected outcome, given the control and covariates in the adjustment set, standardized with respect to the covariate distribution in the population. For our example, $\Psi(P_0)$ can be interpreted as the marginal risk difference: the difference in the mortality risk among patients with early versus delayed ART initiation but the same values of the measured covariates (e.g., baseline CD4 count, age, and sex), averaged with respect to the distribution of these covariates. This estimand can be considered as the best approximation to the causal quantity of interest, given the limitations in the observed data. If the identifiability assumptions hold, our estimate would be endowed with a causal interpretation: a summary of how the distribution of the data would change under a specific intervention. For our example, the causal interpretation would be the difference in the 5-year counterfactual mortality risk if all patients initiated early ART versus if all patients delayed ART initiation. Further interpretation in terms of the impact of a real-world intervention or in terms of a randomized trial requires additional assumptions.

Common Pitfall: Lack of Identifiability Is Different from Statistical Bias

During the identifiability step, we advocate that a clear distinction be made between assumptions based on knowledge, encoded in the structural causal model $M^F$, and those
based on convenience $M^Z$. This delineation emphasizes that the estimand may not equal the causal parameter. The discrepancy depends on unmeasured quantities and non-testable assumptions. In other words, the needed assumptions cannot be evaluated statistically using the observed data alone [1]. Nonetheless, sensitivity analyses can help in evaluating the potential magnitude of the deviations between the causal parameter and the statistical estimand [55–58]. By contrast, the statistical bias of an estimator is a statistical concept, characterizing how an estimator performs on average across multiple repetitions the experiment. Statistical bias can be evaluated through simulations and minimized with data-driven techniques.

20.9 Conclusion

In this chapter, we introduced a formal framework for causal inference [3,4]. Our running example was to estimate the effect of early ART initiation (within 1 month of diagnosis) on 5-year mortality risk among HIV+ adults in sub-Saharan Africa. Our structural causal model $M^Z$ only reflected the causal ordering of our variables; we did not make any exclusion restrictions, independence assumptions, or functional form assumptions. Counterfactual outcomes were generated by deterministically intervening on the data generating system, described by the structural causal model, to set $A = 1$ (i.e., early initiation) and also to set $A = 0$ (i.e., delayed initiation). We focused on the average treatment effect for this static exposure. The observed data $O = (W, A, Y)$ were assumed to be generated by sampling $n$ independent times from a probability distribution compatible with the structural causal model $M^Z$, which implied a non-parametric statistical model $M$. Although our identifiability assumptions did not hold, we still defined a statistical estimand $\Psi(P_0)$ as a best approximation of our wished for causal quantity. We briefly discussed a simple (parametric) substitution estimator and a targeted substitution estimator (TMLE), which allows for data-adaptive estimation while obtaining valid inference. Because our needed identifiability assumptions were not met, we interpreted our estimate as the marginal difference in the mortality risk, given early ART initiation and the measured covariates, and the mortality risk, given delayed ART initiation and the measured covariates, standardized with respect to the covariate distribution.

This framework is easily extended to more complicated data structures. Consider, for example, the following scientific questions, corresponding to interventions on multiple exposure nodes and to alternate counterfactual treatment assignment mechanisms:

- **Longitudinal treatment effects** [31,51,53,54,59–69]: How does cumulative time until ART initiation affect mortality among recently diagnosed HIV+ adults? What is the effect of routine HIV viral load monitoring, compared to routine CD4+ T cell count monitoring, on mortality among patients initiating early ART? What would be impact of early ART initiation on the 5-year mortality if there were no losses to follow-up?

- **Dynamic regimes** (individualized treatment rules) [22–25,61,70–72]: How would mortality have differed if HIV+ adults initiated ART based on HIV RNA viral loads as opposed to CD4+ T cell counts?
• Direct and indirect effects [73–76]: What is the direct effect of early ART initiation on 5-year mortality that is not mediated through changes in HIV RNA viral load?

• Stochastic interventions (nondeterministic interventions) [26]: What would be the 5-year mortality if the distribution of time until ART initiation shifted toward shorter wait times? What is the impact of early ART initiation on 5-year mortality if HIV RNA viral load, the intermediate, remained at the value it would have been in the absence of the exposure (i.e., the natural direct effect [77–79])?

Overall, access to unprecedented amounts of data does not undo the age-old adage: correlation is not causation. Indeed, there are numerous sources of association (dependence) between two variables: direct effects, indirect effects, measured confounding, unmeasured confounding, and selection bias. The methods, introduced here, allow researchers to move from saying drug $X$ is associated with an adverse side effect to saying (under the necessary and transparently stated assumptions) an adverse side effect is caused by drug $X$. Even if the needed identifiability assumptions are not expected to hold, this framework helps us to estimate a statistical parameter, coming as close to the wished causal parameter. In other words, this framework ensures that the scientific question is driving the analysis and not the other way around.

Appendix: Extensions to Multiple Time Point Interventions

As an introduction to causal inference, we focused on causal parameters corresponding to a static intervention on a single node. In this appendix, we step through the causal roadmap for an example of a longitudinal effect, corresponding to a multiple time point intervention.

Step 1—Specify the scientific question: What is the effect of delayed ART initiation on patient outcomes? As before, we want to be specific about the target population: recently diagnosed HIV+ adults in sub-Saharan Africa. We also need to be clear about the definition and timing of the exposures. For simplicity, let us assume that the patients have monthly clinic visits and therefore could initiate ART or not each month. (This framework could easily be extended to shorter or longer time intervals.) Suppose the outcome is viral suppression after 12 months of follow-up.

Step 2—Specify the causal model: Let baseline ($t = 0$) be the time the patient is diagnosed with HIV. Let $L_0$ represent the vector of baseline covariates, including sociodemographics, clinical measurements, and social constructs. Likewise, let $L_t$ represent the vector of time-updated covariates (e.g., clinical measurements). Let $A_t$ be an indicator that the patient initiated ART at time $t$. For example, $A_0 = 1$ represents starting ART on the same day as diagnosis (i.e., month 0), whereas $A_1 = 1$ represents initiation at the first month visit. Finally, let $Y$ be an indicator that the patient had undetectable HIV RNA viral load at the end of follow-up. For simplicity, let us consider only three time points and assume complete follow-up. Our structural causal model $M^F$, only reflecting the causal ordering, is given by

Endogenous nodes: $X = (L_0, A_0, L_1, A_1, Y)$

Exogenous nodes: $U = (U_{L_0}, U_{A_0}, U_{L_1}, U_{A_1}, U_Y)$ with some true joint distribution $P_{U,0}$. We place no assumptions on the set of possible distributions for $U$. (During the identifiability step, we will need to make some independence assumptions. However,
we want to keep our true knowledge, as specified by structural causal model $\mathcal{M}^F$, separate from the additional assumptions needed for identifiability.)

*Structural equations:*

\[
\begin{align*}
L_0 &= f_{L_0}(U_{L_0}) \\
A_0 &= f_{A_0}(L_0, U_{A_0}) \\
L_1 &= f_{L_1}(L_0, A_0, U_{L_1}) \\
A_1 &= f_{A_1}(L_0, A_0, L_1, U_{A_1}) \\
Y &= f_Y(L_0, A_0, L_1, A_1, U_Y).
\end{align*}
\]

We have not made any exclusion restrictions or independence assumptions. The corresponding directed acyclic graph is given in Figure 20.5a.

*Step 3—Specify the target causal quantity:* Let $Y(a_0, a_1)$ denote the counterfactual outcome (viral suppression) if a patient, possibly contrary to fact, had treatment history $(a_0, a_1)$. Counterfactuals are generated by intervening on the structural causal model:

\[
\begin{align*}
L_0 &= f_{L_0}(U_{L_0}) \\
A_0 &= a_0 \\
L_1 &= f_{L_1}(L_0, a_0, U_{L_1}) \\
A_1 &= a_1 \\
Y &= f_Y(L_0, a_0, L_1, a_1, U_Y).
\end{align*}
\]

For the two binary exposures (initiate or not at time $t$), the set of possible exposure combinations is $\mathcal{A} = \{10, 01, 00\}$. For example, $Y(0, 1)$ corresponds to preventing ART initiation at month 0 and starting ART at the 1 month clinic visit. Suppose our goal is to contrast expected counterfactual outcome if, possibly contrary to fact, all patients immediately initiated ART with the the expected counterfactual outcome if, possibly contrary to fact, all patients delayed ART initiation until 1 month after diagnosis:

\[
\Psi^F(P_{U,X,0}) = E_{U,X,0}[Y(1,0) - Y(0,1)].
\]

**FIGURE 20.5**

Directed acyclic graph corresponding to the longitudinal effect when (a) we make no independence assumptions on background factors and (b) when we assume that the background factors are all independent. $L_0$ denotes baseline covariates; $A_0$ denotes whether the patient initiated ART at $t = 0$; $L_1$ denotes time-updated covariates; $A_1$ denotes whether the patient initiated ART at $t = 1$; and $Y$ denotes undetectable viral load.
Step 4—Specify the observed data and its link to the causal model: The observed data consist of $n$ i.i.d. copies of

$$O = (L_0, A_0, L_1, A_1, Y) \sim P_0.$$  

We assume that the observed data were generated by sampling $n$ independent times from a data generating process compatible with $M^F$. The resulting statistical model $M$, describing the possible observed data distributions, is nonparametric.

Step 5—Assess identifiability: For the purposes of discussion, suppose that the unmeasured factors $U = (U_{L_0}, U_{A_0}, U_{L_1}, U_{A_1}, U_Y)$ are all independent (Figure 20.5b). Even if this assumption held, there is not one set of covariates that simultaneously satisfy the back-door criterion for all intervention nodes. The baseline covariates $L_0$ alone fail, because there is an unblocked back-door path from $Y$ through $L_1$ to $A_1$. In other words, the effect of initiation at 1 month $A_1$ on the outcome $Y$ is confounded by time-updated covariates $L_1$. The baseline and time-updated covariates $(L_0, L_1)$ jointly fail, because we are losing (blocking) the effect of early ART initiation $A_0$ on the outcome $Y$ that goes through the covariates $L_1$. This challenge is generally known as time-dependent confounding [27,31,48]: time-varying covariates confound the effect of future exposures on the outcome, but are affected by past exposures.

To identify the effects of longitudinal interventions, we consider the problem sequentially. For each $A_k$ in sequence, we ask if its effect on $Y$ can be identified by conditioning on some subset of the observed past. This leads to the sequential randomization assumption [27]:

$$Y(a_0, a_1) \perp A_0 | L_0 \text{ and } Y(a_0, a_1) \perp A_1 | (L_0, A_0, L_1).$$

In words, we assume that the counterfactual outcome $Y(a_0, a_1)$ is independent from the intervention $A_k$ at time $k$, given the observed past. With the sequential randomization assumption as well a longitudinal version of the positivity assumption, the expectation of counterfactual outcomes, indexed by multiple interventions, can be identified by the longitudinal G-computation formula [27]:

$$E_{U,X,0}[Y(a_0, a_1)] = \sum_{l_0, l_1} E_0(Y | A_1 = a_1, L_1 = l_1, A_0 = a_0, L_0 = l_0) \times P_0(L_1 = l_1 | A_0 = a_0, L_0 = l_0)P_0(L_0 = l_0) = \Psi(P_0).$$

Now we are averaging with respect to the appropriate distribution of covariates and thereby capturing the effect of both exposures $(a_1, a_0)$ on the outcome $Y$ through the covariates $(L_0, L_1)$.

Step 6—Estimation and inference: As with single time point interventions, there are a variety of methods to estimate statistical parameters, corresponding under the necessary assumptions to longitudinal causal effects. Examples include longitudinal IPTW, parametric G-computation (maximum likelihood estimation of the longitudinal G-computation formula), and TMLE [31,51,53,54,59–69].

Step 7—Interpret: As with the single time point setting, the strength of our interpretations depends on rigorous evaluation of the needed assumptions. Even when the identifiability assumptions do not hold, then we always have a statistical interpretation of $\Psi(P_0)$. 
Acknowledgments

This work was supported, in part, by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number R01AI074345. The content is solely the responsibility of the authors and does not necessarily represent the officials views of the National Institutes of Health. Maya Petersen is a recipient of a Doris Duke Clinical Scientist Development Award.

References


